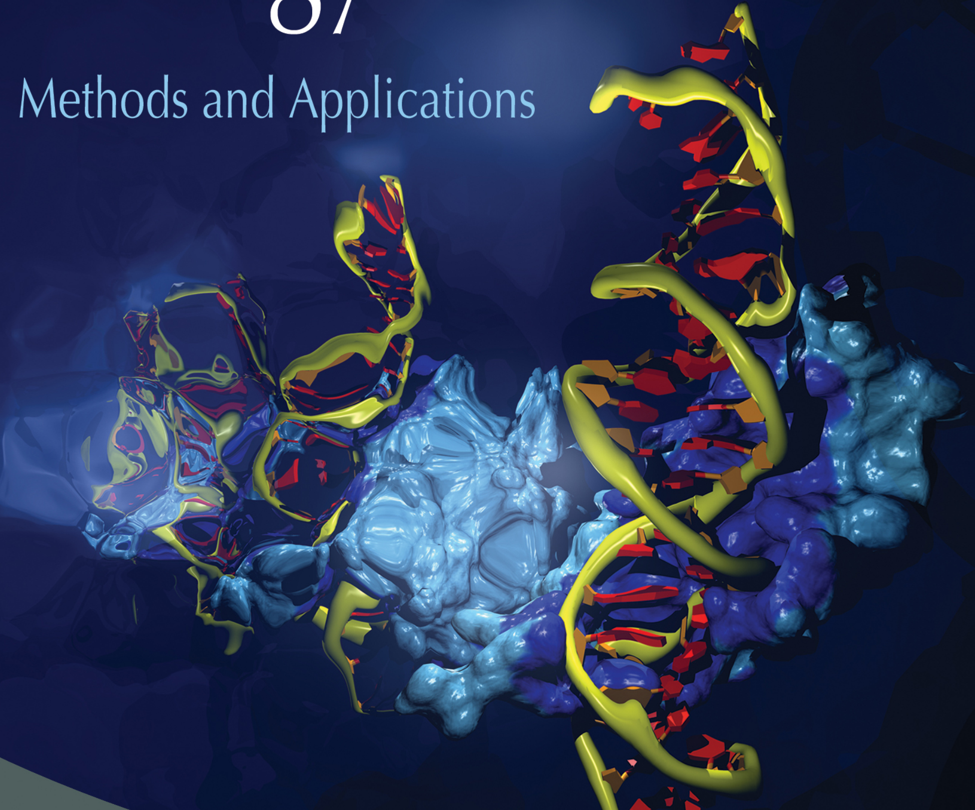
Torsten Schwede • Manuel Peitsch

Computational Structural Biology

Methods and Applications



World Scientific

Computational Structural Biology

Methods and Applications

This page intentionally left blank

Computational Structural Biology

Methods and Applications

Torsten Schwede
Swiss Institute of Bioinformatics
Biozentrum University of Basel, Switzerland

Manuel Peitsch
Novartis Institutes of BioMedical Research, Switzerland

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data
Computational structural biology : methods and applications / editors, Torsten Schwede, Manuel C. Peitsch.
p. ; cm.
Includes bibliographical references and index.
ISBN-13: 978-981-277-877-2 (hardcover : alk. paper)
ISBN-10: 981-277-877-2 (hardcover : alk. paper)
1. Proteins--Structure--Computer simulation. 2. Protein folding--Computer simulation.
3. Proteins--Structure--Mathematical models. 4. Protein folding--Mathematical models.
5. Computational biology. I. Schwede, Torsten. II. Peitsch, Manuel C.
[DNLM: 1. Computational Biology--methods. 2. Protein Conformation. 3. Drug Design.
4. Nucleic Acid Conformation. 5. Protein Folding. 6. RNA. QU 26.5 C7385 2008]
QP551.C714 2008
572'.6330285--dc22
2008001880

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

The cover illustration “Distortions” was created by Ansgar Philippsen, using DINO (www.dino3d.org) and POVray (www.povray.org).

Distortions

The act of visualization aims to reflect the data into a tangible representation, yet, within its contextualized layers of interpretation, a distortion of the original state appears to be an almost unavoidable consequence.

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore.

Preface

Computational structural biology aims primarily at establishing sequence-structure-function relationships for biological molecules using *in silico* techniques. This discipline emerged about 40 years ago (Levitt M. (2001). *Nature Struct Biol* **8**:392–393) and has made much progress in the past decade. The purpose of this book is to provide an overview of the progress in the field and to articulate some of the key challenges for the coming years. By no means could we cover the field comprehensively in just one book, and we thus focused on the structure and function of proteins and RNAs.

The advent of large genome sequencing reinforced the observation that structural information is needed to understand the detailed function and mechanism of biological molecules such as enzyme reactions and molecular recognition events. Furthermore, structures are obviously key to the design of molecules with new or improved functions. In this context, computational structural biology emerged as a discipline to develop computational tools to analyze and predict molecular structures and simulate their dynamical behavior. These theoretical approaches provide valuable insights into the detailed basis of molecular function and enable the effective design of experimental approaches to functional genomics. Major research topics include protein and RNA structure prediction, protein folding, protein and RNA dynamics with emphasis on large complexes and assemblies, molecular recognition, drug discovery and protein engineering.

A key motivation for putting together this book came from our own experience, as 15 years ago we established the Swiss-Model, the first Web-based server for protein structure modeling. One major driver behind our vision was to mask much of the complexity associated with protein modeling behind a simple interface, thereby providing the scientific community with the possibility to gain insights into the 3D structures of proteins of interest, without the need to learn and purchase complex and expensive software. There are probably three major factors contributing to the success of the Swiss-Model. First, the server is easy to use, as the Web-interface removes most of the complexity normally associated with protein modeling. Second, DeepView (also known as the Swiss-PdbViewer), which is available for most relevant computer platforms, has many powerful and easy-to-use features developed by modelers for modelers. Third, the uninterrupted operations for 15 years has allowed us to develop a robust and stable system. Today, well over 60,000 users build in excess of 400,000 models every single year and can access over a million pre-computed models available in the Swiss-Model Repository. Our objective is to continuously improve the performance of the server and the quality of the models it generates.

We are particularly thankful to Nicolas Guex for his many crucial contributions to the development efforts of the Swiss-Model and the DeepView and to Gale Rhodes of the University of Southern Maine for coordinating the active DeepView user community. We also thank Alexander Diemand, Konstatin Arnold, Jürgen Kopp and Lorenza Bordoli for their many contributions to the development and operations for the modeling platform. Furthermore, we are deeply indebted to Jake V. Maizel Jr, Timothy N.C. Wells, Jonathan C.K. Knowles, and Allan Baxter who have provided the necessary environment and resources during the various phases of this project. Finally, we thank Stanley K. Burt, Robert W. Lebherz III, Karol Miaskiewicz and Jack R. Collins of the Advanced Biomedical Computing Center at the National Cancer Institute in Frederick Maryland for their support in operating the US mirror of the Swiss-Model server. We gratefully acknowledge the financial support by GlaxoSmithKline, Novartis, the Swiss National Science Foundation, the Biozentrum of the University of Basel and the Swiss Institute of Bioinformatics.

T. Schwede and M.C. Peitsch

Contents

Section I

Structure Prediction and Assessment Methods

This page intentionally left blank

Chapter 1

Protein Structure Modeling

T. Schwede*,†, A. Sali‡, N. Eswar‡
and M. C. Peitsch§

1.1 Introduction

Knowledge of the three-dimensional (3D) structures of proteins provides invaluable insights into the molecular basis of their functions. Furthermore, the design of experiments aimed at understanding molecular mechanisms — such as site-directed mutagenesis, mapping of disease-related mutations, and the structure-based design of specific inhibitors — are greatly facilitated by the detailed knowledge of the spatial arrangement of key amino acid residues within the overall 3D structure. While great progress has been made in structure determination using experimental methods, such as X-ray crystallography (Chapter 22), high-resolution electron microscopy (Chapter 23) and

*Corresponding author.

†Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland. E-mail: torsten.schwede@unibas.ch.

‡Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, California Institute for Quantitative Biosciences (QB3), University of California at San Francisco, Byers Hall at Mission Bay, Suite 503B, 1700 4th Street, San Francisco, CA 94158-2330, USA.

§Novartis Institutes of BioMedical Research, Basel Klybeckstrasse 141, 4057 Basel, Switzerland.

nuclear magnetic resonance (NMR) spectroscopy (Chapter 24), these approaches are generally still expensive, time consuming, and not always applicable. Currently, less than 50 000 experimental protein structures have been released by the Protein Data Bank PDB[1] (Table 1.1), while another 3500 have been deposited but are still awaiting release. These structures correspond to approximately 17 000 different proteins (sharing less than 90% sequence identity among one another). Nevertheless, the number of structurally characterized proteins is small compared to the 300 000 annotated and curated protein sequences in the Swiss-Prot section of the UniProtKB[2] (http://www.expasy.org/sprot/). This number is even smaller when compared to the 5.2 million known protein sequences in the complete UniProtKB (October 2007). Even after removal of the highly redundant sequences from this database (above), the remaining 3.3 million sequences exceed the number of known 3D structures by more than two orders of magnitude. Thus, no experimental structure is available for the vast majority of protein sequences. This gap has widened over the last decade, despite the high-throughput X-ray crystallography pipelines developed for structural genomics.[3–5] Therefore, the gap in structural knowledge must be bridged by computation.

Computational methods for predicting the 3D structures of proteins enjoy a high degree of interest and are the focus of many

Table 1.1. Current PDB Holdings (October 2007)[a]

		Molecule Type				
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
Experimental Method	X-ray	36847	991	1709	24	39571
	NMR	5929	788	134	7	6858
	EM	106	11	40	0	157
	Other	83	4	4	2	93
	Total	42965	1794	1887	33	46679

[a] The content of the table was obtained from http://www.pdb.org (1). EM: electron microscopy.

research and service development efforts. The prediction of the 3D structure of a protein from its amino acid sequence remains a fundamental scientific problem and it is often considered as one of the *grand challenges* in computational biology and chemistry. Broadly, four different types of approaches are commonly in use. The first and most accurate approach is "comparative" or "homology" modeling that uses experimentally elucidated structures of related protein family members as templates to model the structure of the protein of interest (the "target"). These methods can only be employed when a detectable template of known structure is available. Second, fold recognition and threading methods are used to model proteins that have low or statistically insignificant sequence similarity to proteins of known structure (Chapter 2). Third, *de novo* (or *ab initio*) methods aim to predict the structure of a protein purely from its primary sequence, using principles of physics that govern protein folding and/or using information derived from known structures but without relying on any evolutionary relationship to known folds. Finally, a fourth group of methods, recently receiving a lot of attention, is the "integrative" or "hybrid" methods that combine information from a varied set of computational and experimental sources, including all those listed above.

1.2 Modeling Methods

1.2.1 *Comparative Protein Structure Modeling Techniques*

Template-based protein modeling techniques (aka "homology modeling" or "comparative modeling") exploit the evolutionary relationship between a target protein and templates with known experimental structures, based on the observation that evolutionarily related sequences generally have similar 3D structures. Most comparative modeling procedures consist of several consecutive steps, which can be repeated iteratively until a satisfactory model is obtained: 1) identification of suitable template structures related to the target protein and the alignment of the target and template(s) sequences; 2) modeling of the structurally conserved regions and the prediction of structurally variable regions; 3) refinement of the initial model; and 4) evaluation of the resulting model(s).

1.2.1.1 *Identification of modeling templates and sequence alignments*

Identifying suitable template structures and calculating an accurate alignment of their sequences with that of the target are the key first steps of the comparative modeling process. The sequence identity of the target-template alignment is the most commonly used metric to quantify the similarity between the target and template(s) and is also a good predictor of the quality of the resulting model. It is thus crucial to consider the target-template sequence identity level when selecting template structures (Sections 1.2.2, 1.6 and Chapter 5), as this will have a critical impact on the quality of the resulting model and hence, its potential applications. The overall accuracy of models calculated from alignments with sequence identities of 40% or higher is almost always good (i.e. deviate by less than 2Å RMSD from the experimentally determined structure) (Section 1.2.2). As the target–template sequence identity falls below 30–40%, models that deviate significantly from the average accuracy are frequent (i.e. deviate by more than 2Å RMSD from an experimentally-determined structure). Alignment errors also tend to rapidly increase in this regime and become the most frequent cause of large errors in comparative models even when the correct template is chosen. Moreover, models based on alignments with such low sequence identities may have an entirely incorrect fold.[6]

While identifying and aligning sequences with similarities above 40% is relatively straightforward, more sensitive methods are needed for the lower levels of evolutionary relatedness between sequences. In recent years, significant progress has been made in the development of sensitive methods for sequence homology detection and alignment based on iterative profile searches, e.g. PSI-Blast,[7] Hidden Markov Models, e.g. SAM,[8] HMMER,[9] or profile-profile alignment such as FFAS03,[10] profile.scan,[11] and HHsearch.[12] Furthermore, in the absence of a detectable sequence similarity, fold recognition and threading methods can be used to identify proteins with known structures, that share a common fold with the target sequence (Chapter 2).

1.2.1.2 *Generating all-atom models*

Comparative protein structure modeling yields an all-atom model of a protein, based on its alignment to one or more related template structures. Over the years, two commonly used approaches for model building have emerged and can be described as follows: the first is a rigid fragment assembly approach, in which an initial model is constructed from structurally conserved core regions of the template and from structural fragments obtained from either aligned or unrelated structures.[13,14] The initial model is then subjected to an optimization procedure to refine its geometry and stereochemistry (Section 1.2.1.3). The second approach relies on a single optimization strategy that attempts to maximize the satisfaction of spatial restraints obtained from the target-template alignment, known protein structures, and molecular mechanics force-fields.[15] Such an approach may not require a separate refinement step. However, most model building procedures are usually followed by the application of specialized protocols to enhance the accuracy of the non-conserved regions of the alignment such as loops[16,17] and/or side chains.[18,19]

1.2.1.3 *Model refinement*

Once an atomic model has been obtained, it can potentially be refined to idealize bond geometry and to remove unfavorable contacts that may have been introduced by the initial modeling process. The refinement will generally begin with an energy minimization step using one of the molecular mechanics force fields.[20,21] For further refinement, techniques such as molecular dynamics as well as Monte Carlo and genetic algorithm-based sampling methods[22–24] can be applied. For instance, in certain cases molecular dynamics has been reported to yield some improvement of side chain contacts and rotamer states.[25]

Monte Carlo sampling with focus on regions most likely to contain errors, while allowing the whole structure to relax in a physically realistic all-atom force field, can significantly improve the accuracy of models in terms of both the backbone conformations and the placement of core side chains.[26] Nevertheless, limitations still exist in sampling as well as force field accuracy.

1.2.1.4 *Model evaluation*

Model evaluation aims to recognize the various problems that might have occurred during the modeling process. Furthermore, estimating the overall geometrical accuracy of the individual regions of the model is an essential task of model evaluation. There are two kinds of evaluation schemes that are commonly employed. The first is "fold-assessment" that seeks to ensure the calculated models possess the correct fold and helps in detecting errors in template selection, fold recognition, and target-template alignment.[6,27–29] The second class of methods seeks to identify the model that is closest to the native structure out of a number of alternative models.[30–37] A combination of such assessments is usually employed to select the most accurate model from amongst a set of alternative models, generated based on different templates and/or alignments. In general, addressing these different types of assessment requires specialized scoring systems and classifiers (Chapters 3 and 4).

1.2.2 *Accuracy and Limitations of Comparative Protein Structure Modeling*

Comparative protein structure modeling relies on the evolutionary relationship between the target and template proteins. Consequently, the application of this approach is limited by 1) the availability of suitable template structures; 2) the ability of alignment methods to calculate an accurate alignment between the target and template sequences, even when the relationship between them is remote; and 3) the structural and functional divergence between the target and the template.[38]

The percentage of sequence identity between target and template correlates with model accuracy and often allows for a good first estimate of the model quality. As a rule of thumb, comparative models based on more than 50% sequence identity to their templates can be considered as "high accuracy models" and tend to have about 1 Å root mean square deviation (RMSD)[38] for the main-chain atoms, which is comparable to the accuracy of a medium-resolution NMR-derived structure or a low-resolution X-ray structure.[5,39] Inaccuracies are mainly found in the packing of side chains and loop regions. Comparative models based on 30 to 50% sequence identity can be considered "medium accuracy models", where the most frequent errors include side-chain packing errors, slight distortions of the protein core, inaccurate loop modeling, and sporadic alignment mistakes. Since alignment errors increase rapidly below 30% sequence identity and become the most substantial origin of errors in comparative models, comparative models based on less than 30% sequence identity are considered "low accuracy models".

1.2.2.1 *Template availability and structural diversity*

It has been observed that a very small number of different folds account for the majority of known structures,[40] and a recent study has argued that most sequences could already be modeled using known folds (or fragments of known folds) as templates.[41] Thus, for the majority of target protein domains, a structure with a similar fold would be available within the Protein Data Bank (PDB). However, models based on alignments with low sequence identity often provide accurate information only about the fold of the protein. As stated above, the accuracy of homology models decreases rapidly when the sequence identity between the target and template drops below 30%, mainly due to alignment errors and our inability to model structural differences between the target and the template. While the overall fold of proteins is often well conserved even at undetectable levels of sequence similarity, protein function — such as enzyme function and specificity — shows much higher variability,[42,43] even at high levels of sequence identity (above 50%). New methods beyond simple homology-based assignments are therefore required for functional annotation of new genomic sequences, taking into account specific local structural features.

1.2.2.2 *Natively unstructured proteins*

Intrinsic disorder in proteins, i.e. the presence of unstructured regions, has been a focus of much attention recently, as it has been shown to be implicated in important biological roles, such as translation and transcriptional regulation, cell signaling, and molecular recognition in general. Several studies report examples of disordered proteins implicated in important cellular processes, undergoing transitions to more structured states upon binding to their target ligand, DNA, or other proteins.[44–46] New biological functions linked to native disorder are emerging, such as self-assembly of multi-protein complexes or involvement in RNA and protein chaperones.[47,48] Natively unstructured proteins pose a challenge for experimental structural determination as they can hinder the crystallization of proteins or interfere with NMR spectroscopy. Consequently, such proteins are also not amenable to modeling techniques, as it is unclear to what extent the "correct" conformation can be inferred by comparative modeling, as these protein regions depend on the context of a folded scaffold to assume a defined structure. However, computational approaches for detecting regions in protein sequences with a high propensity for intrinsic disorder have been successfully developed, based on the observation that such protein segments possess characteristic sequence properties.[49–52]

1.2.2.3 *Membrane proteins*

Membrane proteins are involved in a broad range of central cellular processes, including signaling and intercellular communication, vesicle trafficking, ion transport, and protein translocation. It is not surprising that the targets for ~40% of all therapeutic drugs in use today are human membrane proteins. These include targets such as ion channels, reuptake pumps as targets for anti-depressants, and the important group of 7-transmembrane G-protein coupled receptors

(GPCRs). However, membrane proteins pose formidable challenges to experimental structure determination by X-ray crystallography and NMR spectroscopy. Furthermore, human proteins often have no closely related homologs in prokaryotes or *archaea*, which would facilitate expression and crystallization. As a consequence, structures of membrane proteins are significantly underrepresented in the PDB. The 3D structures of only ~135 different membrane proteins are currently publicly available (1 January 2008). Consequently, prediction of membrane protein structures based on physical models that describe intra-protein and protein–solvent interactions in the membrane environment without relying on homologous template structures has been attempted by several groups.[53,54] An important challenge in the modeling of membrane protein structures is the presumed difference relative to the globular proteins. For example, it is believed that membrane proteins are "inside-out" globular proteins, with hydrophobic residues on the outside in contact with the lipid bilayer and polar residues on the inside in the protein core. This design may render the standard scoring functions used for the modeling of globular proteins less suitable for use with membrane proteins. Most recently, a new scoring function was developed in Rosetta to account for such differences.[55]

1.2.3 *De novo Modeling Techniques*

Comparative protein structure modeling methods are only able to produce highly accurate models for protein sequences for which sufficient template information is available on the structures of homologous proteins. However, these methods are not suited to predict parts of sequences that are not aligned with the template sequences, e.g. long variable loop regions, or completely novel folds that have not been observed before. In contrast, *de novo* modeling methods do not explicitly rely on whole known structures as templates. Thus, the structure of any protein can be predicted by these *de novo* methods.

The term *ab initio* prediction often refers to the subset of *de novo* methods that rely on energy functions based solely on physicochemical interactions, not on the PDB. Such approaches, using full-atom simulations with empirical force fields as well as explicit and implicit solvent models, have been successful in predicting the folding of short peptides[56,57] and in discriminating between the native state and a static set of decoys.[58] However, from a practical protein structure prediction perspective, there are still limitations with regards to protein size and accuracy of the predictions.

Most of the successful *de novo* prediction methods that are applicable to larger protein segments (up to ~150 residues) use information from known protein structures.[59] *De novo* methods assume that the native state of a protein is at the global free energy minimum and carry out a large-scale search of conformational space for protein tertiary structures that are particularly low in free energy for the given amino acid sequence. The working hypothesis of this approach is that local amino acid sequence propensities bias each local segment of a polypeptide chain towards a small number of alternative local structures and that non-local interactions preferentially stabilize native-like arrangements of these otherwise transitory local structures. For example, the Rosetta method developed by Baker and coworkers uses an ensemble of short structural fragments extracted from the PDB.[60] These fragments are then assembled in a Monte Carlo search strategy using a scoring function that favors non-local properties of native protein structures such as hydrophobic burial, compactness, and pairing of β-strands.[22,60,61] Using fragments of known structures ensures that the local interactions are close to optimal, thereby reducing the demand on the free energy function. The Rosetta fragment assembly strategy has been successfully applied to *de novo* structure prediction, as well as to modeling of structurally variable regions (loops, insertions) in comparative protein structure models.

The TASSER (Threading/ASSEmbly/Refinement) method developed by Skolnick, Zhang and coworkers uses tertiary restraints derived from threading results to restrict the conformational search space. The query sequence is first threaded through the structures representative of the PDB to identify appropriate local fragments for further structural reassembly. For a given alignment, an initial full-length model is built by connecting the continuous secondary structure fragments through a random walk, followed by parallel-exchange Monte Carlo sampling for refinement.[62,63]

De novo modeling techniques have made tremendous progress over the last decade, and several individual examples of highly accurate predictions have been reported. However, there are still significant limitations that restrict their application for routine use: the computational demand is immense and therefore limits these methods to relatively small systems. In parallel, the overall quality of the resulting models decreases with the increasing size of the protein. As a result, the accuracy of *de novo* predictions is in general still poor, despite a number of positive examples. In CASP7 (Section 1.5), it was generally not possible to correctly predict the overall fold for a majority of the *de novo* modeling targets.[64]

1.3 Protein Modeling and Structural Genomics

Comparative protein structure modeling and experimental protein structure determination complement each other, with the long-term goal of making three-dimensional atomic-level information of most proteins obtainable from their corresponding amino acid sequences. To achieve structural coverage of a majority of sequenced genes, systematic sampling of major protein families with experimental protein structures is essential (unless the *de novo* methods become perfect). Structural genomics is a worldwide initiative aimed at rapidly determining a large number of protein structures using X-ray crystallography and NMR spectroscopy in a high-throughput mode.[65,66] As a result of concerted efforts in technology and methodology development in recent years, each step of experimental structure determination has become more efficient, less expensive, and more likely to succeed.[67] Structural genomics initiatives are making significant contribution to both the scope and depth of our structural knowledge about protein families. Although worldwide structural genomics initiatives only account for ~20% of the new structures, these contribute approximately to three quarters of the new structurally characterized families and over five times as many novel folds as classical structural biology.[68–73]

Most structural genomics consortia follow specific objectives that include focusing on certain protein classes, such as membrane proteins, protein families with special biomedical relevance, enlarging the coverage of sequence space on the domain level, and determining all the proteins in a model genome. They are applying sophisticated bioinformatics strategies for target selection to maximize the gain in novel insights into protein function from a structural perspective.[68,70,71,74–76]

In the light of the ever-growing amount of genome sequencing data, the structure of most of the proteins, even with structural genomics, will be modeled and not elucidated experimentally. From a modeling-centric perspective, the selection of structural genomics targets should thus be such that most of the remaining sequences can be modeled with useful accuracy by comparative modeling. As discussed before, the accuracy of the comparative models currently declines sharply below the 30% sequence identity. Thus, target selection strategies should aim at systematic sampling of protein structures to ensure that most of the remaining sequences are related to at least one experimentally elucidated structure at more than the 30% sequence identity.[5] Using this cutoff, it has been estimated that a minimum of 16 000 targets must be determined to cover 90% of all the protein domain families, including those of membrane proteins.[77] Such estimates show large variations, depending on the level of sequence identity that is assumed to ensure sufficiently accurate model building, and how this coverage is calculated. Recently, it has been proposed to reduce this number to a manageable size by prioritizing structurally uncharacterized protein families from PFAM according to the number of family-members.[78] However, it has been argued that such coarse-grained target selection is suboptimal in terms of reliable structural and functional annotation, and a selection of “fine-grain” targets from within larger coarse-grained families of distantly related proteins would be required to provide a more thorough coverage of functional space as it relates to protein structure.[68]

Until recently, sequence databases were highly biased towards proteins of known function from a relatively small set of model organisms, a result of targeted protein sequencing. However, in the last decade, whole-genome sequencing efforts have presumably reduced or eliminated this bias. We are, however, on the threshold of a new dimension in sequence diversity. The recent meta-genomics projects (which are based on shotgun sequencing of populations of microorganisms) have yielded new insights into the distribution of (mainly microbial) protein families. As there is an approximately linear relationship between the number of sequence clusters and the number of protein sequences, this indicates that there remain many more protein families to be discovered. This, in turn, has direct implications on the selection of targets for structural genomics.[79]

1.4 Integrative (Hybrid) Modeling Techniques

Biological function is seldom effected by a single protein molecule in isolation. It is most often the result of transient or stable interactions among individual proteins in the cell. Most of these interactions remain uncharacterized by traditional structural biology techniques such as X-ray crystallography (Chapter 22) and NMR spectroscopy (Chapter 24). This gap is being bridged by several emerging experimental approaches that vary in terms of the information they provide.[80] For example, the stoichiometry and composition of protein components in an assembly can be determined by methods such as quantitative immunoblotting and mass spectrometry. The shape of the assembly can be revealed by electron microscopy and small angle X-ray scattering. The positions of the components can be elucidated by cryoelectron microscopy and labeling techniques. Whether or not components interact with each other can be measured by mass spectrometry, yeast two-hybrid and affinity purification. The relative orientations of the components and information about interacting residues can be inferred from cryoelectron microscopy, hydrogen/deuterium exchange, hydroxyl radical footprinting, and chemical-crosslinking.

When the approaches dominated by a single source of information fail, simultaneous consideration of all the available information about the composition and structure of a given assembly, irrespective of its source, can sometimes be sufficient to calculate a useful structural model. Thus, integrative modeling methods convert the experimental data derived from the methods listed above into a structural model of a macromolecular assembly through computation[80] (Fig. 1.1). Such an approach can be used to uncover

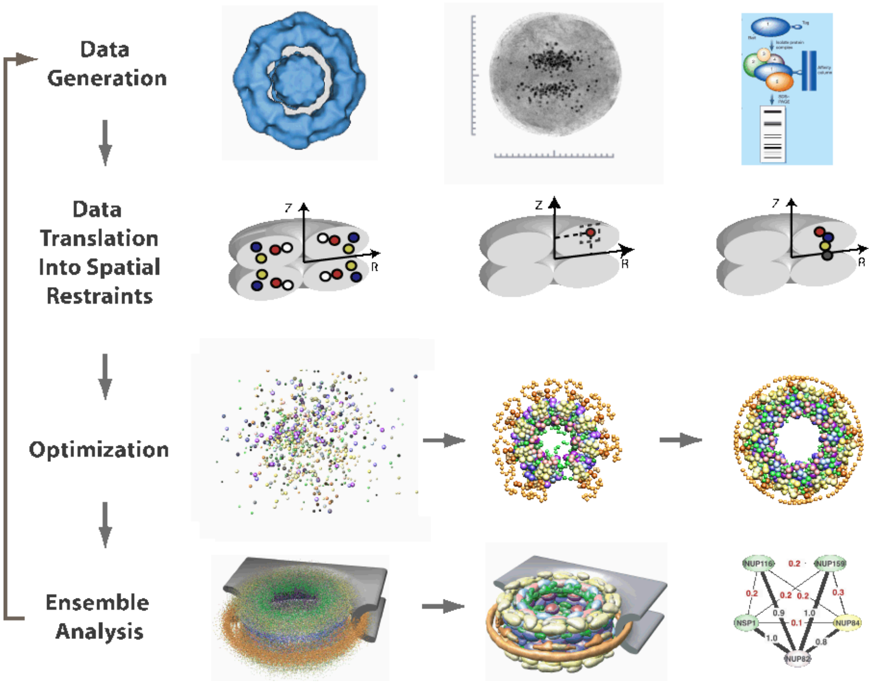


Fig. 1.1 Integrative structure determination. The four steps of determining a structure by integration of varied data are illustrated with the example of the nuclear pore complex.[80,84,132] First, structural data are generated by experiments, such as electron microscopy (*left panel*), immunoelectron microscopy (*middle panel*), and affinity purification of subcomplexes (*right panel*); many other types of information can also be added. Second, the data and theoretical considerations are expressed as spatial restraints ensuring the observed symmetry and shape of the assembly (electron microscopy, *left panel*), positions of constituent gold-labeled proteins (immunoelectron microscopy, *middle panel*), and proximity among the constituent proteins (affinity co-purification, *right panel*). Third, an ensemble of structural solutions that satisfy the data is obtained by minimizing the violations of the spatial restraints (from *left* to *right*). Fourth, the ensemble is clustered into sets of distinct solutions (*left panel*) as well as analyzed in different representations, such as protein positions (*middle panel*) and protein-protein contacts (*right panel*). The integrative approach to structure determination has several advantages: (i) it benefits from the synergy among the input data, minimizing the drawback of incomplete, inaccurate, and/or imprecise data sets (although each individual restraint may contain little structural information, the concurrent satisfaction of all restraints derived from independent experiments may drastically reduce the degeneracy of structural solutions); (ii) it can potentially

the molecular architecture of macromolecular assemblies and even atomic models of protein complexes. Even when this model is of relatively low resolution and accuracy, it can still be helpful for studying the function and evolution of the corresponding assembly; it also provides the necessary starting point for a higher resolution study.

An example of a simple hybrid approach is building a pseudoatomic model of a large assembly by fitting atomic structures of subunits into its cryoelectron microscopy map.[81] Unassigned or partially assigned NMR spectroscopy data and fragment-based modeling approaches have been combined to improve structure refinement in terms of its accuracy, efficiency, and success rate.[82,83] A variety of different types of information, such as symmetry and protein proximity, have been used to characterize large symmetrical assemblies, including the nuclear pore complex,[84,85] EscJ from the type III secretion system,[86] and the AAA+ ring complexes.[87]

1.5 Assessment and Evaluation of Prediction Accuracy

Protein structure modeling is maturing and therefore widely used as a scientific research tool today. Consequently, it is increasingly important to evaluate to what extent the current prediction methods meet the accuracy and requirements of different scientific applications (Chapter 5). A good way to assess the reliability of different protein structure modeling methods *a posteriori* is by evaluating the results of blind predictions after the corresponding protein structures have been determined experimentally. One such effort is the biannual "Community Wide Experiment on the Critical Assessment of

produce all structures that are consistent with the data, not just one; (iii) the variation among the structures consistent with the data allows us to assess the sufficiency of the data and the precision of the representative structure; (iv) it can make the process of structure determination more efficient by indicating what measurements would be the most informative. (This figure was reproduced from Fig. 5 in Ref. 80).

Techniques for Protein Structure Prediction" (CASP).[88,89] During a CASP trial, research groups apply their prediction methods to sequences for which the experimental structure is about to be determined. The accuracy of these blind predictions is then assessed independently once the structures are made available. There are also web servers, LIVEBENCH[90] and EVA,[91] that assess protein structure prediction servers on an automated and continuous basis using sequences from the PDB, before their structures are released, as modeling targets.

1.5.1 *Critical Assessment of Techniques for Protein Structure Prediction (CASP)*

The biannual CASP experiments aim to assess the progress of protein structure prediction methods.[88,92] Besides using classical measures for assessing the accuracy of the Cα positions of the models, several additional criteria were introduced in CASP7 to ensure that the assessment appraises the overall quality of the models, as well as those features of the predictions that are relevant to their usefulness in specific scientific applications, such as the fraction of correctly modeled hydrogen bond interactions (HBscore), the suitability of models for phasing X-ray diffraction data, assessment of the accuracy of predicted cofactor binding sites, and accuracy of the model error estimates provided by the predictors.

In the latest edition of CASP (round 7 in 2006),[39,64,89,93] the general trends observed in the previous years continued: comparative modeling remained by far the most accurate technique for protein structure modeling. However, the majority of predictions submitted in the category of template-based modeling (TBM) were again closer to the template than to the real structure, and only in a few cases, some improvement over a model based on a single best template structure was observed. The fact that no group would outperform a virtual predictor submitting models based on the single best template for each target indicates that template identification and alignment are by no means solved problems and constitute a major bottleneck, besides the challenging question of model refinement. Impressively, successful refinement of model coordinates to a value closer to the experimental structure has been observed, at least in a small number of cases.[22,94]

One of the most remarkable results of CASP7 was that automated prediction servers have matured significantly in the recent years: six of the top 25 groups in the assessment of template-based models were predictors using automated prediction servers, which produce their models without manual intervention. In 29% of a total of 108 cases, the best model for an individual prediction target was submitted by a server. The best prediction server[63] was ranked third over all, i.e. it outperformed all but two of the participating groups.[93,94]

1.5.2 *EVA-CM —— Continuous Automated Assessment of Prediction Servers*

The goal of EVA[91] is to evaluate the sustained performance of protein structure prediction servers through objective measures for prediction accuracy in a fully automated manner. Every week, test sequences are automatically submitted to prediction servers and the results are evaluated and posted on the EVA web sites, thereby providing a continuous, fully automatic and statistically significant analysis of structure prediction servers. Besides comparative modeling, EVA assesses the prediction of secondary structure, inter-residue distances and contacts, and threading.

1.5.3 *Model Quality Evaluation*

Retrospective assessment of the average accuracy of individual modeling methods via projects such as CASP or EVA is invaluable for the development of modeling techniques, but unfortunately does not allow drawing of any conclusions about the accuracy of a specific model, as the correct answer is unknown in a real-life situation. Since the usefulness of predictions crucially depends on their accuracy, a means of reliably predicting the likely accuracy of a protein structure model in the absence of its known 3D structure is an important problem in protein structure prediction (Section 1.2.1.4). Accurate estimates of the errors in a model are an essential component of any predictive method — protein structure prediction not being an exception.

Different scoring schemes have been developed to determine whether or not a model has the correct fold, to differentiate between the native and near-native states, to select the most near-native model in a set of decoys, and to provide quantitative estimates for the coordinate error of the predicted amino acids (Section 1.2.1.4). A variety of methods have been applied to address these tasks, such as physics-based energies, knowledge-based potentials (Chapter 3), combined scoring functions, and clustering approaches. Combined scoring functions integrate several different scores, aiming to extract the most informative features from each of the individual input scores (Chapter 4). Clustering approaches use consensus information from an ensemble of protein structure models provided by different methods.

1.6 Application of Protein Models

1.6.1 *Typical Applications of Protein Models*

The suitability of protein models for specific applications crucially depends on their accuracy. There is a wide range of applications for comparative models, such as designing experiments for site-directed mutagenesis or protein engineering, predicting ligand binding sites and docking small molecules in structure-based drug discovery,[95,96] studying the effect of mutations and SNPs,[97,98] phasing X-ray diffraction data in molecular replacement,[26,99] as well as protein engineering and design.[100] See Chapter 5 for a more detailed discussion about applications of models.

Although the target-template sequence identity generally correlates well with the overall model accuracy, it is often not suitable for making decisions about the usability of models for specific applications. There is a need for new measures to come up with more reliable estimates of model quality. For instance, applications in drug design require a very high accuracy of the local sidechain positions in the binding site, much more so that the overall global accuracy of the backbone.[94,101] Local estimates of the expected model accuracy on a per residue or per atom level would be crucial for many applications, e.g. phasing of crystallographic diffraction data.[39]

1.6.2 *Modeling GPCRs*

Modeling G-protein-coupled receptors has drawn much attention due to their relevance as drug targets. Constraints-based and homology modeling[102–104] has been used as a tool to obtain structural models for GPCRs, at first based on the structures of bacteriorhodopsin,[105,106] and since 2000 using the high resolution X-ray structure of bovine rhodopsin[107] as a template for modeling.[108,109] Only recently the first structure of a GPCR bound to a diffusible ligand, the human β_2-adrenergic G-protein coupled receptor,[110,111] has become available and may now serve as a more suitable template for modeling other members of the class A GPCRs. However, the level of sequence identity within the members of the class A GPCRs is often very low, seriously limiting the accuracy of the local alignment. Especially the conformations of non-conserved inter-helical loops are difficult to model using comparative techniques. Retrospectively, we can analyze the accuracy of the "historic" comparative models built for the human β_2-adrenergic receptor based on the rhodopsin structure as templates. While the overall arrangement of the 7 trans-membrane helix segments is generally correctly represented, significant differences are observed in the relative orientation and shifts of the helices with regard to the center of the receptor (Fig. 1.2).

The ligand-binding pocket is, with regard to rhodopsin, formed by both the structurally conserved and divergent segments. Most deviations are observed for helices III, V, and the extracellular loop ECL2, which connects helices IV and V (Fig. 1.2). While ECL2 is forming a β-sheet structure in rhodopsin, in β_2-adrenergic receptor it contains an unexpected additional α-helical segment and a second disulfide bridge that might stabilize the more solvent exposed conformation. Consequently, specific interactions between the ligand molecule and side chains forming the binding pocket are only partially reproduced by a comparative model based on rhodopsin (Fig. 1.3). See Refs. 110, 111 for a detailed discussion of the individual structural differences, as well as discussion of the activation mechanism.

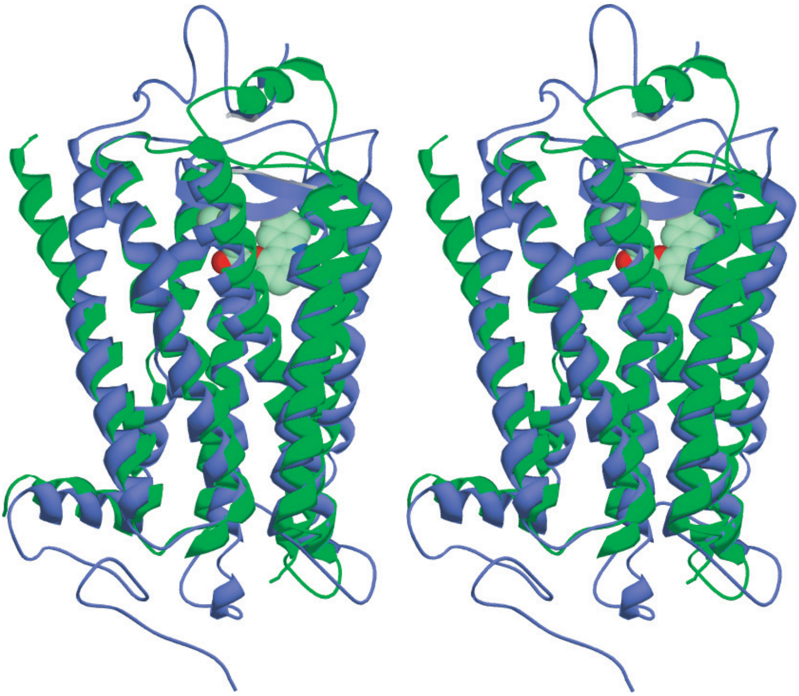


Fig. 1.2 Ribbon representation of the human β_2-adrenergic G-protein coupled receptor with bound ligand carazolol (green, PDB: 2rh1[110]) and the bovine rhodopsin (blue, PDB: 1u19[107]). Bovine rhodopsin has been the only available high resolution template for modeling class A GPCRs until the structure of β_2-adrenergic receptor has been solved in 2007. (Superposition, stereo view).

1.7 Major Protein Modeling Resources

1.7.1 *Protein Modeling Servers and Software Tools*

The huge and constantly growing number of structurally uncharacterized protein sequences, together with the increasing number of available template structures requires the development of automated, stable and reliable modeling methods. Modeling of protein structures usually requires expertise in structural biology and the use of highly specialized computer programs for each of the individual steps of the modeling process. Therefore, automated modeling pipelines

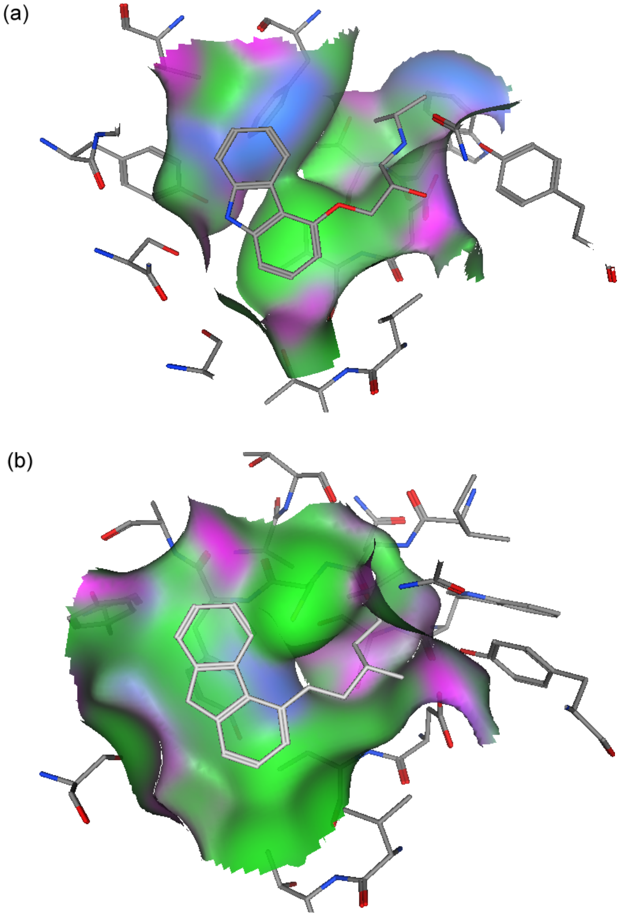


Fig. 1.3 The ligand binding site of the β_2-adrenergic G-protein coupled receptor. The experimentally elucidated structure in panel **(a)** (PDB: 2rh1[110]) as compared to the comparative model based on bovine rhodopsin as template in planel **(b)** (PDB: 1u19[107]).

with integrated expert knowledge such as SWISS-MODEL[14,112–114] and MODPIPE[15,115] were established 15 years ago and have been successfully applied to large data sets.[3,116–120]

Today, there is a plethora of modeling services available on the Internet. Therefore, the question is what is the most appropriate method for a specific target? Meta-servers — methods that use the

Table 1.2. List of Protein Modeling Servers and Software. For a more exhaustive list, see Refs. 93 and 122

	Modeling Server
SwissModel[112–114,124]	http://swissmodel.expasy.org
ModWeb[115]	http://salilab.org/modweb/
I-Tasser[63]	http://zhang.bioinformatics.ku.edu/I-TASSER/
Robetta[133]	http://robetta.bakerlab.org
	Software Tools
HHPred[134]	http://toolkit.tuebingen.mpg.de/hhpred
Modeller[15,115]	http://salilab.org/modeller/
SCWRL3[19]	http://dunbrack.fccc.edu/SCWRL3.php
WhatIf[135]	http://swift.cmbi.ru.nl/whatif/
Rosetta[60]	http://www.rosettacommons.org

results of other servers as input to generate their predictions — are aiming to address this question.[90,121] The general opinion in the community has been that the models generated using a combination of automated predictions and human expertise are superior to those generated using purely automated servers.[90] However, it appears that this view might have to be revised in the near future as the gap between human predictors and servers is closing. Table 1.2 provides examples of the major available resources; see Refs. 93, 122 for a more comprehensive list.

1.7.2 *Protein Model Databases*

Depositions to the PDB are restricted to atomic coordinates that are substantially determined by experimental measurements on specimens containing biological macromolecules.[123]

Currently, the PDB holds approximately 50 000 entries representing 17 000 different proteins. Using these experimentally elucidated structures as templates, several millions of comparative protein models have been generated for the protein sequences contained in the UniProtKB database.[3,4,124,125] Databases of annotated comparative models increase the efficiency for expert users, allow cross-referencing with other (non-structure-centric) resources, and make

Table 1.3. Databases of Automated Comparative Protein Models

Model Database Resources		Refs.
MODBASE	http://www.salilab.org/modbase/	125, 126
SWISS-MODEL Repository	http://swissmodel.expasy.org/repository/	117, 120, 124
Protein Model Portal	http://www.proteinmodelportal.org	

Table 1.4 Protease Models for Entries referenced in the MEROPS Database available in the Protein Model Portal

Group	Number of UniProtKB Entries	Number of Models	Average Sequence Identity with Best Template
Grand Total	6869	28701	39.0%
SWISS-MODEL Repository	3362	5440	69.9%
MODBASE	5001	21471	33.2%
CSMP (Center for Structures of Membrane Proteins)	7	17	19.9%
MCSG (Midwest Center for Structural Genomics)	48	48	28.2%
NESG (Northeast Center for Structural Genomics)	199	244	17.7%
NYSGXRC (New York SGX Center for Structural Genomics)	748	1481	16.9%
PDB[a]	400	2338	N.A.
Protease sequences without structure or model	1342	0	N.A.

[a] Experimentally elucidated protease structures. N.A., not applicable.

comparative models accessible to non-experts. Many specialized efforts exist for specific protein families, or specific organisms. These resources are often manually curated, which poses challenges in terms of maintaining a reasonable update frequency when new template structures and new or updated sequence information become available. Generic model databases such as MODBASE[125,126] and the

SWISS-MODEL Repository[120,124] apply entirely automated techniques for large-scale comparative protein structure modeling.

The Protein Model Portal (http://www.proteinmodelportal.org) has recently been developed as part of the PSI Structural Genomics Knowledge Base to provide an integrated access to the various databases containing structural information and thereby implementing the first step of the community workshop recommendation[123] on archiving structural models of biological macromolecules. Currently, automatically-derived models from six structural genomics centers, MODBASE and SWISS-MODEL Repository are accessible through a single search interface. As an example, we have analyzed all the protease families referenced in the MEROPS database[127] for the number of protein models — and their average sequence identity to the best modeling template — currently available from the Protein Model Portal. It is interesting to note that even in this highly studied class of proteins, there is no structural information available, experimental or modeled, for approximately 20% of the sequences in UniProtKB.

1.8 Future Outlook

1.8.1 *Model Refinement*

Comparative protein structure modeling has matured over the last decade and is now routinely used in many practical applications. There has been a continuous increase in the overall accuracy of protein structure models due to progress in the quality of the sequence-structure alignments as well as the increased availability of high quality template structures. However, comparatively little progress has been made in refining the initial models away from the template closer to the target structure. Model refinement is particularly relevant for models based on alignments with a sequence identity below 30%, which is the typical situation in comparative modeling. Many biomedical applications (Section 1.6) are critically dependent on model accuracy, and the accuracy achieved by comparative modeling based on low sequence identity templates is often insufficient. Improving the accuracy of comparative models beyond the information derived from the template therefore continues to be one of the key questions in the future. Although examples of successful model refinement using molecular dynamics methods have been described occasionally, these methods do not seem to be generally successful.[25,128] The challenges with refinement seem to reside in the limitations of the currently available force fields (which do not accurately represent the energetic interactions of the native state of the protein structure), as well as in the computational effort required for sampling a highly dimensional and rugged energy landscape, which is necessary to identify the global minimum.[22,23,26,129]

1.8.2 *Integrative (Hybrid) Modeling*

Cryoelectron microscopy is emerging as a key technique for studying 3D structures of multi-component macromolecular complexes with masses >250 kDa, such as membrane proteins, cytoskeletal complexes, ribosomes, quasi spherical viruses, molecular chaperones, flagella, ion channels, and oligomeric enzymes. Electron cryotomography even enables the observation of macromolecules inside a living cell in its native state.[130] Various modeling approaches are being developed that utilize cryoelectron microscopy density maps as a constraint in deriving a pseudo-atomic model of the molecular components within a larger complex. Because of the significant likelihood of conformational differences between isolated domains and biological assemblies, additional research leading to the development of reliable hybrid modeling methods, which are able to correctly include structural information from various experimental sources of different resolution and reliability, is essential. The important structural information from hybrid models, generating a synoptic image of the heterogeneous information available for a given macromolecular system, is expected to increase sharply in the coming years. Naturally, this raises the question of whether it will be feasible at one point to combine all these data, together with other data related to the overall cellular structure, to construct a quantitative spatial and temporal model of the cell.[131]

References

1. Berman HM, Westbrook J, Feng Z, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**(1): 235–242.
2. Bairoch A, Apweiler R, Wu CH, *et al.* (2005) The Universal Protein Resource (UniProt). *Nucl Acids Res* **33**(Database Issue): D154–159.
3. Peitsch MC. (1997) Large scale protein modelling and model repository. *Proc Int Conf Intell Syst Mol Biol* **5**: 234–236.
4. Peitsch MC, Schwede T, Guex N. (2000) Automated protein modelling — the proteome in 3D. *Pharmacogenomics* **1**(3): 257–266.
5. Baker D, Sali A. (2001) Protein structure prediction and structural genomics. *Science* **294**(5540): 93–96.
6. Melo F, Sali A. (2007) Fold assessment for comparative protein structure modeling. *Protein Sci* **16**(11): 2412–2426.
7. Altschul SF, Madden TL, Schaffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl Acids Res* **25**(17): 3389–3402.
8. Karplus K, Barrett C, Hughey R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**(10): 846–856.
9. Eddy SR. (1998) Profile hidden Markov models. *Bioinformatics* **14**(9): 755–763.
10. Jaroszewski L, Rychlewski L, Li Z, *et al.* (2005) FFAS03: a server for profile – profile sequence alignments. *Nucl Acids Res* **33**(Web Server Issue): W284–288.
11. Marti-Renom MA, Madhusudhan MS, Sali A. (2004) Alignment of protein sequences by their profiles. *Protein Sci* **13**(4): 1071–1087.
12. Soding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**(7): 951–960.
13. Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM. (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326**(6111): 347–352.
14. Peitsch MC, Jongeneel CV. (1993) A 3-D model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. *Int Immunol* **5**(2): 233–238.
15. Sali A, Blundell TL. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**(3): 779–815.
16. Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B. (2007) Loop modeling: sampling, filtering, and scoring. *Proteins.*
17. Jacobson MP, Pincus DL, Rapp CS, *et al.* (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins* **55**(2): 351–367.
18. Lovell SC, Word JM, Richardson JS, Richardson DC. (2000) The penultimate rotamer library. *Proteins* **40**(3): 389–408.
19. Canutescu AA, Shelenkov AA, Dunbrack RL, Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* **12**(9): 2001–2014.

20. Brooks BR, Bruccoleri RE, Olafson BD, *et al.* (1983) CHARMM: a program for macromolecular energy, minmimization, and dynamics calculations. *J Comput Chem* **4**: 187–217.
21. Cornell WD, Cieplak P, Bayly CI, *et al.* (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* **117**: 5179–5197.
22. Das R, Qian B, Raman S, *et al.* (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69**(S8): 118–128.
23. Han R, Leo-Macias A, Zerbino D, *et al.* (2007) An efficient conformational sampling method for homology modeling. *Proteins* 10.1002/prot.21672.
24. Qian B, Ortiz AR, Baker D. (2004) Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proc Natl Acad Sci USA* **101**(43): 15346–15351.
25. Chen J, Brooks CL, 3rd. (2007) Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins* **67**(4): 922–930.
26. Qian B, Raman S, Das R, *et al.* (2007) High-resolution structure prediction and the crystallographic phase problem. *Nature.*
27. Domingues FS, Koppensteiner WA, Jaritz M, *et al.* (1999) Sustained performance of knowledge-based potentials in fold recognition. *Proteins* (3): 112–120.
28. McGuffin LJ, Jones DT. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19**(7): 874–881.
29. Miyazawa S, Jernigan RL. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **256**(3): 623–644.
30. Lazaridis T, Karplus M. (1999) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* **288**(3): 477–487.
31. Gatchell DW, Dennis S, Vajda S. (2000) Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* **41**(4): 518–534.
32. Vorobjev YN, Hermans J. (2001) Free energies of protein decoys provide insight into determinants of protein stability. *Protein Sci* **10**(12): 2498–2506.
33. Seok C, Rosen JB, Chodera JD, Dill KA. (2003) MOPED: Method for optimizing physical energy parameters using decoys. *J Comput Chem* **24**(1): 89–97.
34. Tsai J, Bonneau R, Morozov AV, *et al.* (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* **53**(1): 76–87.
35. Zhu J, Zhu Q, Shi Y, Liu H. (2003) How well can we predict native contacts in proteins based on decoy structures and their energies? *Proteins* **52**(4): 598–608.
36. Eramian D, Shen MY, Devos D, *et al.* (2006) A composite score for predicting errors in protein structure models. *Protein Sci* **15**(7): 1653–1666.

37. Shen MY, Sali A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**(11): 2507–2524.
38. Chothia C, Lesk AM. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**(4): 823–826.
39. Read RJ, Chavali G. (2007) Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins* **69**(S8): 27–37.
40. Orengo CA, Thornton JM. (2005) Protein families and their evolution-a structural perspective. *Ann Rev Biochem* **74**: 867–900.
41. Zhang Y, Skolnick J. (2005) The protein structure prediction problem could be solved using the current PDB library. *Proc Natl Acad Sci USA* **102**(4): 1029–1034.
42. Rost B. (2002) Enzyme function less conserved than anticipated. *J Mol Biol* **318**(2): 595–608.
43. Tian W, Skolnick J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* **333**(4): 863–882.
44. Dyson HJ, Wright PE. (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**(3): 197–208.
45. Radivojac P, Iakoucheva LM, Oldfield CJ, *et al.* (2007) Intrinsic disorder and functional proteomics. *Biophys J* **92**(5): 1439–1456.
46. Fink AL. (2005) Natively unfolded proteins. *Curr Opin Struct Biol* **15**(1): 35–41.
47. Namba K. (2001) Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells* **6**(1): 1–12.
48. Tompa P, Csermely P. (2004) The role of structural disorder in the function of RNA and protein chaperones. *FASEB J* **18**(11): 1169–1175.
49. Bordoli L, Kiefer F, Schwede T. (2007) Assessment of disorder predictions in CASP7. *Proteins* **69**(S8): 129–136.
50. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* **61**(7): 176–182.
51. Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**(13): 2138–2139.
52. Schlessinger A, Liu J, Rost B. (2007) Natively unstructured loops differ from other loops. *PLoS Comput Biol* **3**(7): e140.
53. Barth P, Schonbrun J, Baker D. (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc Natl Acad Sci USA* **104**(40): 15682–15687.
54. Zhang Y, Devries ME, Skolnick J. (2006) Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Comput Biol* **2**(2): e13.

55. Yarov-Yarovoy V, Schonbrun J, Baker D. (2006) Multipass membrane protein structure prediction using Rosetta. *Proteins* **62**(4): 1010–1025.
56. Jayachandran G, Vishal V, Garcia AE, Pande VS. (2007) Local structure formation in simulations of two small proteins. *J Struct Biol* **157**(3): 491–499.
57. Muff S, Caflisch A. (2007) Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. *Proteins.* 10.1002/prot.21565.
58. Verma A, Wenzel W. (2007) Protein structure prediction by all-atom free-energy refinement. *BMC Struct Biol* 7: 12.
59. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. (2007) The protein folding problem: when will it be solved? *Curr Opin Struct Biol* **17**(3): 342–346.
60. Rohl CA, Strauss CE, Misura KM, Baker D. (2004) Protein structure prediction using Rosetta. *Meth Enzymol* **383**: 66–93.
61. Rohl CA, Strauss CE, Chivian D, Baker D. (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55**(3): 656–677.
62. Wu S, Skolnick J, Zhang Y. (2007) *Ab initio* modeling of small proteins by iterative TASSER simulations. *BMC Biol* **5**: 17.
63. Zhang Y. (2007) Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69**(S8): 108–117.
64. Jauch R, Yeo HC, Kolatkar PR, Clarke ND. (2007) Assessment of CASP7 structure predictions for template free targets. *Proteins* **69**(S8): 57–67.
65. Burley SK. (2000) An overview of structural genomics. *Nat Struct Biol* 7(Suppl): 932–934.
66. Thornton J. (2001) Structural genomics takes off. *Trends Biochem Sci* **26**(2): 88–89.
67. Slabinski L, Jaroszewski L, Rodrigues AP, *et al.* (2007) The challenge of protein structure determination — lessons from structural genomics. *Protein Sci* **16**(11): 2472–2482.
68. Marsden RL, Lewis TA, Orengo CA. (2007) Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinform* **8**: 86.
69. Todd AE, Marsden RL, Thornton JM, Orengo CA. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol* **348**(5): 1235–1260.
70. Chandonia JM, Brenner SE. (2006) The impact of structural genomics: expectations and outcomes. *Science* **311**(5759): 347–351.
71. Liu J, Montelione GT, Rost B. (2007) Novel leverage of structural genomics. *Nat Biotechnol* **25**(8): 849–851.

72. Gileadi O, Knapp S, Lee WH, *et al.* (2007) The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins. *J Struct Funct Genomics* **8**: 107–119.
73. Levitt M. (2007) Growth of novel protein structural data. *Proc Natl Acad Sci USA* **104**(9): 3183–3188.
74. Chen L, Oughtred R, Berman HM, Westbrook J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics* **20**(16): 2860–2862.
75. Liu J, Hegyi H, Acton TB, *et al.* (2004) Automatic target selection for structural genomics on eukaryotes. *Proteins* **56**(2): 188–200.
76. Bussow K, Scheich C, Sievert V, *et al.* (2005) Structural genomics of human proteins — target selection and generation of a public catalogue of expression clones. *Microb Cell Fact* **4**: 21.
77. Vitkup D, Melamud E, Moult J, Sander C. (2001) Completeness in structural genomics. *Nat Struct Biol* **8**(6): 559–566.
78. Chandonia JM, Brenner SE. (2005) Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* **58**(1): 166–179.
79. Yooseph S, Sutton G, Rusch DB, *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**(3): e16.
80. Robinson CV, Sali A, Baumeister W. (2007) Molecular sociology of the cell. *Nature* **450**: 973–982.
81. Topf M, Baker ML, Marti-Renom MA, *et al.* (2006) Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. *J Mol Biol* **357**(5): 1655–1668.
82. Lee SY, Zhang Y, Skolnick J. (2006) TASSER-based refinement of NMR structures. *Proteins* **63**(3): 451–456.
83. Meiler J, Baker D. (2005) The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. *J Magn Reson* **173**(2): 310–316.
84. Alber F, Dokudovskaya S, Veenhoff LM, *et al.* (2007) The molecular architecture of the nuclear pore complex. *Nature* **450**(7170): 695–701.
85. Devos D, Dokudovskaya S, Williams R, *et al.* (2006) Simple fold composition and modular architecture of the nuclear pore complex. *Proc Natl Acad Sci USA* **103**(7): 2172–2177.
86. Andre I, Bradley P, Wang C, Baker D. (2007) Prediction of the structure of symmetrical protein assemblies. *Proc Natl Acad Sci USA* **104**(45): 17656–17661.
87. Diemand AV, Lupas AN. (2006) Modeling AAA+ ring complexes from monomeric structures. *J Struct Biol* **156**(1): 230–243.

88. Moult J. (2005) A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* **15**(3): 285–289.
89. Moult J, Fidelis K, Kryshtafovych A, *et al.* (2007) Critical assessment of methods of protein structure prediction-round VII. *Proteins* **69**(S8): 3–9.
90. Rychlewski L, Fischer D. (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci* **14**(1): 240–245.
91. Koh IY, Eyrich VA, Marti-Renom MA, *et al.* (2003) EVA: evaluation of protein structure prediction servers. *Nucl Acids Res* **31**(13): 3311–3315.
92. Kryshtafovych A, Fidelis K, Moult J. (2007) Progress from CASP6 to CASP7. *Proteins* **69**(S8): 194–207.
93. Battey JN, Kopp J, Bordoli L, *et al.* (2007) Automated server predictions in CASP7. *Proteins* **69**(S8): 68–82.
94. Kopp J, Bordoli L, Battey JND, *et al.* (2007) Assessment of CASP7 predictions for template-based modeling targets. *Proteins: Struct Funct Bioinform* **69**(S8): 38–56.
95. Hillisch A, Pineda LF, Hilgenfeld R. (2004) Utility of homology models in the drug discovery process. *Drug Discov Today* **9**(15): 659–669.
96. Vangrevelinghe E, Zimmermann K, Schoepfer J, *et al.* (2003) Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J Med Chem* **46**(13): 2656–2662.
97. Feyfant E, Sali A, Fiser A. (2007) Modeling mutations in protein structures. *Protein Sci* **16**(9): 2030–2041.
98. Wattenhofer M, Di Iorio MV, Rabionet R, *et al.* (2002) Mutations in the TMPRSS3 gene are a rare cause of childhood nonsyndromic deafness in Caucasian patients. *J Mol Med* **80**(2): 124–131.
99. Raimondo D, Giorgetti A, Giorgetti A, *et al.* (2007) Automatic procedure for using models of proteins in molecular replacement. *Proteins* **66**(3): 689–696.
100. Poole AM, Ranganathan R. (2006) Knowledge-based potentials in protein design. *Curr Opin Struct Biol* **16**(4): 508–513.
101. Thorsteinsdottir HB, Schwede T, Zoete V, Meuwly M. (2006) How inaccuracies in protein structure models affect estimates of protein-ligand interactions: computational analysis of HIV-I protease inhibitor binding. *Proteins* **65**(2): 407–423.
102. Herzyk P, Hubbard RE. (1995) Automated method for modeling seven-helix transmembrane receptors from experimental data. *Biophys J* **69**(6): 2419–2442.
103. Peitsch MC, Herzyk P, Wells TN, Hubbard RE. (1996) Automated modelling of the transmembrane region of G-protein coupled receptor by Swiss-model. *Receptors Channels* **4**(3): 161–164.

104. Dahl SG, Edvardsen O, Sylte I. (1991) Molecular dynamics of dopamine at the D2 receptor. *Proc Natl Acad Sci USA* **88**(18): 8111–8115.
105. Henderson R, Schertler GF. (1990) The structure of bacteriorhodopsin and its relevance to the visual opsins and other seven-helix G-protein coupled receptors. *Philos Trans Roy Soc London B Biol Sci* **326**(1236): 379–389.
106. Pebay-Peyroula E, Rummel G, Rosenbusch JP, Landau EM. (1997) X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science* **277**(5332): 1676–1681.
107. Palczewski K, Kumasaka T, Hori T, *et al.* (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289**(5480): 739–745.
108. Ballesteros J, Palczewski K. (2001) G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin. *Curr Opin Drug Discov Devel* **4**(5): 561–574.
109. Oliveira L, Hulsen T, Lutje Hulsik D, *et al.* (2004) Heavier-than-air flying machines are impossible. *FEBS Lett* **564**(3): 269–273.
110. Cherezov V, Rosenbaum DM, Hanson MA, *et al.* (2007) High-resolution crystal structure of an engineered human {beta}2-adrenergic G protein coupled receptor. *Science* **318**: 1258–1265.
111. Rosenbaum DM, Cherezov V, Hanson MA, *et al.* (2007) GPCR engineering yields high-resolution structural insights into {beta}2 adrenergic receptor function. *Science* **318**: 1266–1273.
112. Peitsch MC. (1995) Protein modelling by E-Mail. *BioTechnology* **13**: 658–660.
113. Guex N, Peitsch MC. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**(15): 2714–2723.
114. Schwede T, Kopp J, Guex N, Peitsch MC. (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucl Acids Res* **31**(13): 3381–3385.
115. Eswar N, John B, Mirkovic N, *et al.* (2003) Tools for comparative protein structure modeling and analysis. *Nucl Acids Res* **31**(13): 3375–3380.
116. Peitsch MC, Tschopp J. (1995) Comparative molecular modelling of the Fas-ligand and other members of the TNF family. *Mol Immunol* **32**(10): 761–772.
117. Peitsch MC, Wilkins MR, Tonella L, *et al.* (1997) Large-scale protein modelling and integration with the SWISS-PROT and SWISS-2DPAGE databases: the example of *Escherichia coli*. *Electrophoresis* **18**(3–4): 498–501.
118. Sanchez R, Sali A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc Natl Acad Sci USA* **95**(23): 13597–13602.
119. Sanchez R, Pieper U, Mirkovic N, *et al.* (2000) MODBASE, a database of annotated comparative protein structure models. *Nucl Acids Res* **28**(1): 250–253.

120. Kopp J, Schwede T. (2006) The SWISS-MODEL Repository: new features and functionalities. *Nucl Acids Res* **34**(Database Issue): D315–318.
121. Wallner B, Larsson P, Elofsson A. (2007) Pcons.net: protein structure prediction meta server. *Nucl Acids Res* **35**(Web Server Issue): W369–W374.
122. Fox JA, McMillan S, Ouellette BF. (2006) A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucl Acids Res* **34**(Web Server Issue): W3–W5.
123. Berman HM, Burley SK, Chiu W, J *et al.* (2006) Outcome of a workshop on archiving structural models of biological macromolecules. *Structure* **14**(8): 1211–1217.
124. Kopp J, Schwede T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucl Acids Res* **32**(Database Issue): D230–D234.
125. Pieper U, Eswar N, Davis FP, *et al.* (2006) MODBASE: A database of annotated comparative protein structure models and associated resources. *Nucl Acids Res* **34**(Database Issue): D291–D295.
126. Sanchez R, Sali A. (1999) ModBase: A database of comparative protein structure models. *Bioinformatics* **15**(12): 1060–1061.
127. Rawlings ND, Morton FR, Barrett AJ. (2006) MEROPS: the peptidase database. *Nucl Acids Res* **34** (Database Issue): D270–D272.
128. Krieger E, Koraimann G, Vriend G. (2002) Increasing the precision of comparative models with YASARA NOVA — a self-parameterizing force field. *Proteins* **47**(3): 393–402.
129. Misura KM, Baker D. (2005) Progress and challenges in high-resolution refinement of protein structure models. *Proteins* **59**(1): 15–29.
130. Baumeister W. (2004) Mapping molecular landscapes inside cells. *Biol Chem* **385**(10): 865–872.
131. Betts MJ, Russell RB. (2007) The hard cell: from proteomics to a whole cell model. *FEBS Lett* **581**(15): 2870–2876.
132. Alber F, Dokudovskaya S, Veenhoff LM, *et al.* (2007) Determining the architectures of macromolecular assemblies. *Nature* **450**(7170): 683–694.
133. Kim DE, Chivian D, Baker D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucl Acids Res* **32**(Web Server Issue): W526–W531.
134. Soding J, Biegert A, Lupas AN. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucl Acids Res* **33** (Web Server Issue): W244–W248.
135. Vriend G. (1990) WHAT IF: a molecular modeling and drug design program. *J Mol Graph* **8**(1): 52–56, 29.

This page intentionally left blank

Chapter 2

Protein Fold Recognition and Threading

L. J. McGuffin*

2.1 Introduction

Fold recognition and threading methods can be used to assign tertiary structures to protein sequences, even in the absence of clear homology. The ongoing development of such methods has had a significant impact on structural biology, providing us with an increasing ability to accurately model 3D protein structures using very evolutionary distant fold templates.

Although fold recognition and threading techniques will not yield equivalent results as those from X-ray crystallography, they are a comparatively fast and inexpensive way to a build a close approximation of a structure from a sequence, without the time and costs of experimental procedures. Using fold recognition we are able to identify proteins with known structures that share common folds with the target sequences. The identified structures can then be used as templates from which the folds of the target sequences are modeled.

For the vast majority of new protein sequences, there will be a structure with a similar fold within the Protein Data Bank (PDB)[1] (see Table 2.1 for a list of relevant Internet resources) from which a suitable model could be constructed.[2] Indeed, in the mid-1990s it was

*The School of Biological Sciences, The University of Reading, Whiteknights, PO Box 221, Reading, RG6 6AS, UK. Email: l.j.mcguffin@reading.ac.uk

Table 2.1 A List of URLs to Some of the Currently Available Web Servers, Databases and Software Resources

Name of Resource	Description	URL
RCSB Protein Data Bank (PDB)	Information portal to biological macromolecular structures (1).	http://www.rcsb.org/
NCBI toolkit	Software development toolkit containing programs such as the sequence-profile PSIBLAST (18) method from National Center for Biotechnology Information.	ftp://ftp.ncbi.nih.gov/toolbox/
UCSC HMM Applications	The SAM (19) profile-HMM servers.	http://www.soe.ucsc.edu/compbio/HMM-apps/
FFAS	The Fold & Function Assignment System (21) profile-profile server.	http://ffas.ljcrf.edu/
HHpred	Server for homology detection and structure prediction by HMM-HMM comparison (23)	http://toolkit.tuebingen.mpg.de/hhpred
THREADER Info Page	Links to download THREADER (5) software and fold library.	http://bioinf.cs.ucl.ac.uk/threader/
PSIPRED	Hybrid methods such as GenTHREADER (33) and mGenTHREADER (40, 57).	http://bioinf.cs.ucl.ac.uk/psipred/
Inub	The INBGU (35) hybrid method.	http://inub.cse.buffalo.edu/
FUGUE	The FUGUE (39) hybrid method.	http://tardis.nibio.go.jp/fugue/
I-TASSER	The iterative version of the TASSER (42) hybrid method.	http://zhang.bioinformatics.ku.edu/I-TASSER/

(*Continued*)

Table 2.1 (*Continued*)

Name of Resource	Description	URL
nFOLD2	The nFOLD (41) hybrid method.	http://www.biocentre.rdg.ac.uk/bioinformatics/nFOLD/
Pcons meta-server	The Pcons (43) meta-server for consensus fold recognition predictions.	http://www.bioinfo.se/pcons/
BioInfobank meta-server	The 3D-Jury (44) meta-server	http://meta.bioinfo.pl/
Robetta meta-server	The Robetta (47) meta-server for full-chain protein structure prediction.	http://www.robetta.org/
Protein Structure Prediction Centre	Information regarding the series of CASP assessments.	http://predictioncenter.org/
LiveBench, Continuous Benchmarking of Structure Prediction Servers	Information regarding the series of LiveBench assessments.	http://meta.bioinfo.pl/livebench.pl
e-Protein	e-Protein project home page: links to structural annotation databases such as 3D-GENOMICS (53), Gene3D (54) and the Genomic Threading Database (56).	http://www.e-protein.org/

discovered by Orengo *et al.*, that just nine different folds accounted for up to 30% of the known structures.[3] More recently, a study by Zhang and Skolnick has argued that most sequences can now be modeled using known folds (or fragments of known folds) as templates.[4] However, it is often the case that for many target sequences templates cannot be found using simple sequence searching alone, due to the low sequence identity to any known structure.

Fold recognition and threading methods aim to assign folds to target sequences that have very low sequence identity to known structures. The original concept of early threading methods was to turn the problem of comparative modeling upside down. In other words, the aim was to calculate how well each potential structure would fit a sequence, rather than how well each sequence fits a structure. In simple terms, fold recognition methods work by comparing each target sequence against a library of potential fold templates using energy potentials and/or other similarity scoring methods. The template with the lowest energy score (or highest similarity score) is then assumed to best fit the fold of the target protein.

So what is the difference between fold recognition and threading? The term "threading" was a neologism coined in the early 1990s by Jones, Taylor and Thornton in order to describe a novel approach to protein fold recognition.[5] Jones *et al.* developed the first true threading method, THREADER, which used the technique of double dynamic programming similar to that of Taylor & Orengo,[6] in order to optimally fit (or "thread") a sequence on to the backbone coordinates of known protein structures. The best fitting models were determined using energy potentials derived from the statistical analysis of known structures. Threading became one of the most successful approaches to fold recognition during the 1990s.

The popularity of the method meant that "threading" became a generic term to describe carrying out protein fold recognition (such as "googling" has become the generic term used to describe web searching) and was often used to differentiate structure-based methods for tertiary structure prediction from sequence-based methods. Technically speaking, "threading" is a specialized sub-class of fold recognition and it is now beginning to fall out of common usage.

Historically, tertiary structure prediction methods were divided into three categories. The first category included simple sequence-based methods for selecting templates prior to Comparative Modeling (CM). In cases where no sequence homologs could be found, then structure-based methods were used for protein fold recognition (FR). Finally, *ab initio* or "new fold" (NF) methods were used where there were no structural templates available. However, in recent years the traditional boundaries have become blurred and the distinction between individual methods has become less clear. Sequence searching has become more powerful and arguably the traditional threading techniques which are based on physical energy potentials are becoming less popular. The term "fold recognition" is now often used to encompass all methods able to carry out template-based modeling beyond the so-called "twilight zone" of sequence identity.

2.2 Sequence-based Fold Recognition Beyond the Twilight Zone

Traditionally, the term comparative modeling (CM) has been used to describe those methods which rely on finding a sequence alignment with relatively high sequence identity (typically >30%) between a target sequence and a template structure. Additionally, the term Fold Recognition (FR) was reserved for methods which did not rely on sequence searching, where the sequence identity between target and template was below the so-called "twilight zone" of between 25–30%.[7] However, ongoing developments in sequence-based methods have allowed accurate fold recognition beyond the traditional sequence identity thresholds.

In 1970, Needleman and Wunsch described one of the first computationally efficient methods for carrying out the optimal *global* alignment of pairs of biological sequences using dynamic programming.[8] The accuracy of pairwise sequence alignments was further improved by Smith and Waterman in 1981, who modified the dynamic programming scoring matrix in order to calculate the optimal *local* alignments.[9] While the dynamic programming approaches were far more computationally efficient than exhaustively searching for the best alignment, as the sequence databases began to increase rapidly in size, it became clear that an even more efficient approach was required.

The FASTA method[10] and BLAST (Basic Local Alignment Search Tool) method[11] were developed in order to perform rapid pairwise searches for homologous sequences within the vast sequence databases. Rather than carrying out optimal alignments on whole sequences, these methods worked by quickly finding matching sub-sequences or "tuples" shared between the target protein sequences and the sequences within the databases.

Sequence searching using pairwise methods was also greatly improved through the use of amino acid comparison matrices. These matrices were developed in order to score the alignment of different pairs of amino acids with different weightings. Different weightings were used to account for the different physical, chemical or structural properties shared by each pair of amino acids, e.g. a leucine-isoleucine match would be scored higher than a leucine-tryptophan. Many sets of matrices have been derived over the years such as the PAM,[12] GCB[13] and JTT[14] matrices, but perhaps the most commonly used set is the BLOSUM set.[15]

A benchmarking study by Brenner *et al.*[16] concluded that for sequences with identities >30%, rapid sequence searching methods such as FASTA and WU-BLAST[17] were found to be comparable in accuracy to the Smith-Waterman[9] based method, SSEARCH.[10] However, when the sequence identity was found to falls below 30%, then conventional pairwise sequence comparison methods were unable to detect relationships.[16]

A major step forward in the ability of sequence searching methods to rapidly detect more distant homologs was made possible through the development of sequence-profile methods such as PSI-BLAST[18] and SAM-T98.[19] Comparisons of sequences against profiles, derived from multiple aligned sequences, allow for the detection of more distant evolutionary relationships than can be achieved using pairwise methods.

The popularity of the PSI-BLAST (Position Specific Iterative — BLAST[18]) method has meant that it has become the universal benchmark against which newly developed sequence searching methods are often compared. The method works by carrying out iterative searches for a target protein on a dataset of sequences using position specific score matrices derived from BLAST profiles. A benchmarking study carried out by Müller *et al.*[20] found that the PSI-BLAST method was able to accurately recognize homologues for 40% of the domains with <20% identity within a model proteome. In addition, a study by Park *et al.*[19] revealed that up to three times as many remote homologues could be detected using profile methods such as PSI-BLAST and their profile hidden Markov model (profile-HMM) based-method, SAM-T98, than could be found using the pairwise methods.

The next landmark in sequence-based fold recognition was the development of profile-profile based methods. In 2000, Rychlewski *et al.* developed the FFAS (Fold and Function Assignment System) method.[21] The FFAS method differed fundamentally from sequence-profile methods such as PSI-BLAST in that it used profiles for both the target and template sequences. In other words, a profile was generated for the target sequence which was then aligned to the template profiles of proteins with known structures. Profile-profile alignment methods have proved to be another major step forward in sequence-based fold recognition and many of the current top performing automated methods now adopt the approach. Ohlson *et al.* have carried out a benchmarking study of some of the best early profile-profile methods.[22] More recently, a new approach using profile-HMM–profile-HMM comparison (HHpred) has been developed which further extends the accuracy of sequence-based fold recognition.[23]

Using profile-based methods allows for the detection of templates beyond the twilight zone threshold of 25–30% sequence identity. Purely sequence-based methods can be used for recognizing the folds of target sequences with very remote but common ancestry, i.e. distant homologues with similar function. Such targets were designated as Fold Recognition Homologous (FR/H) in the CASP6 experiment (see Section 2.6). However, these methods often did not perform adequately at recognizing the relationship between non-homologous protein targets which have similar folds.[21] Such targets were designated as Fold Recognition Analogous (FR/A) in the CASP6 experiment (see Section 2.6). For detecting the analogous proteins which have similar folds but no sequence detectable common ancestry, using a method which made use of additional structural information was often the only clear option.

2.3 Structure-based Fold Recognition and Optimal Sequence Threading

The first real attempt at developing a method which could recognize the fold of a protein in the absence of sequence homology was carried out by Bowie *et al.*, in 1991.[24] The method built upon the ideas from previous studies by Ponder and Richards[25] and Bowie *et al.*,[26] which attempted to relate sequences to folds at low levels of sequence identity by examining the structural environments of the residues within the sequence. The premise of the method was that the structural environment of the residue was more conserved than the actual type of residue; therefore, in the absence of homology, a fold could be predicted by measuring the compatibility of a sequence with template folds in terms of amino acid preferences for certain structural environments. The amino acid preferences for three main types of structural environment were considered: the solvent accessibility, the contact with polar atoms and the secondary structure. These structural environments were reduced to a 1D string which was then aligned using dynamic programming. Following the development of the method, several wrongly traced X-ray structures were identified in the PDB and were subsequently removed from the following release of the database.[27]

The success of the approach of Bowie *et al.* in 1991, highlighted by the study by Luthy *et al.,*[27] sparked an enthusiasm for the further development of fold recognition methods throughout the following decade. Arguably, the most successful fold recognition method during the 1990s was known as optimal sequence threading, or threading for short, which was pioneered by Jones *et al.* in 1992.[5] The threading method differed fundamentally from the approach taken by Bowie *et al.*, in that it considered the detailed network of interactions between residues rather than just consigning them to an individual structural environment. The success of the threading method was built upon two key factors: the development of energy potentials derived from the statistical analysis of known structures[28] and the double dynamic programming algorithm developed by Taylor and Orengo.[6] The method developed by Jones *et al.* worked by optimally fitting (or "threading") target sequences directly onto the backbone coordinates of fold templates using double dynamic programming and then evaluating the fit of each resulting fold using energy and solvation potentials — the premise being that the structure which resulted in the lowest energy was the best fit for the target sequence. In the very first CASP experiment (see Section 2.6), Jones' threading program THREADER was used to successfully identify the folds of 8 out of 11 target sequences which had no discernable sequence homology to known structures.[29]

Throughout the 1990s, the most accurate methods for protein fold recognition were arguably those which built upon the original threading protocol of Jones *et al.* A number of techniques were developed which mostly used iterative dynamic programming to build proposed models, followed by an analysis of structurally adjacent residues to evaluate each model.[30–32]

Whilst threading was an efficient method for assigning folds to sequences with little homology to known structures, there were a number of drawbacks to this technique. Firstly, the double dynamic programming algorithms underlying threading methods were computationally intensive. This became increasingly problematic due to the growth in both the number of targets and the number of structures which could be used as templates. Secondly, the methods often produced multiple output scores for each target and template which required human expertise for accurate interpretation. Thirdly, the methods often produced poor sequence to structure alignments and were limited to predicting structures with single domain folds. In the early CASP competitions, it was sufficient to predict the fold class of a target protein; however, it soon became clear that the accuracy of sequence to structure alignments must also be evaluated to encourage predictors to provide useful 3D models of proteins to the wider community.

All of these factors limited the usefulness of pure threading methods for large scale fully automated structure prediction.

2.4 Hybrid Methods and Fully Automated Servers

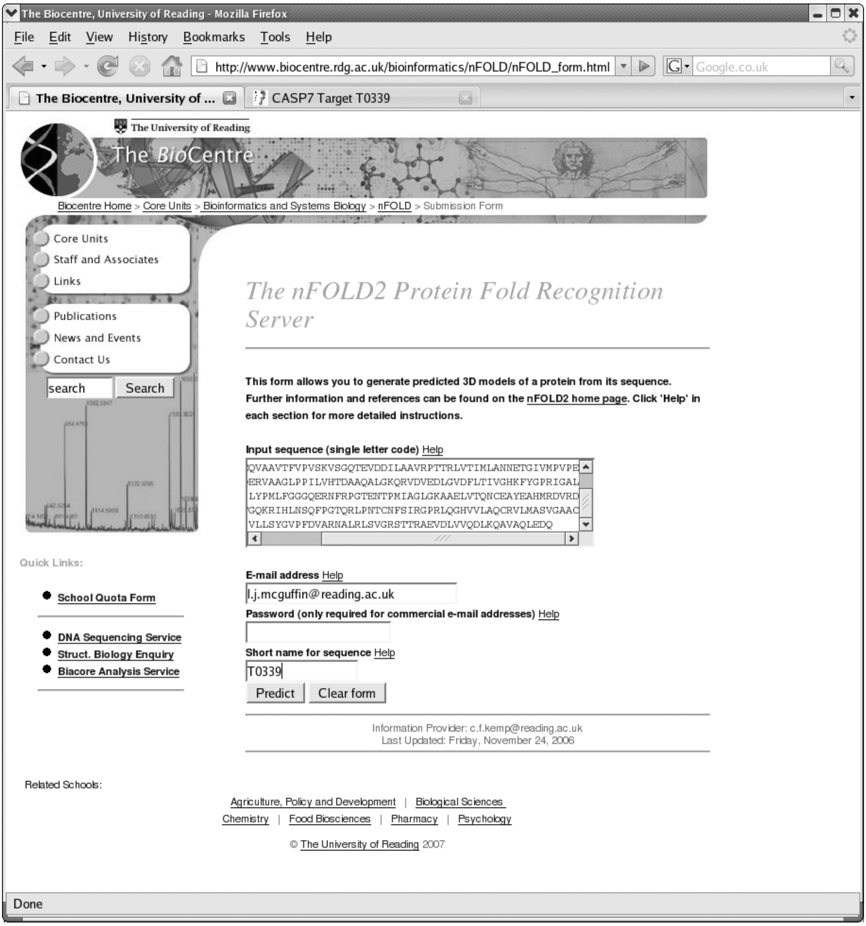
The strengths and limitations of using sequence-based or structure-based fold recognition methods individually led to the development of the so-called hybrid fully automated fold recognition servers. Purely sequence-based methods produced accurate sequence to structure alignments, but were poor at recognizing folds of protein targets with very distant homologues. Conversely, the traditional threading methods were useful for recognizing both distant homologous and analogous folds; however, they were difficult to automate and produced poor models due to inaccurate sequence-structure alignments.

The GenTHREADER method developed by Jones[33] was one of the first methods to combine sequence profile-based searches with energy potentials derived from threading. This hybrid technique was designed in order to perform rapid, fully automated fold recognition on a proteome wide scale. The GenTHREADER protocol initially consisted of sequence-profile based searches against a non-redundant fold library. The resulting sequence-to-structure alignments were then evaluated using the energy potentials from the original THREADER method. The output alignment scores, pairwise energy scores, solvation energy scores and length information were evaluated using an artificial neural network, which was trained to recognize targets and templates with matching folds. The use of a neural network to interpret scores and a computationally efficient protocol allowed GenTHREADER to be incorporated as one of the fully automated methods available via the PSIPRED protein structure prediction server.[34]

In the following years, a number of alternative fully automated fold recognition servers became available, which also employed a hybrid approach. In 2000, Fischer *et al.* developed the INBGU method,[35] which used a combination of sequence profiles and comparisons of the PHD[36] predicted secondary structure of each target with the observed secondary structure of each template. The incorporation of secondary structure scoring allowed for the detection of distant homologues as the secondary structures are better conserved throughout evolution than sequences. In addition Kelley *et al.* developed the 3D-PSSM method,[37] which also incorporated the predicted secondary structure of target proteins using the more accurate PSIPRED method[38] and used PSI-BLAST for sequence-profile alignments. The target profiles were aligned against 3D position-specific scoring matrices (PSSMs), which were generated for templates within the fold library. For each template, PSI-BLAST was used to generate an initial 1D sequence-based PSSM, which was then further enhanced using solvation potentials, secondary structures and structural alignments, resulting in a 3D-PSSM. The FUGUE method, developed by Shi *et al.*,[39] similarly made use of structural alignments, solvent accessibility and secondary structure information in order to produce environment-specific scoring matrices. The method also made use of structure-dependent gap penalties in addition to the score matrices in order to align target sequence profiles against template structural profiles.

Hybrid methods have undergone a number of iterative improvements over the past few years in order to incorporate new innovations in sequence searching and model evaluation. For example, GenTHREADER has been updated to include additional structural information which has resulted in the detection of more remote homologues[40] and the current mGenTHREADER variation of the method also incorporates profile-profile alignments.[41] In addition, the nFOLD method[41] makes use of model quality assessment programs in an attempt to improve the rankings of models built from mGenTHREADER alignments. Other successful autonomous fold recognition servers such as TASSER,[42] also combine the best sequence searching and threading methods along with improvements in the selection of the highest quality models.

Several autonomous fully automated fold recognition web servers have arisen in the recent years through which users are able submit a protein sequence and receive their predicted structures via email. Figure 2.1 shows a screen shot of the nFOLD2 server submission


The Biocentre, University of Reading - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.biocentre.rdg.ac.uk/bioinformatics/nFOLD/nFOLD_form.html
Google.co.uk
The Biocentre, University of ...
CASP7 Target T0339
The University of Reading
The BioCentre
Biocentre Home > Core Units > Bioinformatics and Systems Biology > nFOLD > Submission Form
Core Units
Staff and Associates
Links
Publications
News and Events
Contact Us
search
Search
Quick Links:
School Quota Form
DNA Sequencing Service
Struct. Biology Enquiry
Biacore Analysis Service
Related Schools:
Agriculture, Policy and Development | Biological Sciences
Chemistry | Food Biosciences | Pharmacy | Psychology
© The University of Reading 2007
Done

The nFOLD2 Protein Fold Recognition Server

This form allows you to generate predicted 3D models of a protein from its sequence. Further information and references can be found on the nFOLD2 home page. Click 'Help' in each section for more detailed instructions.

Input sequence (single letter code) Help

```
QVAAVTFVPVSKVSGQTEVDDILAAVRPTTRLVTIMLANNETGIVMPVPE
ERVAAGLPPILVHTDAAQALGKQRVDVEDLGVDFLTIVGHKFYGPRIGAL
LYPMLFGGGQERNFRPGTENTPMIAGLGKAAELVTQNCEAYEAHMRDVRD
GQKRIHLNSQFPGTQRLPNTCNFSIRGPRLQGHVVLAQCRVLMASVGAAC
VLLSYGVPFDVARNALRLSVGRSTTRAEVDLVVQDLKQAVAQLEDQ
```

E-mail address Help

l.j.mcguffin@reading.ac.uk

Password (only required for commercial e-mail addresses) Help

Short name for sequence Help

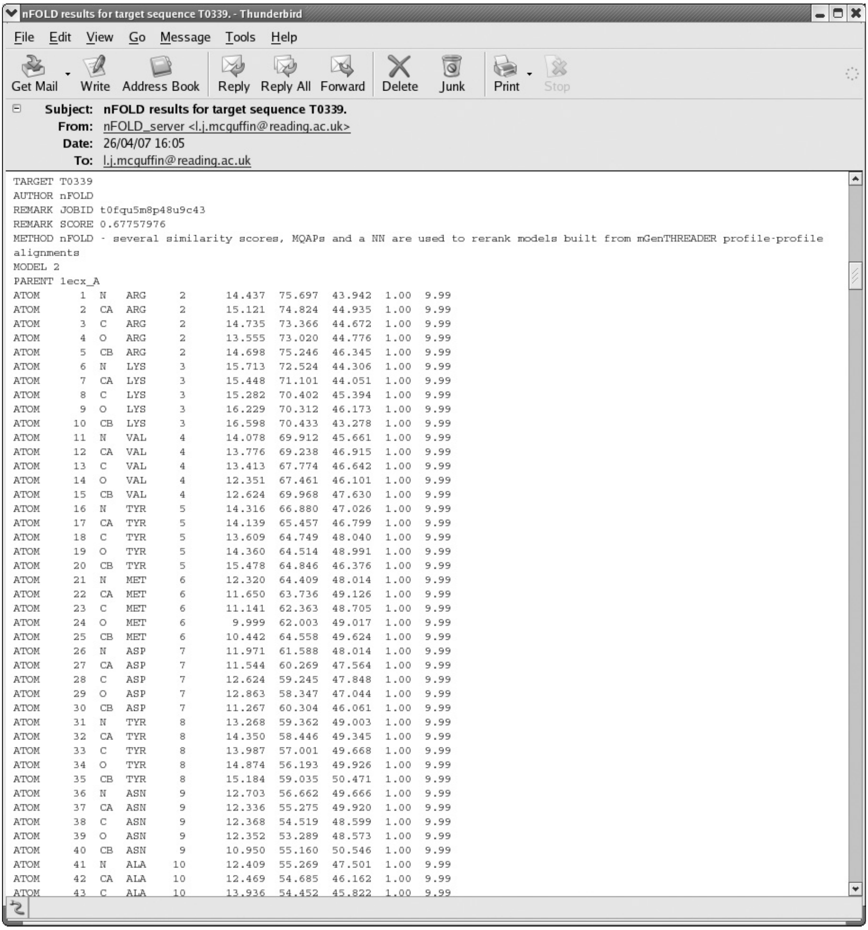
T0339

Predict Clear form

Information Provider: c.f.kemp@reading.ac.uk
Last Updated: Friday, November 24, 2006

Fig. 2.1 The nFOLD2 protein fold recognition server submission form can be found at: http://www.biocentre.rdg.ac.uk/bioinformatics/nFOLD/.

form which is freely available for academic users. The top five predicted models are returned via email in PDB format (Fig. 2.2). Figure 2.3 shows an example of the top model submitted by the nFOLD2 server for CASP7 target T0339 superposed onto the backbone of the native structure. Although this individual server often produces good models, a number of alternative sources of predicted folds are available and often the best strategy is to obtain information from many different servers to form a consensus prediction.


nFOLD results for target sequence T0339. - Thunderbird
File Edit View Go Message Tools Help
Get Mail
Write
Address Book
Reply
Reply All
Forward
Delete
Junk
Print
Stop

Subject: nFOLD results for target sequence T0339.
From: nFOLD_server <l.j.mcguffin@reading.ac.uk>
Date: 26/04/07 16:05
To: l.j.mcguffin@reading.ac.uk

```
TARGET T0339
AUTHOR nFOLD
REMARK JOBID t0fqu5m8p48u9c43
REMARK SCORE 0.67757976
METHOD nFOLD - several similarity scores, MQAPs and a NN are used to rerank models built from mGenTHREADER profile-profile
alignments
MODEL 2
PARENT 1ecx_A
ATOM       1  N   ARG     2      14.437  75.697  43.942  1.00  9.99
ATOM       2  CA  ARG     2      15.121  74.824  44.935  1.00  9.99
ATOM       3  C   ARG     2      14.735  73.366  44.672  1.00  9.99
ATOM       4  O   ARG     2      13.555  73.020  44.776  1.00  9.99
ATOM       5  CB  ARG     2      14.698  75.246  46.345  1.00  9.99
ATOM       6  N   LYS     3      15.713  72.524  44.306  1.00  9.99
ATOM       7  CA  LYS     3      15.448  71.101  44.051  1.00  9.99
ATOM       8  C   LYS     3      15.282  70.402  45.394  1.00  9.99
ATOM       9  O   LYS     3      16.229  70.312  46.173  1.00  9.99
ATOM      10  CB  LYS     3      16.598  70.433  43.278  1.00  9.99
ATOM      11  N   VAL     4      14.078  69.912  45.661  1.00  9.99
ATOM      12  CA  VAL     4      13.776  69.238  46.915  1.00  9.99
ATOM      13  C   VAL     4      13.413  67.774  46.642  1.00  9.99
ATOM      14  O   VAL     4      12.351  67.461  46.101  1.00  9.99
ATOM      15  CB  VAL     4      12.624  69.968  47.630  1.00  9.99
ATOM      16  N   TYR     5      14.316  66.880  47.026  1.00  9.99
ATOM      17  CA  TYR     5      14.139  65.457  46.799  1.00  9.99
ATOM      18  C   TYR     5      13.609  64.749  48.040  1.00  9.99
ATOM      19  O   TYR     5      14.360  64.514  48.991  1.00  9.99
ATOM      20  CB  TYR     5      15.478  64.846  46.376  1.00  9.99
ATOM      21  N   MET     6      12.320  64.409  48.014  1.00  9.99
ATOM      22  CA  MET     6      11.650  63.736  49.126  1.00  9.99
ATOM      23  C   MET     6      11.141  62.363  48.705  1.00  9.99
ATOM      24  O   MET     6       9.999  62.003  49.017  1.00  9.99
ATOM      25  CB  MET     6      10.442  64.558  49.624  1.00  9.99
ATOM      26  N   ASP     7      11.971  61.588  48.014  1.00  9.99
ATOM      27  CA  ASP     7      11.544  60.269  47.564  1.00  9.99
ATOM      28  C   ASP     7      12.624  59.245  47.848  1.00  9.99
ATOM      29  O   ASP     7      12.863  58.347  47.044  1.00  9.99
ATOM      30  CB  ASP     7      11.267  60.304  46.061  1.00  9.99
ATOM      31  N   TYR     8      13.268  59.362  49.003  1.00  9.99
ATOM      32  CA  TYR     8      14.350  58.446  49.345  1.00  9.99
ATOM      33  C   TYR     8      13.987  57.001  49.668  1.00  9.99
ATOM      34  O   TYR     8      14.874  56.193  49.926  1.00  9.99
ATOM      35  CB  TYR     8      15.184  59.035  50.471  1.00  9.99
ATOM      36  N   ASN     9      12.703  56.662  49.666  1.00  9.99
ATOM      37  CA  ASN     9      12.336  55.275  49.920  1.00  9.99
ATOM      38  C   ASN     9      12.368  54.519  48.599  1.00  9.99
ATOM      39  O   ASN     9      12.352  53.289  48.573  1.00  9.99
ATOM      40  CB  ASN     9      10.950  55.160  50.546  1.00  9.99
ATOM      41  N   ALA    10      12.409  55.269  47.501  1.00  9.99
ATOM      42  CA  ALA    10      12.469  54.685  46.162  1.00  9.99
ATOM      43  C   ALA    10      13.936  54.452  45.822  1.00  9.99
```

Fig. 2.2 Results returned via e-mail from the nFOLD2 protein fold recognition server. The atom records for the top five predicted models are returned to users in PDB format. The email also conforms to the CASP TS format.

2.5 Meta-Servers

The success of protein fold recognition can be greatly enhanced by combining the results from many different individual structure prediction servers. During the CASP5 experiment (see Section 2.6), it was clear that the best server methods were the meta-servers, which worked by automatically submitting target sequences to many different autonomous servers and then collating and interpreting the results

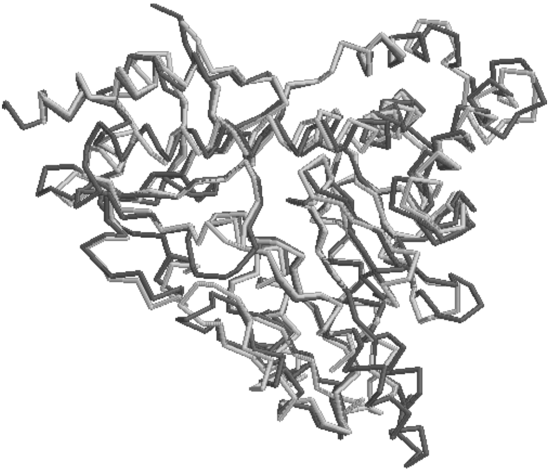


Fig. 2.3 The top predicted nFOLD2 model for CASP7 target 70339 (white) is superposed with the backbone of native structure (dark grey).

to come up with a consensus prediction. Some meta predictors such as the original Pcons method[43] and the 3D-Jury[44] method were purely focused on the selection of the highest quality model built from alignments obtained from many different servers. The Pcons method and the 3D-Jury method both worked by using structure superposition of all the predicted models for each given target. The models with the highest similarity to all other models were given the highest scores. In addition, the Pcons method also attempted to directly predict the quality of individual models and then rank them based on the combined model quality and structure comparison score.

Other meta-server methods such as 3D-SHOTGUN,[45] Pmodeller[46] and ROBETTA[47] were designed to not just simply select the best model but also to refine or improve upon the initial stock of models. For example, the 3D-SHOTGUN method attempted to build hybrid models by splicing together the best fragments from multiple models. In general, the meta-servers which attempt further improvements upon initial models have been shown to outperform those which only carry out model selection.[46]

Despite the major successes of using the meta-servers for protein fold recognition the approach has been criticized for hindering the innovation of new autonomous prediction servers. The success of the meta-servers inherently depends on the underlying methods and there is a danger that the field could stagnate if novel autonomous servers are no longer developed.

2.6 Critical Assessment of Methods: CASP, CAFASP, Livebench and EVA

In 1994, the first meeting was held on the Critical Assessment of techniques for protein Structure Prediction (CASP1).[48] The overall aim of the CASP experiments was to carry out regular blind assessments of our ability to predict protein structures. The experiments can be broken down into three main stages: firstly, the collection of prediction targets from the experimental community; secondly, the collection of predicted models from the predictors; and lastly, the assessment of the predicted models and meeting to discuss the results. For each target, the structural information is known only by the assessors and the experimentalists and the data is not publicly released, so each CASP is truly a blind experiment from the predictors' point of view. The CASP experiments have been held every two years and the latest experiment at the time of writing was CASP7, held in 2006.

During CASP3, a parallel experiment was initiated, called CAFASP1 (the first Critical Assessment of Fully Automated Structure Prediction),[49] which focused purely on the assessment of automated structure prediction servers. The advantage the CAFASP experiments was that they focused on how well each individual method performed, without any added expert human intervention. In running the CASP and CAFASP experiments in parallel, there was also the advantage of directly comparing how well the servers were performing against expert human predictors. As of 2006, the assessment of fully automated prediction methods was officially fully integrated into the main CASP experiment.

The disadvantage of CASP and CAFASP was that they were only held every two years and that, for practical reasons, a limited number of targets (typically ≤ 100) were assessed during each experiment. The LiveBench[50] and EVA[51] experiments were initiated in order to provide continuous benchmarking of structure prediction methods.

Although both LiveBench and EVA have provided benchmarks on most categories of structure prediction, it can be said that LiveBench focuses on benchmarking comparative modeling and fold recognition servers, while EVA focuses more on providing a benchmark of a wide cross section of secondary structure prediction methods and contact order prediction methods. In fact, EVA remains the only official benchmark of secondary structure prediction methods since the decision was made to drop the category during CASP5.

In general, the variety of categories of structure prediction have widened over the years as new methods have been developed to tackle new problems. For example, while the first CASP competition focused on secondary and tertiary structure prediction, the most recent CASP also provided benchmarks for disorder prediction, model quality assessment, model refinement, domain prediction, contact order prediction and even function prediction. Whilst the variety of categories has widened beyond predicting 3D structures, the traditional sub-categories of tertiary structure prediction have become less distinct.

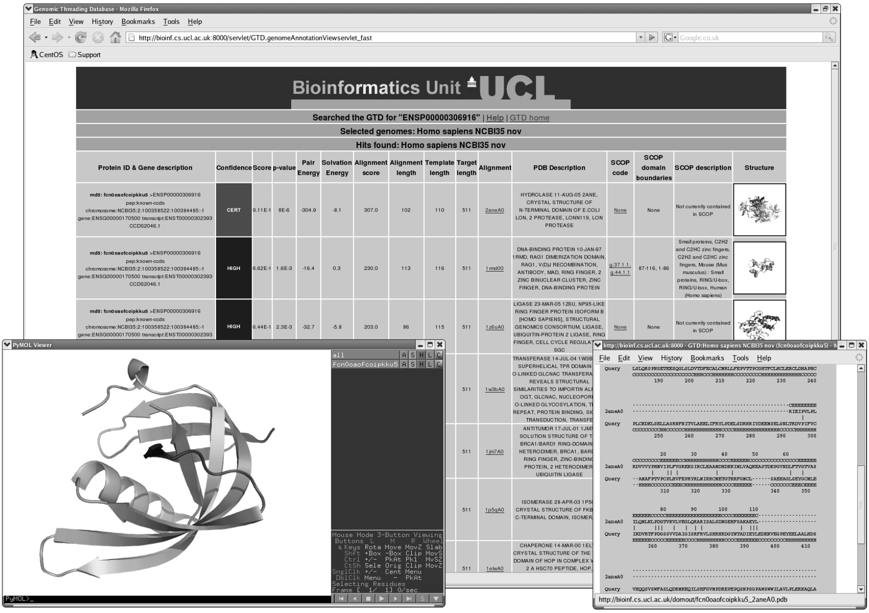
During the first CASP experiment, there were three main categories for tertiary structure prediction, depending on the information available about the target sequence. The first category — Comparative Modeling (CM) — was assigned to targets where there was a high sequence identity to a homologous template within the PDB. The second category was Fold Recognition (FR), which was reserved for targets with templates of known structure but for which little or no sequence homology could be detected. Finally, the term *ab initio* was reserved for targets which could not be modeled using any available structures as templates and so predictors would have to start "from first principles."

The boundaries between these categories have become increasingly blurred as CASP has evolved over the years. The *ab initio* category became the New Fold (NF) category as innovative methods were developed, which meant that proteins with novel folds could be modeled by assembling fragments of structures based on known folds.[52] Due to the improvement in sequence-based searching, the FR category became further subdivided into FR/H and FR/A and improved to distinguish targets with weak homologues from targets with analogous fold templates. In the most recent experiment (CASP7), the traditional boundaries have now been dropped altogether and each target is now more simply categorized as either a Template-Based Modeling (TBM) target or as Template-Free Modeling (FM) target. Of course there is still a distinction between easy and hard TBM targets; some multi-domain targets fall into both TBM and FM categories and there is still often contention, despite the simplification.

2.7 Proteome Scale Fold Recognition

The improvement of the accuracy in protein fold recognition strategies and the development of fully automated methods has meant that it is now possible to carry out tertiary structure predictions for entire proteomes. A number of databases have been developed which serve as models built from sequence-to-structure alignments for all the proteins encoded within key genomes. Over the years, several databases have become available such as 3D-GENOMICS,[53] Gene3D[54] and SUPERFAMILY,[55] most of which have used a sequence-profile based method in order to structurally annotate whole proteomes. However, the Genomic Threading Database (GTD)[56] differed from most other databases in that the GenTHREADER method was used in order to detect more remote evolutionary relationships between proteome sequences and structures.

Despite the efficiency of fully automated fold recognition methods, carrying out predictions for whole proteomes is nevertheless very computationally intensive. However, the task is parallelizable and prediction jobs can be easily distributed across clusters of processors using Grid technology in order to speed up the computation. Indeed, recently McGuffin *et al.* developed a meta-scheduling software called JYDE which was used to distribute the load of proteome scale intensive fold recognition.[57] Using their JYDE software, McGuffin *et al.* were able to structurally annotate the entire human proteome in about 24 hours using the very latest profile-profile version of mGenTHREADER (Fig. 2.4).


Genomic Threading Database - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://bioinf.cs.ucl.ac.uk:8000/servlet/GTD.genomeAnnotationViewservlet_fast
Google.co.uk
CentOS
Support
Bioinformatics Unit UCL
PyMOL Viewer
all A S H L C
fcn0oaofcoipkku5 A S H L C
PyMOL>_
http://bioinf.cs.ucl.ac.uk:8000 - GTD:Homo sapiens NCBI35 nov (fcn0oaofcoipkku5) - M
File Edit View History Bookmarks Tools Help
http://bioinf.cs.ucl.ac.uk/domout/fcn0oaofcoipkku5_2aneA0.pdb

Searched the GTD for "ENSP00000306916" | Help | GTD home

Selected genomes: Homo sapiens NCBI35 nov

Hits found: Homo sapiens NCBI35 nov

Protein ID & Gene description	Confidence	Score	p-value	Pair Energy	Solvation Energy	Alignment score	Alignment length	Template length	Target length	Alignment	PDB Description	SCOP code	SCOP domain boundaries	SCOP description	Structure
md5: fcn0oaofcoipkku5 >ENSP00000306916 pep:known-ccds chromosome:NCBI35:2:100358522:100384485:-1 gene:ENSG00000170500 transcript:ENST00000302393 CCDS2046.1	CERT	9.11E-1	8E-6	-304.9	-8.1	307.0	102	110	511	2aneA0	HYDROLASE 11-AUG-05 2ANE, CRYSTAL STRUCTURE OF N-TERMINAL DOMAIN OF E.COLI LON, 2 PROTEASE, LONN119, LON PROTEASE	None	None	Not currently contained in SCOP	
md5: fcn0oaofcoipkku5 >ENSP00000306916 pep:known-ccds chromosome:NCBI35:2:100358522:100384485:-1 gene:ENSG00000170500 transcript:ENST00000302393 CCDS2046.1	HIGH	6.62E-1	1.6E-3	-16.4	0.3	230.0	113	116	511	1rmd00	DNA-BINDING PROTEIN 10-JAN-97 1RMD, RAG1 DIMERIZATION DOMAIN, RAG1, V(D)J RECOMBINATION, ANTIBODY, MAD, RING FINGER, 2 ZINC BINUCLEAR CLUSTER, ZINC FINGER, DNA-BINDING PROTEIN	g.37.1.1, g.44.1.1	87-116, 1-86	Small proteins, C2H2 and C2HC zinc fingers, C2H2 and C2HC zinc fingers, Mouse (Mus musculus) : Small proteins, RING/U-box, RING/U-box, Human (Homo sapiens)	
md5: fcn0oaofcoipkku5 >ENSP00000306916 pep:known-ccds chromosome:NCBI35:2:100358522:100384485:-1 gene:ENSG00000170500 transcript:ENST00000302393	HIGH	6.44E-1	2.3E-3	-32.7	-5.8	203.0	86	115	511	1z6uA0	LIGASE 23-MAR-05 1Z6U, NP95-LIKE RING FINGER PROTEIN ISOFORM B [HOMO SAPIENS], STRUCTURAL GENOMICS CONSORTIUM, LIGASE, UBIQUITIN-PROTEIN 2 LIGASE, RING FINGER, CELL CYCLE REGULAT SGC	None	None	Not currently contained in SCOP	
									511	1w3bA0	TRANSFERASE 14-JUL-04 1W3B SUPERHELICAL TPR DOMAIN O-LINKED GLCNAC TRANSFERA REVEALS STRUCTURAL SIMILARITIES TO IMPORTIN AL OGT, GLCNAC, NUCLEOPORI O-LINKED GLYCOSYLATION, T REPEAT, PROTEIN BINDING, SI TRANSDUCTION, TRANSFE				
									511	1jm7A0	ANTITUMOR 17-JUL-01 1JM7 SOLUTION STRUCTURE OF T BRCA1/BARD1 RING-DOMAIN HETERODIMER, BRCA1, BAR RING FINGER, ZINC-BINDIN PROTEIN, 2 HETERODIMER UBIQUITIN LIGASE				
									511	1p5qA0	ISOMERASE 28-APR-03 1P5 CRYSTAL STRUCTURE OF FKB C-TERMINAL DOMAIN, ISOMER				
									511	1elwA0	CHAPERONE 14-MAR-00 1EL CRYSTAL STRUCTURE OF THE DOMAIN OF HOP IN COMPLEX W 2 A HSC70 PEPTIDE, HOP,				

```
Mouse Mode 3-Button Viewing
 Buttons  L    M    R  Wheel
  & Keys Rota Move MovZ Slab
    Shft +Box -Box Clip MovS
    Ctrl +/-  PkAt Pk1  MvSZ
    CtSh Sele Orig Clip MovZ
SnglClk +/-  Cent Menu
 DblClk Menu   -  PkAt
Selecting Residues
Frame [  1/  1] 0/sec
```

```
Query    LSLQRSPNSETEESQGLSLDVTDFECALCMRLLFEPVTTPCGHTFCLKCLERCLDHAPHC
         CCCCCCCCCCCCCCCCCCCCHHHCCCHHHHHHHCCCCCCCCCHHHHHHHHHHHHHCCCCC
                 190       200       210       220       230       240


         -----------------------------------------------------CEEEEEEEE
2aneA0   -----------------------------------------------------RIEIPVLPL
                                                                  |
Query    PLCKDKLSELLASRNFNITVLAEELIFRYLPDELSDRKRIYDEEMSELSNLTRDVPIFVC
         CCCCCCCCCHHCCCCCCHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHCCCCCHHHHHHH
                 250       260       270       280       290       300

                 20        30        40        50        60
         CCCCCCCCCEEEEEEECCHHHHHHHHHHHHCCCCEEEEEEECCCCCCCCCCCCCCCCEEEEEEE
2aneA0   RDVVVYPHMVIPLFVGREKSIRCLEAAMDHDKKIMLVAQKEASTDEPGVNDLFTVGTVAS
             |    || |          |    |                 |   |
Query    --AMAFPTVPCPLHVFEPRYRLMIRRCMETGTKRFGMCL-----SAEHAGLSEYGCMLE
         --HHHHCCCCCCCEEECHHHHHHHHHHHHHCCCCEEEEEEE-----CCCCCCCEEECEEEE
                 310       320       330            340       350

                 80        90        100       110
         EEEEEEECCCCCEEEEEEEEEEEEEEEEEEEEEECCCCEEEEEEEEEC-----------------
2aneA0   ILQMLKLPDGTVKVLVEGLQRARISALSDNGEHFSAKAEYL-----------------
         |      |||   |   |  | |               |  |||
Query    IKDVRTFPDGSSVVDAIGISRFRVLSHRHRDGYNTADIEYLEDEKVEGPEYEELAALHDS
         EEEEEEECCCCCEEEEEEECCCCHHHHHHHHHCCCCEEEEEEEEECCCCCCCCCHHHHHHHHHH
             360       370       380       390       400       410


         ------------------------------------------------------------
2aneA0   ------------------------------------------------------------

Query    VHQQSVSWFASLQDRMKEQILSHFGVMPDREPEPQSNPSGPAWSWWILAVLPLERKAQLA
```

Fig. 2.4 The Genomic Threading Database web interface.[56] Results show an example of a confident model built for human protein domain, which could not be structurally annotated prior to the high throughput fold recognition carried out by McGuffin *et al.*[57]

The development of the JYDE Grid middleware and the Genomic Threading Database was in conjunction with the e-Protein project. The e-Protein project was part of the UK e-Science initiative and was set up in order to bring together the software and hardware resources of Imperial, UCL and the European Bioinformatics Institute in order to provide a fully automated pipeline for structural and functional annotation of key proteomes.

2.8 Future Outlook

The traditional categories of tertiary structure prediction are becoming less useful for classifying methods. Sequence-based methods are able detect homology between targets and templates beyond the twilight zone and the most successful fold recognition methods now incorporate profile-profile based searches in order to produce accurate sequence to structure alignments. In addition, fragment assembly methods are increasingly being used to model the folds of larger proteins and the traditional threading methods are falling out of common usage. It is conceivable that in the near future, template-based modeling will be the only technique required to model the folds of almost all new protein sequences as our knowledge of "fold space" becomes complete. Indeed, a study by Zhang *et al.*,[4] speculates that we may already be reaching that point. Template-free modeling methods will, therefore, be required less often for the modeling of complete folds. It is likely that template-free modeling techniques will become more widely used for model refinement i.e. for modeling loops and unstructured regions where the sequence-to-structure alignments are poor.

As more targets become available at CASP experiments and the ongoing benchmarking experiments such as LiveBench continue to drive the development of methods, it will be increasingly impractical for humans to keep up with fully automated prediction servers. The accuracy of fully automated methods has greatly increased to the extent that the meta-servers are now becoming more successful predictors than humans.[58] Conceivably, in the near future all of the top prediction methods will be server methods and there will be less to be gained in the modeling process from human intervention. Despite the usefulness and success of the meta-server methods, developers should continue to carry out incremental improvements in the underlying autonomous methods. The recent success of the iterative version of the TASSER[42] method developed by Zhang in CASP7, has highlighted the fact that autonomous methods can still show significant improvements. With further efforts, new autonomous methods will continue to improve upon model quality and outperform the meta-servers which rely on predictions from older methods. Increasingly sophisticated algorithms which can run in hours on a single server should continue to become more competitive against the intensive "brute force" approaches which require thousands of CPU days to provide a single model. Despite this, Grid technology will continue to play a necessary role in high throughput proteome-wide template-based modeling.[57]

One area of structure prediction which should also grow in popularity is the development of model quality assessment programs (MQAPs). As the number of methods and available templates continue to increase, predictors will be left with a choice of potentially hundreds of models per target. Increasingly, the problem will be how to select the best alignments which produce the most accurate models, rather than one of identifying an appropriate template or getting a reasonably good alignment.

It is conceivable that the MQAPs and meta-servers, which purely carry out model selection, will be categorized differently from methods which further refine models. Indeed, in CASP7 there are now separate categories for quality assessment (QA) methods and model refinement methods (CASPR). The CASPR experiment focuses on a few targets and encourages predictors to build models closer to native structures rather than constructing them from the best available template. However, it is clear that both model selection and model refinement techniques will be essential for the development of the most accurate structure prediction pipelines.

The long-term future outlook for template-based modeling will be to continue taking the next logical step from tertiary structure modeling towards quaternary structure modeling. This will continue to bring computational structural biology closer to the realm of systems biology. It will become increasingly important to verify and improve the quality of high throughput protein interaction data being produced by a number of experimental and predictive methods. The detailed 3D modeling of protein interactions will enable us to understand how novel interactions may be occurring between proteins at the atomic level rather than simply inferring which proteins interact. Preliminary steps towards template-based modeling of interactions have already been made and it is likely that we will continue to see further progress in this area.[59]

References

1. Berman HM, Westbrook J, Feng Z, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**: 235–242.

2. Jones DT. (2000) A practical guide to protein structure prediction. *Meth Mol Biol* **143**: 131–154.
3. Orengo CA, Jones DT, Thornton JM. (1994) Protein superfamilies and domain superfolds. *Nature* **372**: 631–634.
4. Zhang Y, Skolnick J. (2005) The protein structure prediction problem could be solved using the current PDB library. *Proc Natl Acad Sci USA* **102**: 1029–1034.
5. Jones DT, Taylor WR, Thornton JM. (1992) A new approach to protein fold recognition. *Nature* **358**: 86–89.
6. Taylor WR, Orengo CA. (1989) Protein structure alignment. *J Mol Biol* **208**: 1–22.
7. Rost B. (1999) Twilight zone of protein sequence alignments. *Protein Eng* **12**: 85–94.
8. Needleman SB, Wunsch CD. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–453.
9. Smith TF, Waterman MS. (1981) Identification of common molecular subsequences. *J Mol Biol* **147**: 195–197.
10. Pearson WR, Lipman DJ. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**: 2444–2448.
11. Altschul SF, Gish W, Miller W, *et al.* (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
12. Dayhoff MO, Schwartz RM, Orcutt BC. (1978) A model of evolutionary change in proteins, In M. Dayhoff (ed.), *Atlas of Protein Sequence and Structure,* pp. 345–352. Silver Springs: National Biomedical Research Foundation, Washington DC.
13. Gonnet GH, Cohen MA, Benner SA. (1992) Exhaustive matching of the entire protein sequence database. *Science* **256**: 1443–1445.
14. Jones DT, Taylor WR, Thornton JM. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**: 275–282.
15. Henikoff S, Henikoff JG. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**: 10915–10919.
16. Brenner SE, Chothia C, Hubbard TJ. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA* **95**: 6073–6078.
17. Altschul SF, Gish W. (1996) Local alignment statistics. *Meth Enzymol* **266**: 460–480.
18. Altschul SF, Madden TL, Schaffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**: 3389–3402.
19. Park J, Karplus K, Barrett C, *et al.* (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* **284**: 1201–1210.

20. Muller A, MacCallum RM, Sternberg MJ. (1999) Benchmarking PSI-BLAST in genome annotation. *J Mol Biol* **293**: 1257–1271.
21. Rychlewski L, Jaroszewski L, Li W, Godzik A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* **9**: 232–241.
22. Ohlson T, Wallner B, Elofsson A. (2004) Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins* **57**: 188–197.
23. Soding J, Biegert A, Lupas AN. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucl Acids Res* **33**: W244–W248.
24. Bowie JU, Luthy R, Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**: 164–170.
25. Ponder JW, Richards FM. (1987) Internal packing and protein structural classes. *Cold Spring Harb Symp Quant Biol* **52**: 421–428.
26. Bowie JU, Clarke ND, Pabo CO, Sauer RT. (1990) Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins* **7**: 257–264.
27. Luthy R, Bowie JU, Eisenberg D. (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356**: 83–85.
28. Sippl MJ. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**: 859–883.
29. Lemer CM, Rooman MJ, Wodak SJ. (1995) Protein structure prediction by threading methods: evaluation of current techniques. *Proteins* **23**: 337–355.
30. Godzik A, Kolinski A, Skolnick J. (1992) Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol* **227**: 227–238.
31. Bryant SH. (1996) Evaluation of threading specificity and accuracy. *Proteins* **26**: 172–185.
32. Thiele R, Zimmer R, Lengauer T. (1999) Protein threading by recursive dynamic programming. *J Mol Biol* **290**: 757–779.
33. Jones DT. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* **287**: 797–815.
34. McGuffin LJ, Bryson K, Jones DT. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**: 404–405.
35. Fischer D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac Symp Biocomput* 119–130.
36. Rost B, Sander C, Schneider R. (1994) PHD — an automatic mail server for protein secondary structure prediction. *Comput Appl Biosci* **10**: 53–60.
37. Kelley LA, MacCallum RM, Sternberg MJ. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* **299**: 499–520.

38. Jones DT. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**: 195–202.
39. Shi J, Blundell TL, Mizuguchi K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* **310**: 243–257.
40. McGuffin LJ, Jones DT. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19**: 874–881.
41. Jones DT, Bryson K, Coleman A, *et al.* (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins* **61**(7): 143–151.
42. Zhang Y, Arakaki AK, Skolnick J. (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* **61**(7): 91–98.
43. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci* **10**: 2354–2362.
44. Ginalski K, Elofsson A, Fischer D, Rychlewski L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* **19**: 1015–1018.
45. Fischer D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins* **51**: 434–441.
46. Wallner B, Fang H, Elofsson A. (2003) Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins* **53**(6): 534–541.
47. Chivian D, Kim DE, Malmstrom L, *et al.* (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins* **53**(6): 524–533.
48. Moult J, Pedersen JT, Judson R, Fidelis K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**: ii–v.
49. Fischer D, Barret C, Bryson K, *et al.* (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins* (3): 209–217.
50. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci* **10**: 352–361.
51. Rost B, Eyrich VA. (2001) EVA: large-scale analysis of secondary structure prediction. *Proteins* (5): 192–199.
52. Jones DT, McGuffin LJ. (2003) Assembling novel protein folds from super-secondary structural fragments. *Proteins* **53**(6): 480–485.
53. Fleming K, Muller A, MacCallum RM, Sternberg MJ. (2004) 3D-GENOMICS: a database to compare structural and functional annotations of proteins between sequenced genomes. *Nucl Acids Res* **32**: D245–D250.
54. Buchan DW, Rison SC, Bray JE, *et al.* (2003) Gene3D: structural assignments for the biologist and bioinformaticist alike. *Nucl Acids Res* **31**: 469–473.
55. Gough J. (2002) The SUPERFAMILY database in structural genomics. *Acta Crystallogr D Biol Crystallogr* **58**: 1897–1900.

56. McGuffin LJ, Street SA, Bryson K, *et al.* (2004) The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucl Acids Res* **32**: D196–D199.
57. McGuffin LJ, Smith RT, Bryson K, *et al.* (2006) High throughput profile-profile based fold recognition for the entire human proteome. *BMC Bioinform* **7**: 288.
58. Fischer D. (2006) Servers for protein structure prediction. *Curr Opin Struct Biol* **16**: 178–182.
59. Grimm V, Zhang Y, Skolnick J. (2006) Benchmarking of dimeric threading and structure refinement. *Proteins* **63**: 457–465.

Chapter 3

Scoring Functions for Protein Structure Prediction

Francisco Melo* and Ernest Feytmans†

3.1 Introduction

A potential energy function is an essential tool to predict the three-dimensional structure of a protein. Two fundamentally different approaches exist to obtain a potential energy function. In the first one, which is of an inductive nature, a mathematical model that describes the system is assumed without previous knowledge about the system's properties. In this approach, spectroscopic and thermodynamic experimental data and results from quantum mechanical calculations in simple molecules are used to fit the mathematical model adopted. The resulting potential is directly extrapolated to more complex molecules by assuming that a common behavior will exist in both cases. The potentials obtained by an inductive approach are called semi-empirical potentials or classical force fields. The second approach is deductive or knowledge-based, and assumes the

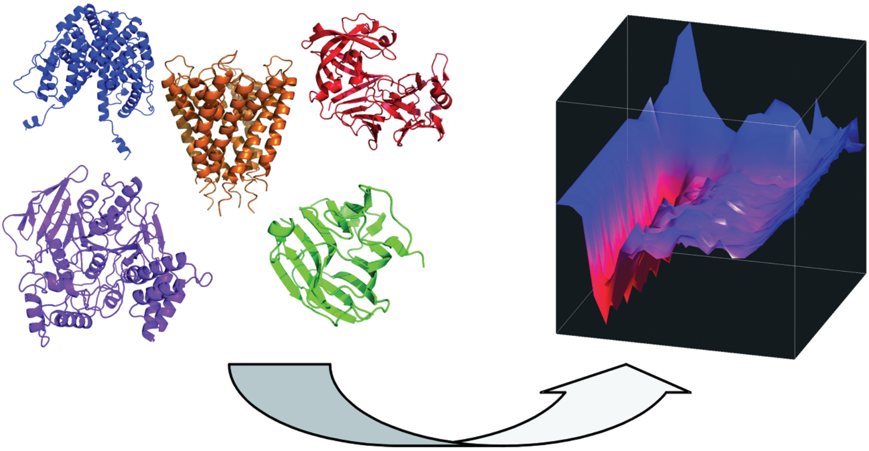
*Corresponding author.
Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile Alameda 340. Santiago, Chile. Email: fmelo@bio.puc.cl.
†Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Génopode, CH-1015 Lausanne, Switzerland.

opposite scenario: the potential energy function of a large macromolecular-solvent system is complex and thus cannot be modeled by a simple and pre-conceived mathematical model. In order to obtain an accurate description of the potential energy function, experimental data from large macro-molecular-solvent systems should be used. Given their deductive nature, potentials obtained by a deductive approach are also called empirical potentials, knowledge-based potentials, statistical potentials or scoring functions, and constitute the main subject of this chapter. Here, we will refer to those as scoring functions.

In this chapter, we attempt to give a simple and general overview of scoring functions and their typical applications to protein structure prediction. There are some excellent reviews on the literature that provide more details about most of the topics covered here.[1–5] By no means does this chapter attempt to be an extensive revision of the state-of-the-art on the field.

Scoring functions are widely used in protein structure prediction because of their relative simplicity, accuracy, and computational efficiency. Among their applications we found the assessment of experimentally determined and computationally predicted protein structures,[1] *ab initio* protein structure prediction,[2] fold recognition or threading,[3] detection of native-like protein conformations[4] and prediction of protein stability.[5]

Scoring functions do not classify forces, but instead, based on geometrical descriptors (e.g. distance, angles, etc.) extract information from experimental data of known protein structures, by deriving the propensities for the interaction of two or more bodies (Fig. 3.1). Using principles of statistical mechanics, these scoring functions describe microstates of atomic interactions within protein structures as probabilities of discrete events normalized in reference to the whole system (i.e. all possible microstates). Based on the holistic nature of scoring functions, which accounts for atom-atom interactions as well as solvation effects, they are also commonly referred to as effective energy functions. Furthermore, their strong foundations in statistical mechanics allow us to recognize a physical basis in a phenomena otherwise purely statistical.[6]



$$\Delta E(s) = -RT \ln\left[\frac{P(s)}{P(r)}\right]$$

Fig. 3.1 Overall scheme for the derivation of scoring functions. From a non-redundant set of experimental native protein structures that represent the known folded state, the probability of occurrence of a particular interaction between bodies defined by a restrained geometrical variable *s*, described in the equation as *P*(*s*), is first calculated. Then, the probability of occurrence of a reference state *r* (described as *P*(*r*) in the equation), which represents an average state of all interactions by assuming a null-interaction model (i.e. an ideal gas mixture of bodies), is also calculated from the same experimental data. Finally, the inverse Boltzmann law is used to infer a scoring function that contains the difference of pseudo-energies or scores between the folded and unfolded states in proteins upon a given set of geometrical restraints. In this illustration, the score difference obtained for main chain hydrogen bonds is given in the *Z*-axis, as a function of sequence separation between the interacting atoms (*Y*-axis) and the Euclidean distance that separates them in three-dimensional space (*X*-axis), as described by the left-hand rule.

Scoring functions are informatic functions.[7] Their capacity to properly describe the atomic interactions that are recurrent in native protein conformations depends on many parameters and on how the data is compressed and classified. In addition, it is also important to emphasize that the performance of a scoring function does not only depend on how the information is extracted, compressed and classified, but also, on how the information is used.

3.2 Structure and Components of Scoring Functions

Scoring functions are multi-dimensional matrices that hold a compressed and simplified representation of the existing experimental data. Irrespectively of the application, all scoring functions have four main components that define their structure and capabilities: a body definition, a geometrical descriptor, a reference system, and a set of restraints. Most of these components can be constituted by a single type or multiple types and can be designed to provide a rough or detailed description of the data.

The body definition consists of selecting the type of objects that will be treated differently from the experimental data. For example, a body definition can consist of single atoms or centroids of groups of atoms. Additionally, different atoms or centroids can be further grouped into a single body type if they share some common features, such as their physicochemical properties.[8] The definition of body types is adopted to reduce the size of the matrix because the experimental data available is always limited. Additionally, a proper representation of body types helps to reduce the dependency between bodies, which in theory is a condition that should be fulfilled. In practice this is very difficult to achieve because many dependencies between bodies arise, mostly as a consequence of atom-atom connectivity issues.

The geometrical descriptors typically adopted to describe the interactions between the defined bodies are pair-wise distances, simple angles, dihedral or torsion angles, radial or angular densities, or any combination thereof.[9]

The ideal reference system should consist of a weighted average representation of all possible subsystems.[6] It is used as a reference frame to calculate a meaningful score for a particular state. Because of the lack of experimental data, a null interaction model is normally adopted as a reference system, which should approximate to the weighted average state of all possible states of a system. In Section 3.3.2, we describe in more detail this important component of the scoring functions and its current limitations.

The set of restraints are used to define the limits or the scope of the scoring function, to split it into different portions that require specific treatment, to avoid the emergence of some artifacts and to improve its performance in particular applications. Typical restraints of scoring functions include a minimum and a maximal distance range, the splitting of local and non-local interactions, varying resolutions to store the data, symmetry or asymmetry, distinct derivation and utilization of the scoring function, and different criteria to select the source experimental data that is used to derive them.

With these main ingredients, simple or complex scoring functions can be derived from experimental data, and used for different applications with a varying degree of success. In the following subsections we describe some of the most typical scoring functions used for protein structure prediction.

3.2.1 *Contact Scoring Functions*

A contact scoring function constitutes the simplest and the more coarse-grained representation of pair-wise interactions between bodies in native protein structures.[10] These scoring functions were the first to be developed because they consist of a minimum size matrix that can be properly filled with few experimental data. Typically, a contact potential is represented by a squared bi-dimensional matrix [A][A], where A specifies the total number of different bodies defined (Fig. 3.2). For example, a contact potential relying on a body definition that consists of the alpha carbons of the 20 standard amino acids will have a value of A equals to 20. If derived asymmetrically, this potential will represent the propensities of interaction of 400 body pairs at a given distance threshold in 3D space.

Contact scoring functions are typically derived symmetrically for the beta carbons or side chain centroids of standard amino acids. Therefore, they contain a total of 210 different pseudo-energy, score or propensity terms (i.e. $N \times (N + 1)/2$, where N is the total number of bodies). A relevant parameter in these scoring functions, which is fixed at different values depending on the particular application, is constituted by the maximum distance threshold used to

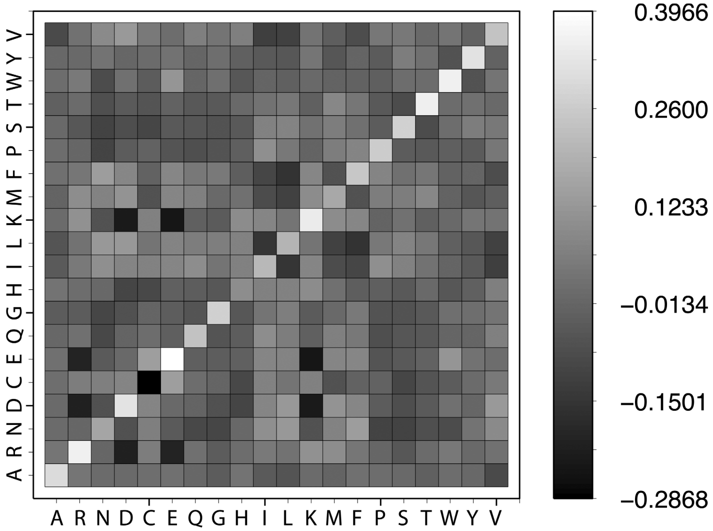


Fig. 3.2 A contact scoring function. Amino acid-based contact scoring function symmetrically calculated for pairs of non-local interacting beta carbons under 10.0 Å. In the case of glycine, a virtual beta carbon with the correct chirality and assuming standard stereochemistry was built and used to calculate the contacts. The approximated energy score of each pair-wise interaction can be obtained from the spectrum bar.

define a contact. Pair-wise interactions occurring at larger distances are not considered. Additionally, a contact between two bodies in the structure could be further restrained by other geometrical parameters such as angles and/or orientations that are defined on a given reference frame in the same structure. These additional restraints are often included in an effort to capture only the effective interactions between bodies.

3.2.2 *Distance-dependent Scoring Functions*

Distance-dependent scoring functions capture the propensities of pair-wise interactions between defined bodies up to a maximum distance threshold in a binned fashion.[11] Each bin contains the score for a particular distance range defined by a pair of minimum and

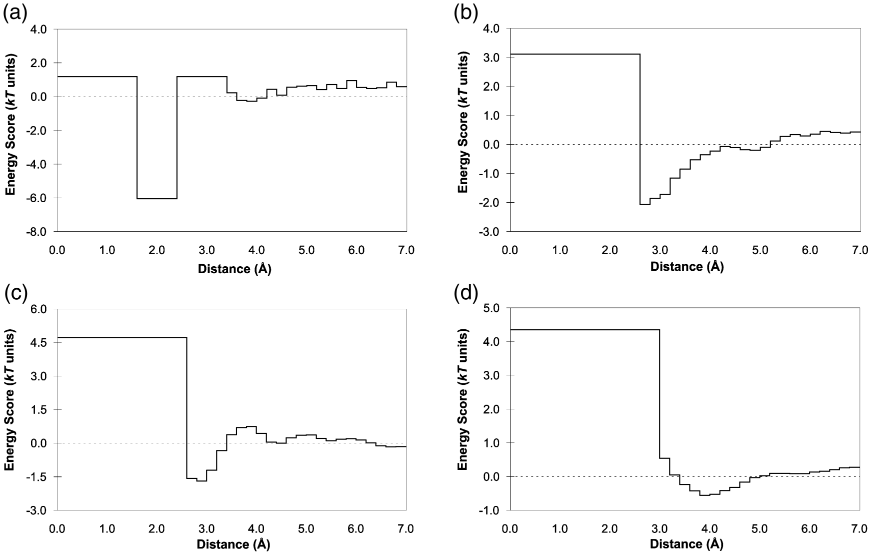


Fig. 3.3 Examples of distance-dependent scoring functions. Atomic non-local distance-dependent scoring functions in the range 0.0–7.0 Å for (a) the sulfur-sulfur interaction between two non-local cysteine side chains; (b) salt bridge interaction between lysine Nε and carboxylic side chain oxygens of aspartic and glutamic acids, respectively; (c) hydrogen bond interaction between N and O main chain atoms; and (d) the interaction between methyl-methyl side chain groups.

maximum values. Typically, the scoring functions of this type have repulsive values at short distances, one or more minima at intermediate distances, and values near zero for large distances (Fig. 3.3).

Distance-dependent scoring functions currently are the most widely used for protein structure prediction because of their favorable tradeoff between geometric simplicity and amount of information content they are capable of encoding. Such scoring functions have been successfully used in many different applications of protein structure prediction, which include: protein fold assessment or detection of miss-folded proteins, the ranking of protein conformations, the selection of native-like conformations, the detailed detection of local errors in protein models, and the assessment of the stability of single mutant proteins. These scoring functions often constitute an important component of different protein structure prediction methods (i.e. *ab initio* protein structure prediction, fold recognition, and comparative modeling).

A distance-dependent scoring function is typically represented by a matrix of four dimensions: [A][A][K][D], where A specifies the total number of different bodies defined, K provides a distinction between local and non-local interactions, and D describes the different distance ranges or bins to represent the interactions.

Many variants of distance-dependent scoring functions have been described. These can be calculated at the atomic or at the amino acid level. They can also describe only the local, only the non-local, or both types of interactions. The bins, used to convert a continuous space into a discrete representation of the pair-wise interactions, can be homogeneous or heterogeneous. Any combination of these variants can be adopted to calculate a distance-dependent scoring function depending on the particular application intended.

These scoring functions are normally derived asymmetrically, because the shape of distance-dependent pair-wise scoring functions for a common body pair are different (i.e. [i][j] and [j][i]). This means that these scoring functions are sensitive to the order along the protein chain of the interacting bodies, which makes them more suitable than contact scoring functions to capture some detailed structural aspects, such as those that occur both at the core and at the boundaries of regular secondary structures in proteins.

3.2.3 *Accessible Surface Scoring Functions*

Accessible surface scoring functions attempt to capture the propensity of interaction of the defined bodies with the solvent (Fig. 3.4). They typically include a residue-based solvent accessibility[3] or an atomic solvent accessibility description.[12] In both cases, these scoring functions offer a simplified representation of solvent exposure by an implicit model (i.e. solvent is not explicitly included when the scoring function is calculated nor when it is used). Often, the accessible surface of a body is represented by the total number of other bodies that are found surrounding it at a given and fixed distance threshold (i.e. within a fixed radius sphere that is centered on a body).

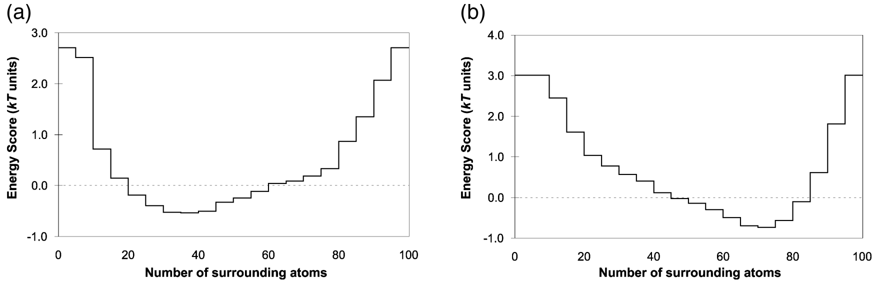


Fig. 3.4 Examples of accessible surface scoring functions. Atomic-based accessible surface scoring functions for (a) carboxylic side chain oxygens of aspartic and glutamic acids; and (b) aromatic carbons. The total number of surrounding atoms are counted within a 10.0 Å sphere radius centered on the atom.

3.2.4 *Combined Scoring Functions*

Because accessible surface scoring functions complement the information of contact and distance-dependent scoring functions, they are sometimes used in a combined fashion.[1] In this scenario, accessible surface terms capture the interactions between the protein and the solvent, and the contact and distance-dependent terms capture the intramolecular protein interactions.

To properly combine these two independent scoring functions, a normalization scheme is required.[6] This is because accessible surface scoring functions only calculate a single term for each body, while the contact and distance-dependent scoring functions often consider many. Therefore, if directly combined (i.e. simply added), the total weight of the accessible surface score in the resulting combined scoring function is quite low.

3.3 How is a Scoring Function Derived?

The derivation of scoring functions for protein structure prediction is carried out by using the inverse Boltzmann law, which in its general form states:

$$\Delta E(s) = -RT \ln\left[\frac{p^{F}(s)}{p^{U}(s)}\right] \tag{3.1}$$

where $\Delta E\ (s)$ represents the change of energy associated with the transition between the unfolded and the folded states defined by the variable s; $p^F(s)$ represents the probability of occurrence of the subsystem defined by s in the folded state F; $p^U(s)$ represents the probability of occurrence of the subsystem defined by s in the unfolded state U; R is the gas constant and T the absolute temperature measured in Kelvin. Unfortunately, the term $p^U(s)$, which describes the reference state, cannot be directly calculated because a homogeneous and unbiased experimental sample of the unfolded state of proteins is not available. As an attempt to circumvent this problem, the observed interactions among any pair of atoms by assuming a null interaction model are often used as a reference system to derive the scoring function.[11]

More specifically, and using as an example the derivation proposed by Sippl,[11] the distance-dependent scoring function can be calculated using the following expression:

$$\Delta S_k^{ij}(d) = RT\ln\left[1 + M_{ijk}\cdot\sigma\right] - RT\ln\left[1 + M_{ijk}\cdot\sigma.\frac{f_k^{ij}(d)}{f_k^{xx}(d)}\right] \tag{3.2}$$

where M_{ijk} is the number of occurrences for the interaction of atom types i and j at sequence separation k and is calculated as follows:

$$M_{ijk} = \sum_{d=1}^{r} f(i,j,k,d) \tag{3.3}$$

where r is the number of classes of distance. σ is the weight given to each observation. $\sigma = 0.02$ is generally used, so that with 50 observations, $f_k^{ij}(d)$ and $f_k^{xx}(d)$ have equal weights for the calculation of $\Delta S_k^{ij}(d)$ (which represents the change in score between the state defined by i, j, k, and d, and the average score of the reference state with the same parameters k and d). $f_k^{ij}(d)$ is the relative frequency of occurrence for the interaction of atom types i and j at sequence separation k in the class of distance d, and is calculated by the following expression:

$$f_k^{ij}(d) = \frac{f(i,j,k,d)}{M_{ijk}} \tag{3.4}$$

where $f_k^{xx}(d)$ is the relative frequency of occurrence for all the interactions of any two atom types at sequence separation k in the class of distance d and is calculated as follows:

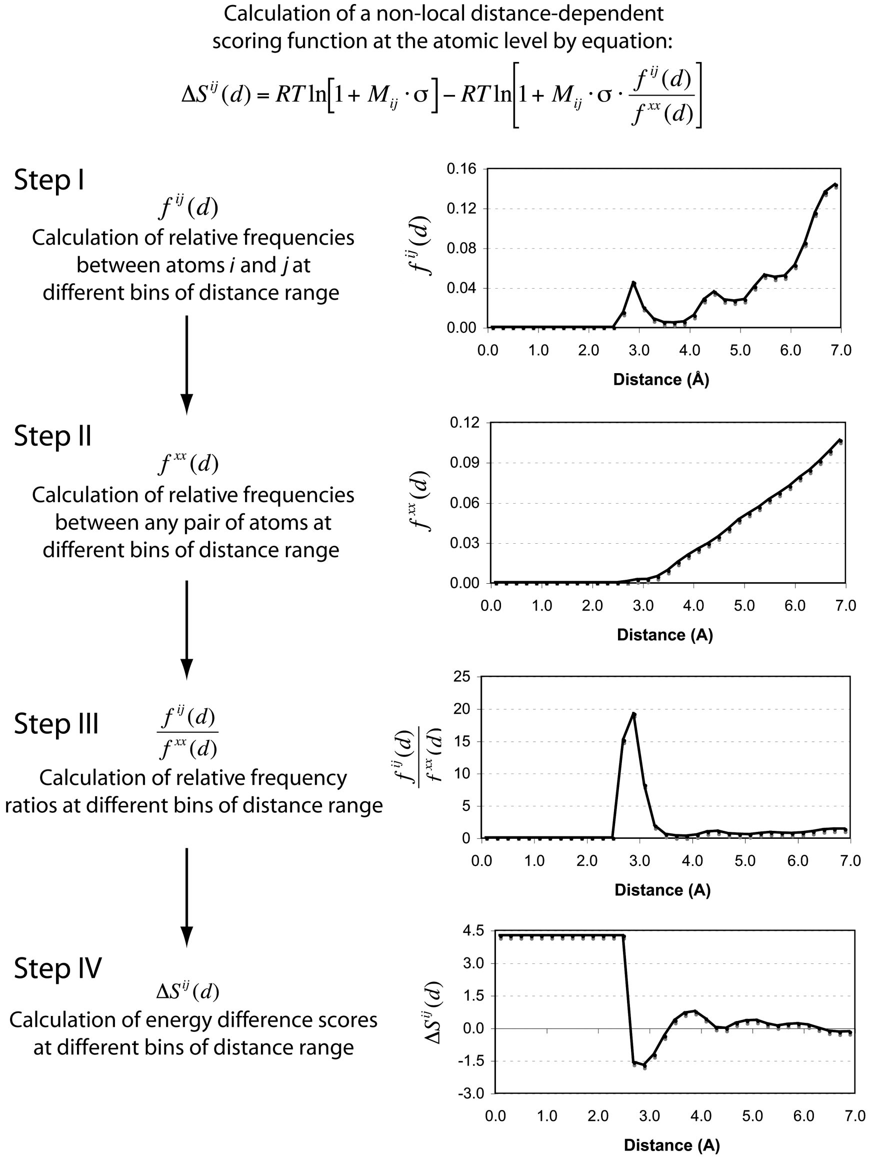
$$f_k^{xx}(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} f(i,j,k,d)}{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{d=1}^{n} f(i,j,k,d)} \tag{3.5}$$

where n is the number of different atom types and r is the number of distance classes. The temperature T was set to 293 K, so that RT is equivalent to 0.582 kcal/mole.

The distance-dependent scoring functions are typically derived for a maximal distance range and divided into a fixed number of homogeneous or heterogeneous bins or distance classes; both parameters being arbitrarily defined. Finally, different atom types for all non-hydrogen atoms in the 20 standard amino acids are defined in various ways. These definitions are typically performed in order to minimize the size of the matrix that stores a finite number of observations from experimental data, by collapsing different atoms or amino acids that are similar from a physicochemical point of view into a same type.[8,13] An example showing the key steps in the derivation of a distance-dependent scoring function for the non-local interactions is given in Fig. 3.5.[12] This example uses the uniform density reference system (see below) described by Sippl.[11]

3.3.1 *Selection of Source Experimental Data*

To avoid any possible bias in the statistical representation of data and their subsequent utilization, the scoring functions should be derived



Calculation of a non-local distance-dependent scoring function at the atomic level by equation:

$$\Delta S^{ij}(d) = RT\ln\left[1 + M_{ij} \cdot \sigma\right] - RT\ln\left[1 + M_{ij} \cdot \sigma \cdot \frac{f^{ij}(d)}{f^{xx}(d)}\right]$$

Step I

$f^{ij}(d)$

Calculation of relative frequencies between atoms *i* and *j* at different bins of distance range

Step II

$f^{xx}(d)$

Calculation of relative frequencies between any pair of atoms at different bins of distance range

Step III $\frac{f^{ij}(d)}{f^{xx}(d)}$

Calculation of relative frequency ratios at different bins of distance range

Step IV

$\Delta S^{ij}(d)$

Calculation of energy difference scores at different bins of distance range

Fig. 3.5 Derivation of a scoring function. The critical quantities required to derive a scoring function are shown. In this example, a non-local distance-dependent scoring function for the hydrogen bond between main chain atoms N and O is calculated. A total of 138 native proteins were used to calculate this scoring function. The term *Mij* in this particular example corresponds to a total of 73 855 interactions and σ is 0.02.

from a non-redundant set of experimental protein structures. The most typical approach to achieve this consists of using a set of unrelated proteins, which is defined upon a fixed sequence similarity threshold after an all-against-all pair-wise sequence comparison.[14] A sequence similarity threshold is adopted based on the known sequence to structure relationship first observed and described by Chothia and Lesk.[15]

Since the source data contains proteins of known experimental structure, another option that is more valid, but also more expensive in terms of calculations, is to define the non-redundant set based on the structural comparison of the structures.[16] Currently, there is no need to calculate these sets, since several non-redundant data sets obtained at different structural similarity thresholds are easily available and frequently updated.[17]

However, all the existing sets of non-redundant protein structures have been calculated through automated procedures, which are unable to detect all possible sources of errors. For that reason, it is still important to filter them by checking some important features of the protein structures, therein defined such as the resolution, duplicated or missing atoms, structural gaps, atom clashes, and size. This issue is extremely important, because the incorporation of protein with errors in the data set used to derive a scoring function can have a significant impact on its final performance. As the size of the PDB database increases, this issue becomes more relevant, since a semi-automated method for the selection of the non-redundant data set is more difficult. Therefore, special attention should be paid to this issue in the future.

3.3.2 *Reference Systems*

The calculation of scoring functions requires the definition of a reference state that represents an average interaction of the system. Typically, a null interaction model is assumed to calculate such an average. Two main reference states have been proposed and used: the uniform density[11,18,19] and the distance-scaled finite ideal-gas.[5] In the uniform density model, the total number of pairs in any given distance shell for a reference state will be the same as that for folded proteins. Therefore, this reference state model could not necessarily constitute a truly non-interacting ideal-gas system. In the distance-scaled finite ideal-gas model, the protein atoms are treated as non-interacting and uniformly distributed points in finite spheres.

Irrespectively from the reference system used to calculate a scoring function, it is important to mention that it constitutes a critical feature that will largely determine its successful application.

The best solution to the problem of the reference system definition would be to have a representative and unbiased experimental sample of the unfolded state of proteins (see the discussion in Section 3.7).

3.4 How is a Scoring Function Used?

Scoring functions can be either used to calculate a single overall score of a protein conformation or a detailed score per residue. When a protein conformation is evaluated with a scoring function, the list of relevant geometries between bodies must be built first. Then, each observed geometry will have a score term associated with it. For example, in the case of the distance-dependent scoring functions, all body-body distances are calculated and their corresponding scores obtained based on the specific body pair and the distance that separates them in three-dimensional space. These individual scores can be added to obtain a total sum for the complete protein or added for each amino acid independently. The utilization of overall or detailed scores depends on the particular application of the scoring function and is described in more detail next.

3.4.1 *Classical Overall Score Calculation*

The overall score calculation consists of adding up all observed terms corresponding to specific geometries in a protein that fulfill the restraints of the scoring function. In the case of scoring functions at the atomic level, the resulting score can be normalized by the total number of contributing terms or simply kept as a raw score. For scoring functions at the amino acid level, more sophisticated normalization schemes such as the calculation of a z-score can be adopted, which is useful to compare scores between proteins that are different in size, structure, and amino acid composition.[6]

The overall z-score of a protein is calculated based on the total score of the particular protein conformation and also in a distribution of scores obtained for a set of different proteins of the same size and composition as the protein being evaluated. The general mathematical expression of the z-score is:

$$Z = \frac{S - \mu}{\sigma} \tag{3.6}$$

where Z is the z-score, S the total score of the protein, μ the average score, and σ the standard deviation of the scores obtained for a set of different proteins.

Two different approaches can be used to calculate the overall z-score of a protein with scoring functions at the amino acid level. The first uses as a reference frame the sequence space of proteins. The second is based on using distinct known native protein structures as a reference frame.

3.4.1.1 *Sequence space reference frame*

The overall z-score of a protein is here calculated using the total score of the particular protein conformation and a distribution of scores obtained from a set of random proteins of the same size and composition as the protein being evaluated (Fig. 3.6). The sequence of the protein is randomized hundred or thousand times and threaded into the same conformation, thus generating many random proteins with the same structure. By using this normalization model, the z-score gives a reference value in relation to the score expected by chance for a protein of this size and amino acid composition that adopts this conformation.[9]

3.4.1.2 *Structure space reference frame*

Instead of using a single conformation and many different sequences of the same size and amino acid composition, the z-score could also

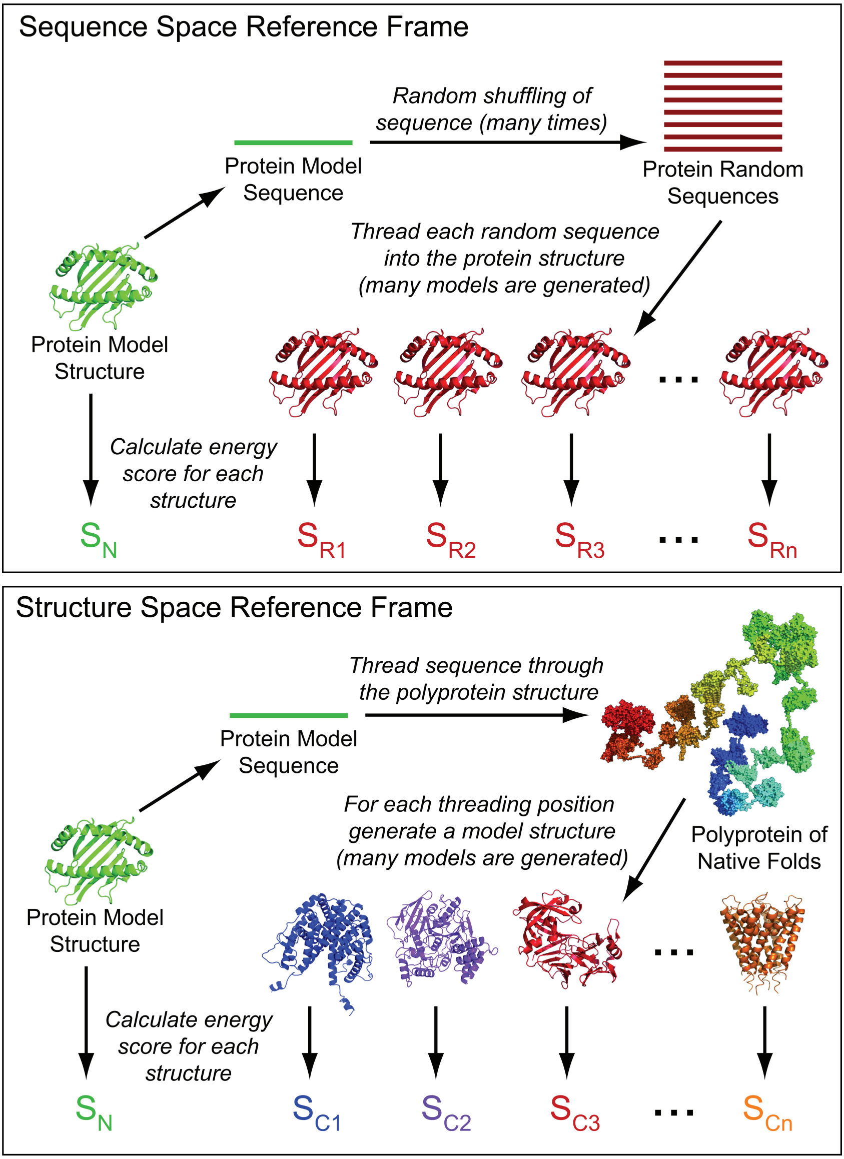

Sequence Space Reference Frame
Random shuffling of sequence (many times)
Protein Model Sequence
Protein Random Sequences
Thread each random sequence into the protein structure (many models are generated)
Protein Model Structure
Calculate energy score for each structure
...
S_N
S_{R1}
S_{R2}
S_{R3}
...
S_{Rn}
Structure Space Reference Frame
Thread sequence through the polyprotein structure
Protein Model Sequence
For each threading position generate a model structure (many models are generated)
Polyprotein of Native Folds
Protein Model Structure
Calculate energy score for each structure
...
S_N
S_{C1}
S_{C2}
S_{C3}
...
S_{Cn}

Fig. 3.6 Reference frames for the calculation of z-scores. In the top panel, several random (Rn) protein models with the same conformation are built, their scores calculated and used as a reference distribution for the calculation of the z-score. In the bottom panel, many alternative native conformations (Cn) are tested for a single protein sequence.

be calculated based on a single sequence that adopts many different conformations (Fig. 3.6). This approach requires an artificial polyprotein that consists of a set of distinct or non-redundant known native structures connected together.[1] The sequence of the protein being evaluated is threaded at different starting positions through the polyprotein, thus generating many alternative native conformations for the same protein sequence. By using this normalization model, the z-score gives a reference value in relation to the score expected for a given protein when adopting known native conformations.

3.4.2 *Classical Detailed Score Calculation*

In addition to the overall score, there are some applications that require a more detailed score calculation such as the total score value for each amino acid. These include the detection of errors in particular regions of a protein structure or the prediction of mutation effects in protein stability. For example, the detection of particular residues or local regions with structural errors is typically carried out by means of smoothed and normalized score profiles. The normalized total score per residue (S^R) is defined as follows:

$$S^R = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} \Delta S_k^{ij}(d)}{T^R} \tag{3.7}$$

where N is the total number of i atoms belonging to a given residue, and M is the total number of atoms j that interact with atom i at sequence separation k and below the distance range defined in the scoring function. $\Delta S_k^{ij}(d)$ corresponds to the score assigned by the scoring function to the interaction between atoms i and j, at sequence separation k and at distance d. T^R is the total sum of i,j interactions recorded for residue R and lies in the range $0 \leq T^R \leq N \times M$. The normalized score profiles can also be further smoothed by a sliding window with a fixed length of residues (typically in the range 5–15 residues). These normalized and smoothed score profiles can be finally used to locally assess the structural errors in a protein structure.

3.4.3 *Variations on Score Calculations*

Though most of the existing scoring functions are used with the same set of parameters that were adopted to derive them, in some particular cases it would be more convenient to modify this scheme.[20–22] The scoring function is derived with some parameters and used with a different set of parameters due to some limitations about the existing experimental data and the methodologies used to derive them. These limitations mostly arise because there is a lack of unbiased experimental data representing the unfolded state of proteins and also because of atom-atom connectivity issues, which result in some assumptions clearly not being valid (i.e. different atom-atom interactions being independent among them).

An example of this hybrid approach consists of the derivation of a scoring function based only on the non-local interactions and its subsequent utilization to calculate the scores of the non-bonded interactions.[20,21] Another example is to derive a scoring function for all interactions occurring up to a given distance range, and then using it to calculate only the effective interactions that are not shielded by other atoms.[22] These new approaches exhibit an improved performance at discriminating between native and near-native protein conformations.

It has been proposed, but not yet demonstrated, that these hybrid derivation/utilization approaches lead to a minimization of the existing mutual information among atom pair interactions, thus leading to scores that relate more closely to the true energies of a protein-solvent system.[21,22]

3.5 Typical Applications of Scoring Functions in Protein Structure Prediction

Many different applications of scoring functions in protein structure prediction have been described. These include evaluating whether or not a given protein model has the correct fold (i.e. fold assessment), picking the most accurate model out of many alternative models (i.e. detection of native-like protein conformations), estimating the overall geometrical accuracy of a model (i.e. model ranking), estimating the geometrical accuracy of the individual regions of the model (i.e. error detection), and prediction of protein stability (i.e. mutant screening). Next, we refer to these in more detail and provide examples for some of them.

3.5.1 *Fold Assessment*

Fold assessment consists of assessing whether or not a given protein model has the correct fold and constitutes the first step in protein structure assessment. For example, a protein model that has been built by comparative modeling will have the correct fold if the correct template was picked and if the template was aligned at least approximately correctly with the target sequence.[23]

Residue-based combined accessible surface and distance-dependent scoring functions have shown the best performance for this particular task.[9] Therefore, a combined z-score is typically calculated for a protein structure model, and if the z-score is below a fixed and optimized threshold, the correct fold is predicted. Otherwise, the protein model is rebuilt after modifying several input parameters of the protein structure prediction software, and the fold assessment process is carried out again. Accurate fold assessment is particularly useful in large-scale protein structure prediction and also in predicting the structure of proteins that do not exhibit a clear sequence similarity to known protein structures (i.e. fold recognition and *ab initio* protein structure prediction).

3.5.2 *Model Ranking*

The model ranking problem consists of sorting many different protein conformations according to their expected accuracy or deviation from a native structure. This is often necessary when many models for the same target protein are built, which is typically the case in comparative modeling. Several models can be generated by subtle changes on the input sequence-structure alignment, by selecting a different template structure or by different runs of molecular dynamics and energy minimization applied to some hypervariable regions in a protein (i.e. typically the loop regions that result from insertions or deletions in the template structure).

Though the ranking of models constitutes a difficult challenge, atom-based distance-dependent scoring functions have proved to be useful for this particular task in some cases.[18,24] However, the correlation between scores and structural deviation can still be significantly improved. It seems that more complex multi-variate scoring functions will be required in order to improve the results in this particular application (see Sec. 3.7.6 below).

3.5.3 *Error Detection*

The detection of localized protein structure errors is an important problem in protein structure prediction. The ability of a scoring function to detect wrongly modeled regions constitutes the first requirement to be fulfilled in order to improve the prediction of these regions. In accurate comparative models, localized errors are typically found at the loop regions. Unfortunately, these regions often determine some key aspects such as the substrate selectivity of enzymes or the specificity of antigen binding in antibodies. Detection of small and localized errors can also be important to assess the quality of experimentally solved protein structures.

The most successful scoring functions for tackling this problem are atom-based distance-dependent functions. The detection of local errors is typically carried out by the calculation of score profiles and the subsequent selection of local clusters of amino acids that are found above a certain score threshold.[1,12] The major difficulty to improve the correct detection of wrongly modeled regions arises because these zones are normally interconnected in three-dimensional space (i.e. they interact with each other). Scoring functions with a shorter distance range and a stronger local component are less sensitive to this problem and thus perform better.[20]

3.5.4 *Folding and Molecular Simulations*

Scoring functions can also be used for molecular dynamics and energy minimization simulations. Though they are discrete functions, through direct interpolation from the data a first derivative can be easily obtained. Despite of this possibility, scoring functions have not been extensively used for energy minimization and molecular dynamics simulations. One successful application of scoring functions on this subject has been carried out for the modeling of loop conformations.[25] In this work, a non-local scoring function[12] was combined with CHARMM-22 molecular mechanics force field[26,27] by replacing Lennard-Jones and Coulomb non-bonded terms from CHARMM by the complete set of terms from the scoring function. Therefore, the 1–2 bond, 1–3 angle, and 1–4 dihedral and improper dihedral energies were obtained from CHARMM; and the 1–5 and above non-bonded pseudo-energies were obtained by cubic spline interpolation from the discrete scoring functions. Thus, all 1–5 or higher non-bonded interactions were obtained from a single pseudo-energy function that was derived non-locally for each atomic pair. Contributions from both potentials, as well as residue side chain dihedral angle pseudo-energies derived from the observed statistical preferences in experimental data,[28] were equally weighted and combined to get the total pseudo-energy of the system. It turned out that this approach resulted in the most accurate predictions of loop conformations.[25] These results suggest that more effort at incorporating scoring functions not only to assess some fixed or static conformations of proteins, but also to generate or to predict them through molecular dynamics simulations, should be attempted.

3.6 Other Applications of Scoring Functions

The scoring functions described in this chapter are not necessarily restricted to the particular application of protein structure prediction. Just to name some, we have recently been successfully applying this kind of scoring functions to other relevant problems that include: (1) Sequence-based fold assignment of proteins in absence of sequence similarity to known protein structures, (2) sequence-based gene prediction, (3) sequence-based prediction of structure in RNA molecules (i.e. stable secondary structure formation and internal ribosome entry sites existence), (4) structure-based ligand binding site prediction, and (5) text-based assignment of authorship of anonymous or disputed documents.

3.7 Future Outlook

There are many aspects in the derivation and use of scoring functions that could be improved. In this section, we proceed to briefly describe those, which, in our opinion, are the most relevant.

3.7.1 *Reference Systems and Atom Type Definitions*

Scoring functions rely on a reference system or reference state to calculate the difference of score between two states (i.e. the folded and unfolded state). Unfortunately, the reference state is also calculated from the same dataset of native protein structures solved by experiment and do not necessarily correspond to the real probabilities that exist in the unfolded state. The reason for doing this is that an unbiased and representative experimental data source for the unfolded state of proteins is not yet available. Both the uniform density or the distance-scaled finite ideal-gas reference states cannot explicitly deal with some relevant issues such as atom-atom connectivity side effects and the independence of pair-wise interactions, which become extremely important for non-bonded interactions of a short-distance range.[20,21]

As it has been recently proposed, one possibility to minimize this problem is to derive scoring functions only for the effective atom-atom interactions.[22] The effective scoring functions show a significant improvement in the difficult challenge of discriminating between native and near-native conformations. Another option would be to improve the atom type definitions, in order to minimize the mutual information between different atom pairs.[7] Scoring functions are informatic functions built under the assumption that all interactions are independent, which is clearly not true in some cases such as in the close non-bonded interactions and the long-distance range interactions. Most of the problems arise due to atom-atom connectivity issues. The use of effective interactions could be a solution for the derivation of long-distance range scoring functions and a proper definition of atom types could aid in overcoming the observed dependency effects among different atom pairs.

3.7.2 *Solvation Models*

Most of the described solvation models obtained by knowledge-based approaches are very coarse-grained. They only provide a general description of the burial propensity by a single term. More strictly, solvation effects should be better described by a scoring function containing pair-wise terms between solvent and protein atoms. However, the modeling of water atoms in this hypothetical explicit solvent model would be a difficult task. We propose that a combined explicit-implicit solvent model should be developed. A possible strategy for doing this will be the following: first, an effective solvation propensity needs to be derived for each protein atom from experimental data of native proteins. This would be possible based on the fact that known protein structures contain many solvent molecules. Second, an estimation of the total number of expected interactions between a protein atom and the solvent could be derived based on the exposure of a protein atom. Though this would be more difficult to calculate directly and accurately for all atom types, the problem could be initially addressed for polar atoms and then extrapolated to the non-polar ones. With these two ingredients, the solvation score of a particular protein conformation could be calculated in a pair-wise fashion without explicitly considering the solvent molecules (i.e. just by multiplying the expected number of water molecules surrounding a protein atom based on its exposure by the solvation propensity score of that protein atom). This strategy would allow many solvation terms to be obtained for each protein atom found at the surface, thus eliminating the need for normalization of accessible surface terms and pair-wise terms in the combined scoring functions. In this scenario, the amount of information obtained from both scoring functions would be properly balanced: (1) Buried atoms would be assessed based on the pair-wise protein terms, (2) exposed atoms would be assessed based on their interaction propensity with solvent molecules, and (3) intermediate buried atoms would be assessed by a balanced sum of the terms from the two scoring functions.

3.7.3 *Evolutionary Information*

Scoring functions are currently derived from structural data alone, totally neglecting experimental data about protein sequences. However, the total number of available sequences constitutes a rich source of information about homologue proteins to known experimental structures that are able to fold. Based on the findings of Chothia and Lesk in 1986, we know that structure is more conserved than sequence, and thus more or less accurate structural models could be built for many of the available sequences by comparative modeling. These models could be then used to derive residue-based scoring functions for each specific structural space, thus capturing not only the thermodynamic contributions of the protein folding, but also some kinetics constraints of the process that are present in a particular fold space.[29] The proper knowledge and usage of this kind of "evolutionary scoring functions" could be very useful for protein structure prediction, and particularly for fold recognition and protein design. It can also have a positive impact in the structural genomics project by aiding the selection of new protein folds.[30]

3.7.4 *Multivariate Scoring Functions*

There are many problems that cannot be properly tackled with single scoring functions. Among these we found the fold assessment of short and incomplete protein models.[9,31] Most large-scale protein structure efforts produce many short and incomplete protein models.[32] The major problem of assessing this kind of protein models with current scoring functions resides in the fact that they are artificial and thus not found in nature in that way (i.e. an accessible surface scoring function will poorly assess them). However, on the other hand, these types of protein models still contain some useful information for functional inference (i.e. structural matching of functional template sites) and/or for the rational design of experiments. Therefore, in this case, a multivariate scoring function is needed, which should be capable of selecting or weighting properly the individual terms to give an accurate assessment. Some multivariate scoring functions have already been described for this particular problem, which have proved to be significantly more accurate in the assessment of the difficult cases described above.[31,33,34]

Acknowledgements

Due to space constraints, citations in this chapter to the previous work done by others are not as extensive as intended. In many cases we have selected just a few previous publications in each covered subject. The selection of the citations in this chapter has been done sometimes based on the impact of old work, and at other times, priority has been given to novel and recent work. We acknowledge everyone who has contributed in general to the development of scoring functions, and also particularly those who have worked on the subject but with a focus on the specific field of protein structure prediction. FM would like to dedicate this work to Manfred Sippl for his enormous contribution to the field, and also to Evandro Ferrada for the many illuminating discussions on the topic. This work has been supported in part by grant 1051112 from FONDECYT to FM.

References

1. Sippl MJ. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**: 355–362.
2. Kihara D, Zhang Y, Lu H, Kolinski A, Skolnick J. (2002) *Ab initio* protein structure prediction on a genomic scale: application to the Mycoplasma genitalium genome. *Proc Natl Acad Sci USA* **99**: 5993–5998.
3. Jones DT, Taylor WR, Thornton JM. (1992) A new approach to protein fold recognition. *Nature* **358**: 86–89.
4. Casari G, Sippl MJ. (1992) Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native proteins. *J Mol Biol* **224**: 725–732.
5. Zhou H, Zhou Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* **11**: 2714–2726.

6. Sippl MJ. (1993) Boltzmann principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aid Mol Des* 7: 473–501.
7. Solis AD, Rackovsky S. (2006) Improvement of statistical potentials and threading score functions using information maximization. *Proteins* **62**: 892–908.
8. Melo F, Marti-Renom M. (2004) Accuracy of sequence allignment and fold assessment using reduced amino acid alphabets. *Proteins* **63**: 986–995.
9. Melo F, Sanchez R, Sali A. (2002) Statistical potentials for fold assessment. *Protein Sci* **11**: 430–448.
10. Miyazawa S, Jernigan RL. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**: 534–552.
11. Sippl MJ. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**: 859–883.
12. Melo F, Feytmans E. (1998) Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol* **277**: 1141–1152.
13. Melo F, Feytmans E. (1997) Novel knowledge-based mean force potential at atomic level. *J Mol Biol* **267**: 207–222.
14. Berman HM, Battistuz T, Bhat TN, *et al.* (2002) The Protein Data Bank. *Acta Cryst D* **58**: 899–907.
15. Chothia C, Lesk AM. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**: 823–826.
16. Holm L, Ouzounis C, Sander C, Tuparev G, Vriend G. (1992) A database of protein structure families with common folding motifs. *Protein Sci* **1**: 1691–1698.
17. Marti-Renom MA, Ilyin VA, Sali A. (2001) DBAli: a database of protein structure alignments. *Bioinformatics* **17**: 746–747.
18. Samudrala R, Moult J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol* **275**: 895–916.
19. Lu H, Skolnick J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* **44**: 223–232.
20. Ferrada E, Melo F. (2007) Non-bonded terms extrapolated from non-local knowledge based energy functions improve error detection in near native protein structure models. *Protein Sci* **16**: 1410–1421.
21. Ferrada E, Vergara IA, Melo F. (2007) A knowledge-based potential with an accurate description of local interactions improves discrimination between native and near-native protein conformations. *Cell Biochem Biophys* **49**: 111–124.
22. Ferrada E, Melo F. (2007) Knowledge-based energy functions and effective atomic interactions. *Protein Sci*, **submitted**.

23. Marti-Renom MA, Stuart A, Fiser A, Sanchez R, Melo F, Sali A. (2000) Comparative protein structure modeling of genes and genomes. *Ann Rev Biophys Biomol Struct* **29**: 291–325.
24. Samudrala R, Levitt M. (2000) Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci* **9**: 1399–1401.
25. Fiser A, Do RK, Sali A. (2000) Modeling of loops in protein structures. *Protein Sci* **9**: 1753–1773.
26. Brooks B, Bruccoleri R, Olafsonand B, *et al.* (1983) CHARMM: a program for macromolecular energy, minimizations and dynamic calculations. *J Comput Chem* **4**: 187–217.
27. MacKerell AD, Jr., Bashford D, Bellott M, *et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102**: 3586–3616.
28. Sali A, Blundell TL. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**: 779–815.
29. Panjkovich A, Melo F, Marti-Renom MA. (2007) Evolutionary potentials: structure specific potentials exploiting the evolutionary record of sequence homologs. *Protein Sci.*
30. Baker D, Sali A. (2001) Protein structure prediction and structural genomics. *Science* **294**: 93–96.
31. Melo F, Sali A. (2007) Fold assessment for comparative protein structure modeling. *Protein Sci* **16**: 2142–2426.
32. Pieper U, Eswar N, Braberg H, *et al.* (2006) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucl Acids Res* **33**: 291–295.
33. Jones DT. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* **287**: 797–815.
34. McGuffin LJ, Jones DT. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19**: 874–881.

This page intentionally left blank

Chapter 4

Assessment of Protein Structure Predictions

E. Capriotti and M. A. Marti-Renom*

4.1 Introduction

Since the beginning of the 1980s, protein structure prediction and simulation have been one of the most challenging tasks for computational structure biology. Although progress has been made, one can openly say that reliably predicting the fold of all known protein sequences is still far from reach. Current approaches can predict a three-dimensional (3D) protein structure for parts of ~60% of the sequences of an average genome.[1–3] Recently, automatic large-scale predictions of 3D structure models are being made available on the web. For example, the ModBase database[2] currently stores more than 4.2 million models, the SwissModel repository[1] stores ~1.3 million models, and the PMDB database[3] stores ~75 000 models. Therefore, comparative protein structure modeling is filling the gap between the known sequence and structure spaces.

In the post-genomic era, a more difficult task lies ahead in annotating, understanding, and modifying the function of proteins. This task is greatly aided by the knowledge of the protein structures, as the biochemical function of a protein is determined by its structure and

*Corresponding author.

Structural Genomics Unit, Bioinformatics Department, Prince Felipe Research Center, Avda. Autopista del Saler, 16 , 46013 Valencia, Spain. Email: mmarti@cipf.es.

dynamics. In the absence of an experimentally determined structure, 3D models are often valuable for rationalizing existing evidence and guiding new experiments.[4] However, the accuracy of a model determines its utility (Chapter 5), making a means of reliably determining the accuracy of a model an important problem in protein structure prediction.[4,5] Model assessment aims to predict the likely accuracy of a protein structure model in the absence of its known 3D structure.

Model assessment has been previously applied to: (i) determine whether or not a model has the correct fold,[6–9] (ii) discriminate between the native and near-native states,[10–19] and (iii) select the most near-native model in a set of decoys that does not contain the native structure.[16–18,20–23] Several scoring schemes have been developed for these tasks, including physics-based energies, knowledge-based potentials, combined scoring functions, and clustering approaches. Physics-based energy functions are true energy functions describing the interactions acting upon all atoms in a protein structure and are typically developed for and used in molecular dynamic simulations. Statistical or knowledge-based potentials are derived from known protein structures by applying the inverse of the Boltzmann's equation and comparing a system in the thermodynamic equilibrium with the database of folded protein structures. Combined scoring functions usually integrate several different scores with the aim of extracting the most informative features from each of the individual input scores. Finally, the so-called clustering approaches use consensus information from an ensemble of protein structure models provided by one or more methods.

We begin this chapter by introducing the problem of protein structure prediction (Chapter 1). Next, we describe the four main approaches to model assessment. Details describing some of the most widely used scoring function are also provided together with a table of Internet resources for model assessment (Table 4.1). Finally, some of the results from the recent evaluation of model assessment methods carried out at the seventh Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment are introduced before a final outlook of the future of model assessment.

Table 4.1 A List of URLs to Some Relevant Internet Resources

Title	Refs.	URL
DECOY SETS		
Decoys "R" Us	72	http://dd.compbio.washington.edu
RAPPER	73	http://mordred.bioc.cam.ac.uk/~rapper/decoys.php
Skolnick lab	17	http://cssb.biology.gatech.edu/skolnick/files/all-atom/
ROSETTA	N/A	http://www.bakerlab.org
Sali lab	N/A	http://www.salilab.org
PHYSICS-BASED ENERGIES		
CHARMM	26	http://www.charmm.org
AMBER	25	http://amber.scripps.edu
GROMOS	74	http://www.igc.ethz.ch/gromos/
KNOWLEDGE-BASED POTENTIALS		
VERIFY3D	75	http://nihserver.mbi.ucla.edu/Verify_3D/
TAP	58	http://protein.cribi.unipd.it/tap/
FRST	76	http://protein.cribi.unipd.it/frst/
ANOLEA	77	http://protein.bio.puc.cl/cardex/servers/anolea/
DFIRE	41	http://sparks.informatics.iupui.edu/hzhou/dfire.html
PROSA-Web	78	https://prosa.services.came.sbg.ac.at/prosa.php
PROQ	21	http://www.sbc.su.se/~bjornw/ProQ/ProQ.cgi
SIFT	79	http://sift.cchmc.org
HOPPscore	80	http://hoppscore.lbl.gov/run.html
HARMONY	81	http://caps.ncbs.res.in/harmony/

4.2 Protein Structure Prediction

The aim of protein structure prediction is to build a 3D model for a protein of unknown structure (target) either using *ab initio* methods (i.e. template-free approaches) or on the basis of sequence similarity to proteins of known structure (i.e. template-based approaches such as comparative modeling or threading). Chapter 1 in this book provides a comprehensive introduction to protein structure prediction.

Since the accuracy of a protein structure model determines its usability,[4] two basic conditions must be met to build a useful 3D model. First, an accurate model needs to be built based on the correct template and approximate correct alignment. Second, a reliable score for the model has to be computed to assess its accuracy. Thus, the aim of the second step is to predict errors in models produced in the first step. Next, we outline some of the typical errors in protein structure models. The first two types or errors are specific of template-based approaches while the rest also apply to template-free approaches:

Template selection. The initial step in template-based protein structure prediction is the selection of a template structure. Although selecting the incorrect template is a major error affecting models based on very low sequence identity to their templates (i.e. under ~25% sequence identity), current model assessment methods are usually able to reliably detect it.

Misalignments. One of the largest sources of errors in models from template-based approaches is the incorrect alignment between the target and the template sequences. Such errors affect models based on ~40% or less sequence identity to the closest template(s). The use of multiple sequence alignments, multiple templates and iterative model-building and target-template alignment modification may alleviate such errors.

Template-free modeling. Segments of the target sequence that have no equivalent region in the template structure (i.e. whole protein for template-free modeling or insertions in template-based modeling) are the most difficult regions to model.

Rigid body shifts. As a consequence of sequence divergence there is a natural diversity between two homologous sequences. One type of structural diversity is the rigid distortion of parts of the models. The use of multiple templates may reduce such error.

Side chain packing. The correct packing of side-chain atoms is essential for high-resolution modeling where the resulting models may be used for docking of small molecules. Therefore, methods for predicting the detailed accuracy of a model are becoming even more important in the advent of a large number of determined structures and the use of models for the docking of small molecules.[24]

Fortunately, during the last decade, the development of more accurate fold assignment and target-template alignment methods together with the use of multiple sources or structural information are mitigating the errors in protein structure models.

4.3 Model Assessment

Protein structure model assessment addresses the general question of how accurate a model is. More specialized questions include: (i) evaluating whether or not the model has the correct fold, (ii) selecting the most accurate model from a set of decoys or alternative solutions, (iii) estimating the overall accuracy of a model (i.e. defining a score that correlates with the RMSD after superimposing a model and its native structure), and (iv) estimating the accuracy of different regions in a model. In the next sections we introduce the four types of available approaches for model assessment, which are used to address some or all of the problems mentioned above: physics-based energies, knowledge-based potentials, combined scoring functions, and clustering approaches. Table 4.1 provides a list of relevant accessible Internet resources.

4.3.1 *Physics-based Energies*

Molecular mechanics energy functions with solvation models are the usual components of physics-based energies. Generally speaking, chemical force fields are functional forms encoding a set of parameters for describing the energy of a system of particles. The function and the parameters describing a force field are usually derived both from experimental observations and quantum mechanical calculations. A basic representation of a force field energy function depends on two main contributions: a term describing the energy from chemical bonds between the atoms in the system and a term describing the interactions between non-bonded atoms in the system. The first term depends on the distances, angles, and dihedral angles between bonded atoms in the molecule. The second term depends on the electrostatic and van der Waals interactions between non-bonded atoms in the molecule. Such energy scores have been classically developed as part of molecular mechanics simulation packages such as AMBER,[25] CHARMM,[26] MM-PBSA,[27] or GROMOS.[28] However, some physics-based approaches, which are outlined next, have also been used for ranking structural decoys.[10,17,19,29–32]

Lazaridis and Karplus applied an effective energy function (EEF1) combining the CHARMM 19 force field with a Gaussian model for solvation free energy to discriminate native structure on a dataset of 650 decoys for six proteins.[10] The results showed that the native state was always more stable than any of the misfolded structures and molecular dynamics simulation reduced the free energy gap between near-native and misfolded structures.

The all-atom version of the Optimized Potential for Liquid Simulations (OLPS),[33] combined with the Surface Generalized Born (SGB) method, was used to discriminate near-native conformations in a set of 49 000 minimized decoy structures for 32 proteins.[18] This energy function was able to correctly identify the native structure within the decoy set in 70% of the tested proteins. The analysis also highlighted the contribution of the solvation free energy in the detection of the native-like structure.

A Molecular Mechanics-Poisson Boltzmann Solvent Accessible Surface Area (MM-PBSA) model was recently used to calculate the free energy of a protein loop structure as a surrogate of the similarity of the decoy to its native structure.[27] The results from such simulations indicated that the MM-PBSA free energy estimator was able to detect native-like structures for 81% of the decoy sets. Moreover the use of the colony energy approach[34] reduced the MM/energy dependency on minor conformational changes. Thus, the authors were able to correlate free energy scores with the root mean square deviation (RMSD) of a decoy set with respect to the native structure.

Recently, Maupetit and colleagues[22] proposed a coarse-grained optimized potential for efficient structure prediction (OPEP). Their method was able to detect native conformations in 83% of the 29 test proteins with more than 28 000 decoy structures.

Finally, an updated version of the AMBER force field[17] with terms representing the solvation contribution was tested for its ability in identifying near-native structures for 150 target proteins within a set of 14 000 decoy structures. The authors concluded that the ability of the method for identifying near-native structures in the decoy set decreased with the time of molecular simulation of the decoys. This version of AMBER was able to detect 100% of near-native structures after only minimizing the structural decoys (i.e. with no molecular mechanics simulation). However, the accuracy decreased to ~70% after a small simulation of 200 picoseconds and to ~30% for 2 nanoseconds simulations. Therefore, such results indicate that molecular mechanics force fields are able to identify near native structures but cannot drive the simulation towards the native conformation of the protein. The authors of the study also concluded that the native structure often does not appear to be in the lowest free energy state.[17] As of today, the refinement problem (i.e. the ability to move the coordinates of a protein structure prediction towards its native conformation) has no generally applicable solution.

In summary, physics-based scoring functions provide good means for selecting near-native structure models in a set of predicted decoys. The introduction of solvation terms clearly improved the ability of such force fields to discriminate between near-native and no-native conformations. However, a universal energy function for model refinement is still far from reach and the relative weight for each energy term contribution may need to be optimized for each decoy set under consideration.

4.3.2 *Knowledge-based Potentials*

Statistical potentials, also called potentials of mean force, constitute the main implementation of the knowledge-based potentials for model assessment. In general terms, such potentials encode the statistical preferences of different residues or atom types to be exposed to the solvent, or to interact with each other in a pair-wise or higher order fashion. Such preferences are normally extracted from a set of selected structures, which represent the known structural space for globular proteins. The basic hypothesis is that protein crystal structures contain a large amount of information describing the stabilizing forces of protein folding, which can be extracted by using the following three assumptions of statistical mechanics: (i) protein folding can be described by a free energy function, (ii) the conformation of a system can be approximated by two-body interactions, and (iii) high frequency conformations should correspond to low free energy structures. If such assumptions are true, it is then possible to derive an atomic energy function for which the global minimum corresponds to the observed native crystal structure.

Since the end of the 1970s, several authors have used such approximations to derive statistical rules from known protein structures.[8,35–45] The main characteristic shared by most knowledge-based potentials is the use of the inverse Boltzmann distribution to derive pseudo-energies from a non-redundant set of protein structures, which states that the probability ($p(x)$) of state x with energy $\varepsilon(x)$ is:

$$p(x) = \frac{1}{Z} e^{-\varepsilon(x)/kT} \tag{1.1}$$

where k is the Boltzmann's constant and T is the absolute temperature. The partition function Z, which can be considered the ground state energy, is defined as:

$$Z = \sum_{x} e^{-\varepsilon(x)/kT} \tag{1.2}$$

Thus, a general representation of the energy function is:

$$\varepsilon(x) = -kT \log\left(\frac{p(x)_{obs}}{p(x)_{exp}} \right) \tag{1.3}$$

where $p(x)_{obs}$ and $p(x)_{exp}$ are the observed and expected occurrences of the state x respectively. The inverse of the Equation (1.3) is then the pseudo-energy score of the knowledge-based potential, which calculates the energy relative to state x ($\varepsilon(x)$) using the distribution function $p(x)$:

$$\varepsilon(x) = -kT \log p(x) - kT \log Z \tag{1.4}$$

Although there has been debate about the physics basis of statistical potentials,[46–48] it is assumed that the database of protein structures represents the conformational space of globular proteins in thermodynamic equilibrium.

Several types of statistical potentials have been derived which assess different structural features of models. Such potentials include contact,[8,23,37] distance,[16,40,41,45,49] solvent accessibility,[8,42,50] and a combination of solvent accessibility and pair-wise interaction.[16,41,44,45,49,51] Next, we summarize a few particular implementations and applications of knowledge-based potentials for model assessment. Our list is not exhaustive nor complete, but it highlights different approaches for model assessment using knowledge-based potentials. For a recent evaluation and reviews of such methods, see Section 4.5 within this chapter and references.[52,53]

Although significant work was done beforehand, knowledge-based potentials became more widely used after the work of Sippl in the beginning of the 1990s.[42,54] Sippl's PROSA, a Cα/Cβ distance-dependence potential that used a poly-protein of 230 different folds for calculating the final Z-score of a model, was originally benchmarked using a set of 163 protein structures. The author concluded that such potentials were accurate for detecting the native structure for most available globular proteins. A new atomic-level statistical potential based on atom-type definitions was later developed by Melo and Feytmans.[49] Using such an approach, it was possible to obtain average frequencies of pair-wise contacts about 15 times higher than the ones obtained using reduced representations for each amino acid. Similarly, Samudrala and Moult[45] developed a residue specific all-atom probability discriminatory ratio (RAPDF), which resulted in a better discrimination of native models compared to other simplified protein representations and illustrated the importance of using a detailed atomic description of the system. Similar conclusions were obtained by Lu and Skolnick[16] using their heavy-atom potential for discriminating native from near-native structures in a set of decoys. The authors pointed out that their atomic potential tended to pick lower RMSD structures being able to discriminate the native structure in 87% of 119 protein decoy sets. A significant improvement of such atomic-based potentials was later obtained by using a mathematical programming approach.[55] Qiu and Elber compared the performances of their potential with other existing methods, concluding that their potential reached similar accuracy using a much smaller number of parameters.

More recently, the development of alternative reference states (i.e. $p(x)_{exp}$ in Equation (1.6) for assessing random interactions has lead to significant improvements in the final accuracy for assessing a protein structure model. For example, a distance-scaled finite ideal-gas reference state was used to derive the DFIRE potential.[41] On average, DFIRE all-atom potential identified the native structure for 84% of the 32 decoys sets used in its benchmark. In a subsequent work, Zhou and co-workers showed that a reduced description of the original DFIRE potential resulted in a similar success rate as its all-atom potential for ranking native structures in a benchmark of 96 decoy sets.[56] Similarly, the DOPE potential,[44] which uses a reference state based on non-interacting atoms in a homogeneous sphere with the radius dependent on a sample native structures, resulted in a higher accuracy than DFIRE, recognizing 87% of the native structures in the 32 decoy sets.

Finally, a new kind of knowledge-based potentials considering the relative orientation of different residue atoms has recently flourished.[57–60] Although the orientation-dependence potentials have not yet been extensively tested, their large-scale application together with coarse-grained protein representations could be very promising.

In summary, knowledge-based potentials, which use empirical observations of proteins of known structures, have proved useful for assessing the accuracy of protein structure models. The use of different reference states together with multi-body representations of protein structures may finally meet the needed accuracy for large-scale protein structure assessment.

4.3.3 *Combined Scoring Functions*

To improve the accuracy of methods for assessing the accuracy of protein structure models, several scoring functions have been developed using a weighted combination of individual scores from physics and/or knowledge-based approaches.[21,55,60–64] Such scores have been shown to increase the ability to discriminate incorrect models from correct models compared to their individual input scores.[64,65] However, combined scoring functions require the optimization of weights and parameters for each individual input score. As a result, the optimized scoring functions are very dependent on the training set of models used for their derivation. Next, we outline some such approaches developed in the last few years.

The ProQ program implements a neural-network that combines several structural features calculated from the assessed model.[21] Such features include: atom- and residue-based contact potentials, predicted and model secondary structure agreement, solvent accessible surface, fraction of modeled protein, Cα-Cα distance discrepancy between the model and the used template, and protein shape. ProQ was able to detect the correct protein structure model for 62% to 77% of several LiveBench decoy sets.[66]

A Support Vector Machine learning approach was implemented in the SVMod score.[64] SVMod was trained in regression mode taking into account different individual input scores including: three MODPIPE scores, two secondary structure agreement scores, and the DOPE all heavy atom score. The optimal SVMod score was able to select protein structure models on average ~0.45Å apart from the closest model to the native structure in a set of 300 protein structure decoys from 20 target proteins.

Similar to the work by Eramian and co-workers, Benkert and coworkers recently developed a linear combination of individual scores in the QMEAN program.[60] QMEAN combined a coarse-grained torsion angle potential, a secondary structure specific distance-dependent pairwise potential, a solvation potential and two terms accounting for the agreement between the model and predicted solvent accessibility and secondary structure from sequence. The QMEAN score was tested on a large set of 22 420 models of 95 target proteins from CASP.[67] QMEAN was favorably compared with other existing methods showing a statistically significant improvement in the detection of the native structure and in discriminating between correct and incorrect protein structure models.

In summary, combined scoring functions are able to capture particular structural features from models that may have been detected by each individual score. Therefore, such approaches leverage the input information towards the final goal of detecting the most accurate model in a pool of possible solutions.

4.3.4 *Clustering Approaches*

One of the most challenging tasks in protein structure model assessment is to devise a score that correlates with the actual accuracy of the model. One would hope that a perfect scoring function would assign favorable scores to models that are structurally similar to their native structure. Unfortunately, this is not usually the case, and current scoring functions, either physics-based, knowledge-based, or a combination of both, do not always favorably score models close to the native structure. However, when some correlation between the score and the model accuracy exists, structurally comparing all models from independent structure predictions of the same sequence may help in selecting the most accurate model in a set of possible solutions. In other words, an accurate scoring function should more often produce a structural conformation near the native structure than a misfolded structure. This hypothesis has recently been exploited in different implementations of the so-called clustering approaches.[20,68–70]

Shortle and co-workers first applied a clustering approach for predicting the accuracy of models from 10 small proteins in sets of 500 to 1000 ensemble models of low-energy.[20] The authors demonstrated that the conformation with the largest number of models within 4Å RMSD was closer to the native structure than were the majority of models from other clusters in the ensemble. The same approach was later efficiently used in the 4th CASP experiment.[69]

More recently, a new cluster-density method, which weighted the final score of a model using the mean RMSD of its conformation respect to the other ones in the decoy set, was implemented in the self-RAPDF method.[68] The results demonstrated that the use of the density scoring function increased the number of selected near-native conformations from 75% to 92% with respect to the RAPDF method.

A large-scale benchmarking of a clustering-based approach was recently carried out using the SPICKER strategy[70] over a 1489 decoy sets of up to 280 000 models generated by the TASSER program.[62] The results indicated that the top five identified conformations had RMSD values in the top 1.4% of all decoys. The results also indicated that for 78% of the 1489 target proteins, the difference in RMSD from their native conformation to the selected model and RMSD from native to the absolutely best individual model in the decoy was below 1Å.

In summary, the information from an ensemble of decoy conformations can be used to derive statistical probabilities, which facilitate the identification of near-native structures in a set of possible solutions. However, as it is evident from the conceptual implementation of clustering approaches, their final accuracy depends on the quality of the scoring functions used to generate the ensemble of conformations. In other words, an inaccurate scoring function will result in an inaccurate conformation selection by clustering. Another limitation of clustering approaches is their inability to assess the quality of a model on its own.

4.4 Evaluation of Model Quality Assessment Methods

In the seventh edition of the CASP experiment, a new category was introduced with the aim to blindly evaluate methods for model assessment.[52] The new category, named Model Quality Assessment, introduced two measures, evaluating assessment methods at the whole model and the reside levels.

A total of 23 864 models from 95 different target sequences were assessed by 28 different model assessment methods. The model quality assessors concluded that the Pcons program (65 and CASP7 special issue), which uses the ProQ method for model assessment, was able to constantly select models with near-native conformation. Although not with statistical significance, due to the limited number of groups participating in the residue-by-residue assessment category, the same method seemed able to identify reliable regions of models in a reasonable number of cases. However, as mentioned above, consensus-based methods, such as Pcons, rely in an ensemble of solutions to score individual models. Lee and co-workers (see CASP7 special issue) were able to identify good quality models based only on a non-relative score. Their strategy consisted of blindly relying on their own protein structure predictions. All assessed models were structurally compared with the model of the same target produced by their own method. The final assessment score corresponded to the structural distance of the assessed model to their model.

The strategy by Lee and co-workers resulted in good assessments because their models were consistently accurate for most of the targets in CASP7. Current methods can effectively select accurate models from a set of decoys or ensemble conformations. However, a substantial improvement of methods for evaluating regions of a model as well as assessing the absolute quality of a model on its own is still needed.

4.5 Future Outlook

Despite the large amount of sequence and structure information available and the ever increasing interest of the protein structure community in a reliable 3D structure assessment method, the quest for a perfect scoring function is still open.[53] Currently, the most reliable methods can reasonably select near-native conformations from a decoy set of possible conformations. However, three unsolved tasks lie ahead, to: (i) refine a protein structure model towards its native structure, (ii) reliably predict the absolute accuracy of a model, and (iii) identify regions or residues in a model most likely to contain errors.

During the past years, the most successful approaches for model assessment have relied on either combining individual scores

(Section 4.3.3) and/or clustering by structure similarity the resulting ensemble of predicted models (Section 4.3.4). However, such approaches do not add to our basic knowledge of the molecular mechanisms by which proteins adopt their native conformations. One can expect that imperfect individual score functions will hamper both, combined and clustering-based scoring functions. Therefore, a more reliable scoring function based on physics (Section 4.3.1) and/or empirical observations (Section 4.3.2) is clearly needed. Such a scoring function will have to address outstanding open problems such as:

Solvation. One of the major forces towards the folding of a protein is the initial hydrophobic collapse. However, none of the existing methods for protein structure simulation (including those for model assessment) are able to accurately model the effect of water on the protein structure.

Topological determinants. Characterizing the properties of short-to-medium range interactions is needed for elucidating the topological determinants of a protein fold.

Side-chain packing. Although successful methods for protein structure prediction may use a reduced representation of protein structures, which usually simplifies side-chains as single pseudo-atoms, the correct modeling of interaction in the core of proteins will be required for high-resolution protein structure prediction. Such level of details will also be needed for assessing the accuracy of models for protein-protein and/or protein-ligand interactions.

Protein structure flexibility and disorder. Current methods for model assessment do not include data about unstructured parts of proteins. Therefore, the use of such information will likely result in more accurate scoring functions for model assessment.

Small, multi-domain, and non-globular proteins. Most of the methods introduced here were developed to assess single domain globular protein models. Therefore, the average accuracy of such methods significantly drops when they are applied to very small or multi-domain proteins and to disordered or transmembrane proteins.

Detailed knowledge-based potentials. The rapid increase of structures deposited in the Protein Data Bank,[71] due in part to the Structural Genomics initiatives, allows the inclusion of multi-body terms in statistical potentials. Such detailed knowledge-based potentials are likely to result in more accurate methods for model assessment.

The increasing interest of the computational structural biologist in addressing such problems and the opportunity to blindly test their methods in an automatic, large-scale, and (hopefully) continuous manner will push the fields of model assessment and protein structure prediction towards very interesting and challenging times.

Acknowledgements

We are grateful to Pascal Benkert, Francisco Melo, and Anna Tramontano for providing access to their work prior to publication. Our research is supported by the European Union (Marie Curie Program FP6-039722), the Spanish Ministerio de Educación y Ciencia (BIO2007/66670), and the Generalitat Valenciana (GV/2007/065).

References

1. Kopp J, Schwede T. (2006) The SWISS-MODEL Repository: new features and functionalities. *Nucl Acids Res* **34**: D315–D318.
2. Pieper U, Eswar N, Davis FP, *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucl Acids Res* **34**: D291–D295.
3. Castrignano T, De Meo PD, Cozzetto D, Talamo IG, Tramontano A. (2006) The PMDB Protein Model Database. *Nucl Acids Res* **34**: D306–D309.
4. Baker D, Sali A. (2001) Protein structure prediction and structural genomics. *Science* **294**: 93–96.
5. Ginalski K, Grishin NV, Godzik A, Rychlewski L. (2005) Practical lessons from protein structure prediction. *Nucl Acids Res* **33**: 1874–1891.

6. Miyazawa S, Jernigan RL. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* **256**: 623–644.
7. Domingues FS, Koppensteiner WA, Jaritz M, *et al.* (1999) Sustained performance of knowledge-based potentials in fold recognition. *Proteins* (3): 112–120.
8. Melo F, Sanchez R, Sali A. (2002) Statistical potentials for fold assessment. *Protein Sci* **11**: 430–448.
9. McGuffin LJ, Jones DT. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19**: 874–881.
10. Lazaridis T, Karplus M. (1999) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J Mol Biol* **288**: 477–487.
11. Gatchell DW, Dennis S, Vajda S. (2000) Discrimination of near-native protein structures from misfolded models by empirical free energy functions. *Proteins* **41**: 518–534.
12. Vorobjev YN, Hermans J. (2001) Free energies of protein decoys provide insight into determinants of protein stability. *Protein Sci* **10**: 2498–2506.
13. Seok C, Rosen JB, Chodera JD, Dill KA. (2003) MOPED: method for optimizing physical energy parameters using decoys. *J Comput Chem* **24**: 89–97.
14. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. (2003) An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins* **53**: 76–87.
15. Zhu J, Zhu Q, Shi Y, Liu H. (2003) How well can we predict native contacts in proteins based on decoy structures and their energies? *Proteins* **52**: 598–608.
16. Lu H, Skolnick J. (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins* **44**: 223–232.
17. Wroblewska L, Skolnick J. (2007) Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. *J Comput Chem* **28**: 2059–2066.
18. Felts AK, Gallicchio E, Wallqvist A, Levy RM. (2002) Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. *Proteins* **48**: 404–422.
19. Dominy BN, Brooks CL. (2002) Identifying native-like protein structures using physics-based potentials. *J Comput Chem* **23**: 147–160.
20. Shortle D, Simons KT, Baker D. (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc Natl Acad Sci USA* **95**: 11158–11162.
21. Wallner B, Elofsson A. (2003) Can correct protein models be identified? *Protein Sci* **12**: 1073–1086.
22. Maupetit J, Tuffery P, Derreumaux P. (2007) A coarse-grained protein force field for folding and structure prediction. *Proteins* **69**(2): 394–408.

23. Park B, Levitt M. (1996) Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol* **258**: 367–392.
24. Allen KN. (2007) Form finds function. *Nat Chem Biol* **3**: 452–453.
25. Case DA, Cheatham TE, 3rd, Darden T, *et al.* (2005) The Amber biomolecular simulation programs. *J Comput Chem* **26**: 1668–1688.
26. Brooks BR, Bruccoleri RE, Olafson BD, *et al.* (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* **4**: 187–217.
27. Fogolari F, Brigo A, Molinari H. (2003) Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophys J* **85**: 159–166.
28. Soares TA, Daura X, Oostenbrink C, Smith LJ, van Gunsteren WF. (2004) Validation of the GROMOS force-field parameter set 45Alpha3 against nuclear magnetic resonance data of hen egg lysozyme. *J Biomol NMR* **30**: 407–422.
29. Lee MC, Duan Y. (2004) Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. *Proteins* **55**: 620–634.
30. Hsieh MJ, Luo R. (2004) Physical scoring function based on AMBER force field and Poisson-Boltzmann implicit solvent for protein structure prediction. *Proteins* **56**: 475–486.
31. Lee MR, Tsai J, Baker D, Kollman PA. (2001) Molecular dynamics in the endgame of protein structure prediction. *J Mol Biol* **313**: 417–430.
32. Fogolari F, Tosatto SC. (2005) Application of MM/PBSA colony free energy to loop decoy discrimination: toward correlation between energy and root mean square deviation. *Protein Sci* **14**: 889–901.
33. Jorgensen WL, Maxwell DS, Tirado-Rives J. (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **118**: 11225–11236.
34. Xiang Z, Soto CS, Honig B. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci USA* **99**: 7432–7437.
35. Tanaka S, Scheraga HA. (1976) Statistical mechanical treatment of protein conformation. I. Conformational properties of amino acids in proteins. *Macromolecules* **9**: 142–159.
36. Bowie JU, Luthy R, Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**: 164–170.
37. Miyazawa S, Jernigan RL. (1985) Estimation of effective interresidue contact energies from protein crystal-structures — quasi-chemical approximation. *Macromolecules* **18**: 534–552.
38. Levitt M, Chothia C. (1976) Structural patterns in globular proteins. *Nature* **261**: 552–558.

39. Eisenberg D, McLachlan AD. (1986) Solvation energy in protein folding and binding. *Nature* **319**: 199–203.
40. Sippl MJ. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**: 859–883.
41. Zhou H, Zhou Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* **11**: 2714–2726.
42. Sippl MJ. (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des* 7: 473–501.
43. Melo F, Feytmans E. (1998) Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol* **277**: 1141–1152.
44. Shen MY, Sali A. (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**: 2507–2524.
45. Samudrala R, Moult J. (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol* **275**: 895–916.
46. Finkelstein AV, Badretdinov A, Gutin AM. (1995) Why do protein architectures have Boltzmann-like statistics? *Proteins* **23**: 142–150.
47. Thomas PD, Dill KA. (1996) Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol* **257**: 457–469.
48. Rooman MJ, Wodak SJ. (1995) Are database-derived potentials valid for scoring both forward and inverted protein folding? *Protein Eng* **8**: 849–858.
49. Melo F, Feytmans E. (1997) Novel knowledge-based mean force potential at atomic level. *J Mol Biol* **267**: 207–222.
50. Jones DT, Taylor WR, Thornton JM. (1992) A new approach to protein fold recognition. *Nature* **358**: 86–89.
51. Zhang C, Liu S, Zhou H, Zhou Y. (2004) The dependence of all-atom statistical potentials on structural training database. *Biophys J* **86**: 3349–3358.
52. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A. (2007) Assessment of predictions in the model quality assessment category. *Proteins*, **in press.**
53. Skolnick J. (2006) In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol* **16**: 166–171.
54. Sippl MJ. (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol* **5**: 229–235.
55. Qiu J, Elber R. (2005) Atomically detailed potentials to recognize native and approximate protein structures. *Proteins* **61**: 44–55.
56. Zhang C, Liu S, Zhou H, Zhou Y. (2004) An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci* **13**: 400–411.

57. Buchete NV, Straub JE, Thirumalai D. (2004) Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci* **13**: 862–874.
58. Tosatto SC, Battistutta R. (2007) TAP score: Torsion angle propensity normalization applied to local protein structure evaluation. *BMC Bioinform* **8**: 155.
59. Betancourt MR, Skolnick J. (2004) Local propensities and statistical potentials of backbone dihedral angles in proteins. *J Mol Biol* **342**: 635–649.
60. Benkert P, Tosatto SCE, Schomburg D. (2007) QMEAN: a comprehensive scoring function for model quality assessment. *Proteins*, **in press**.
61. Bradley P, Misura KM, Baker D. (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* **309**: 1868–1871.
62. Zhang Y, Arakaki AK, Skolnick J. (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* **61**(7): 91–98.
63. Zhang Y, Skolnick J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* **57**: 702–710.
64. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA. (2006) A composite score for predicting errors in protein structure models. *Protein Sci* **15**: 1653–1666.
65. Wallner B, Fang H, Elofsson A. (2003) Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins* **53**(6): 534–541.
66. Rychlewski L, Fischer D. (2005) LiveBench-8: The large-scale, continuous assessment of automated protein structure prediction. *Protein Sci* **14**: 240–245.
67. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. (2005) Progress over the first decade of CASP experiments. *Proteins* **61**(7): 225–236.
68. Wang K, Fain B, Levitt M, Samudrala R. (2004) Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct Biol* **4**: 8.
69. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D. (2001) Rosetta in CASP4: Progress in *ab initio* protein structure prediction. *Proteins* (5): 119–126.
70. Zhang Y, Skolnick J. (2004) SPICKER: a clustering approach to identify near-native protein folds. *J Comput Chem* **25**: 865–871.
71. Berman HM, Westbrook J, Feng Z, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**: 235–242.
72. Samudrala R, Levitt M. (2000) Decoys “R” Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci* **9**: 1399–1401.
73. DePristo MA, de Bakker PI, Lovell SC, Blundell TL. (2003) *Ab initio* construction of polypeptide fragments: efficient generation of accurate, representative ensembles. *Proteins* **51**: 41–55.
74. van Gunsteren WF, Bakowies D, Baron R, *et al.* (2006) Biomolecular modeling: Goals, problems, perspectives. *Angew Chem Int Ed Engl* **45**: 4064–4092.

75. Bowie JU, Zhang K, Wilmanns M, Eisenberg D. (1996) Three-dimensional profiles for measuring compatibility of amino acid sequence with three-dimensional structure. *Meth Enzymol* **266**: 598–616.
76. Tosatto SC. (2005) The victor/FRST function for model quality estimation. *J Comput Biol* **12**: 1316–1327.
77. Melo F, Devos D, Depiereux E, Feytmans E. (1997) ANOLEA: a www server to assess protein structures. *Ismb* **5**: 187–190.
78. Wiederstein M, Sippl MJ. (2007) ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucl Acids Res* **35**: W407–410.
79. Adamczak R, Porollo A, Meller J. (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* **56**: 753–767.
80. Sims GE, Kim SH. (2006) A method for evaluating the structural quality of protein models by using higher-order phi-psi pairs scoring. *Proc Natl Acad Sci USA* **103**: 4428–4432.
81. Pugalenthi G, Shameer K, Srinivasan N, Sowdhamini R. (2006) HARMONY: a server for the assessment of protein structures. *Nucl Acids Res* **34**: W231–W234.

This page intentionally left blank

Chapter 5

The Biological Applications of Protein Models

A. Tramontano*

5.1 Introduction

One key question about protein structure modeling is whether it is useful in real life applications. The answer to this question is undoubtedly positive: the cases where protein structure modeling has provided precious information to biologists are so many that it is impossible to give even a short description of a significant fraction of them.[1] Therefore, we will rather focus on the biological problems that they can help solve and give references to some of the specific examples with the caveat that they are just some of the many possible ones that could be listed.

There is, however, one issue that we need to address first: we have to discuss how we can estimate the quality of a model *a priori* and evaluate whether it is sufficient for a given application. While this is not a trivial task, it is an exceptionally important one. If we want protein structure modeling to move out from the computers of the experts and make its way to the labs, we cannot just provide users with a model without a quality estimate attached to it. This is indispensable before the three-dimensional modeling techniques can be effectively

*Department of Biochemical Sciences, University "La Sapienza", P.le A. Moro, 5, Rome, 00187, Italy. Email: anna.tramontano@uniroma1.it.

added to the many tools available and used for designing, interpreting, and evaluating wet experiments in the life sciences.

We will first review the methodologies that can be used to assess the quality of a three-dimensional model of a protein, and next, how this can be exploited to decide whether the expected use of a model is within the realm of reality.

5.2 The Expected Quality of a Model

The different methods that can be used to build a model of a protein have been discussed in previous chapters. For the purpose of this discussion, we can distinguish them into two broad categories: template-based, which include comparative modeling and fold recognition techniques, and template-free methods. The latter can be real *ab initio* techniques where virtually no information is taken from experimentally determined protein structures (although of course many stereo-chemical parameters are estimated from the analysis of known structures), or methods that take advantage of known protein structures more directly, for example, by selecting their fragments to build the model. As in all theoretical distinctions, the boundaries between the two categories are rather blurred. However, for the purpose of what we are going to discuss here, all we need to know is whether or not a model can be associated to one, or very few, protein structures deemed to have a similar structure as the target protein.

The importance of this distinction lies in the observation that in the first case we can use some measure of the expected structural difference between target and template as an estimate of the expected quality of an average model of the target, while in the second case we can only analyze the coordinates of the produced model and try to infer the quality from them.

The simplest case is homology modeling. In a seminal work, Lesk and Chothia[2] analyzed 32 pairs of homologous proteins of known structure and asked the question of how much the *core* of the structures diverged as a function of the sequence identity (a rough measure of the evolutionary distance). There are several definitions of the core of a protein structure. In their work, Chothia and Lesk used an almost tautological definition of *core* as the part of the protein structures that is most conserved between the two homologous proteins under study. Regardless the specific definition, we can intuitively understand what the core of a protein is: the part of the structure that is not peripheral to the folded nucleus of the protein, i.e. the protein without external "decorations" such as loops and small domains that are usually not very well conserved in evolution. In the same paper, Chothia and Lesk also analyzed the extent to which the core is conserved as a function of sequence identity.

This historical piece of work can be discussed from many perspectives. For example, it might be interpreted as a tool to predict the average coordinate error in the core of the model of a protein built by homology.

One could make the following reasoning: since proteins with an average sequence identity of 50% have about 1.0 Å RMSD between corresponding atoms of the main chain of their core, if we build a model by just copying the coordinates of the core of a protein sharing 50% sequence identity with the target protein, the expected error of the main chain of the core of the model is about 1.0 Å. This holds if the correct correspondence between the atoms of the proteins has been used, i.e. if the sequence alignment correctly reflects the optimal structural superposition between the two proteins (note that this does not necessarily correspond to the alignment correctly reflecting the evolutionary relationship).

The next step is to estimate the expected quality of the alignment and whether or not some refinement can be applied to the initial model to move it away from the template and make it more similar to the real target structure, thereby reducing the coordinate error.

We will not discuss refinement here, and will instead say a few words about the problem of estimating the correctness of an alignment *a priori*.

In principle, one expects that the closer two sequences, i.e. the higher the sequence identity or similarity between them, the higher the likelihood of obtaining a correct alignment. This effect, if one could quantify it, should be combined with the results of the Chothia and Lesk[2] analysis to obtain the expected accuracy of a model.

The results of the CASP experiments[3] could help us in estimating this error by computing the average alignment errors of the models submitted for proteins as a function of their pair-wise sequence identity with the template. However, in the last decade, pair-wise alignment has been very rarely used as the basis of a modeling experiment because of the advent of multiple sequence alignment methods and the availability of many more related sequences. In the large majority of the cases, the alignment used for building a comparative model is obtained by extracting the pair-wise alignment from a multiple sequence alignment. We need to address the problem of the quality of the alignment taking into account the number and distribution of the sequences used in the multiple sequence alignment.

One possible method is to consider which is the most difficult pair-wise alignment (i.e. the one involving the two sequences with the lowest sequence identity) in the multiple sequence alignment and use this parameter as an estimator of the difficulty of the alignment.[4] This procedure is based on the hypothesis that the multiple sequence alignment is built iteratively by subsequent pair-wise alignments (the technique used, for example, by ClustalW[5]) but methods based on different rationales do exist[6] and it is unclear whether this measure is appropriate in all cases and whether alternative and more appropriate methods can be devised.

The take-home message of this short digression is that, even in comparative modeling where we can analyze the underlying sequence alignment and know the relationship between sequence identity and structural similarity, predicting the expected quality of models is far from trivial.

Even more complex is the case when the model building method is not based on a template. In this case, we should analyze the coordinates and assess the quality of the model itself. Most methods are based on some energetic evaluation of the model, often based on an estimate of the likelihood of the interactions observed in the model. A blind assessment of the ability of different methods to give a quality estimate to models and to their residues was included in CASP in 2006.[7]

Models produced by servers participating in CASP were made available to all predictors. Predictors were asked to submit estimates for the quality of these models before the corresponding experimental structure was available. They were given the opportunity of submitting quality predictions for the model structure as a whole and/or on a residue-by-residue basis. At the end of the experiment, the observed quality of the server models (according to the results of the CASP automatic evaluation) was compared with the values submitted by the quality predictors.[8]

Unfortunately, the results were not very exciting. Some methods, for example, Pcons,[9] were able to rank a set of models for the same structure according to their relative quality moderately well, but no method was reasonably good in assigning a value to a single model. In fact, a naïve predictor assigning the distance between the structure of the model and that of the template as a quality estimate performed better than, or as well as, essentially all methods of this kind.[8]

The take-home message, therefore, is that we are not yet able to look at a model and compute its expected accuracy with sufficient reliability, and consequently, we cannot equip the end user of a model with a value that can tell him or her whether the model is good enough for a given application.

There are nevertheless rules of thumb that can be applied and that we will try to discuss in the next section. Although they cannot provide a definite answer to the question, most of the time they can be very useful.

5.2.1 *Some Useful Definitions*

In the following, we will use two parameters to describe how close a model is to the corresponding experimental structure, the Root Mean Squared Deviation (RMSD) and the GDT-TS.

Given two sets of n atom coordinates of two proteins a and b, the RMSD is defined as:

$$rmsd(a,b)=\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(a_{ix}-b_{ix})^2+(a_{iy}-b_{iy})^2+(a_{iz}-b_{iz})^2} \qquad (5.1)$$

and is expressed in Ångstrom (Å).

The RMSD might not be the best parameter for comparing a model and a structure because of its quadratic form; if a part of the protein is incorrectly predicted, it is often sufficient to know that it is farther than some threshold from the correct conformation, and this is why a different measure, the GDT-TS, is often used:

$$\text{GDT-TS} = \frac{(\text{GDT-1} + \text{GDT-2} + \text{GDT-4} + \text{GDT-8})}{4} \qquad (5.2)$$

where GDT-1, GDT-2, GDT-4, and GDT-8 are the percentage of Cα atoms of the model that are distant less than 1, 2, 4, and 8 Å from the corresponding atoms of the structure, respectively.

Experimentally determined structures can be obtained, by and large, by NMR and X-ray crystallography. In the first case, the experiment produces an ensemble of structures whose dispersion, expressed, for example, as the average RMSD among the structures, can be considered an indication of the accuracy of the structure. In the second, parameters such as resolution, R- and R-free factors play this role. To orient the reader, in very low resolution structures (let's say around 6 Å), the location of alpha helices (but not beta sheets) can usually be determined, and in low resolution structures (around 3 Å), the backbone and (if the data set is good) the side chains can be seen. At the other end, in very high resolution structure (around 1 Å), even hydrogen atoms can be located. For an accurate structure, one needs a resolution of at least 2–2.5 Å.

For the sake of conciseness, we will not dwell on the relationship between NMR and X-ray crystallography accuracy and the relationship between them. At risk of offending both NMR spectroscopists and X-ray crystallographers, we can assume that an NMR structure corresponds to a low resolution X-ray structure.

5.3 Biological Applications

Let us now explore some of the many biological applications that can be envisaged for a model, namely the possibility of using a model for solving an experimental structure, for understanding the function of the protein, for predicting the domain boundaries in multi-domain proteins, for designing site specific mutants and chimeric proteins, for modifying the stability or solubility, and finally, for docking and/or designing an inhibitor.

Obviously, it would be completely unreasonable to ask that a model provide more information than an experimental structure, therefore we will discuss which information we can expect to derive from a model compared to what we could infer if we had the experimental structure.

5.3.1 *Solving the Phase Problem in Crystallography by Molecular Replacement*

In a crystallography experiment (see Chapter 22), a crystal is irradiated with X-rays whose diffracted waves are collected and measured. The reconstruction of the structure of the molecule in the crystal requires knowledge of the phase of the diffracted waves, which cannot be directly measured. The information is lost in the passage from the three-dimensional structure of the molecule to its two-dimensional diffraction pattern. It can be recovered using experimental methods such as heavy-atom isomorphous replacement and anomalous scattering or by molecular replacement, which relies on the availability of an atomic model of the target structure. This involves taking the model and rotating and translating it into the new crystal system until there is a good match with the experimental data. If this is successful, the amplitudes and phases from this solution can be computed and combined with the data to produce an electron density map. This is obtained using a Fourier transform and is equivalent to focusing the diffraction pattern in other forms of microscopy.

It has been shown[10] that models of reasonable accuracy (GDT-TS above 80–85) can be used for molecular replacement, also in cases where the homologous protein used to build the model fails. Notice that in this case the RMSD between the model and the structure is not a good indicator of whether a model will work in a molecular replacement experiment. Interestingly, the procedure can easily be automated,[11] and therefore, the solution of protein structures by crystallography in structural genomics projects can be sped up.

5.3.2 Prediction of Biological Function

Probably the most important aim of the genomics projects, and of most of modern biology, is the elucidation of the function of the many proteins whose sequence is becoming available.

Function can be defined at very different levels of detail,[12] and clearly, a structure can help, if at all, in elucidating the molecular function of a protein, while its biological function and localization are much more difficult to relate to a structure. As an example, even if we know or are able to predict that a given protein is an enzyme of the hydrolase class, this by itself is not sufficient to decide whether it is involved in digestion, blood coagulation, or apoptosis, etc.

Unfortunately, in general, not even the molecular function can be predicted on the basis of the three-dimensional structure. If the protein contains an active site that we have already seen,[13–15] or if it has been co-crystallized with a mimic of its substrate, we can make educated guesses at some level of detail although rarely can we go as far as inferring its specificity.[16–18] Predicting the function of a protein when its structure does not resemble anything we have already seen, is a very challenging task, whatever the resolution of the structure.

Not surprisingly, the detection of a known active or binding site can only be attempted for models of good quality since we need to search for a subset of residues with similar coordinate sets in proteins of known function.[19–21] Most likely this can only be achieved for homology-derived models. The relationship between sequence identity and structural conservation that we discussed above is computed on average, i.e. over the whole protein core, but a comparative or homology model has the advantage of performing better for regions of the proteins that are evolutionarily conserved. These include the active site, and therefore, we can often use models built using a template sharing a level of identity above, say, 30% with the template.

There is added value in a model built by homology with respect to the simple detection of the homology.[17] Even if we only consider the simpler case of enzymes, only proteins sharing more than 85% sequence identity have strictly conserved function (up to the fourth digit of the Enzyme Classification scheme, EC). The third digit is shared by proteins having at least 55% sequence identity. At 25% sequence identity, not more than 60% protein pairs have the same EC code, an additional 20% only share the first three digits of the classification, and a few percent of pairs have no common EC digit at all. If we compare SwissProt keywords, and therefore can include also non-enzymes, matters become even less hopeful: not even 95% sequence identity guarantees strict conservation of annotation, and at a level of sequence identity of 25%, no more than 45% of the pairs have the same keywords. Clearly, the numbers listed here are bound to change as methods and databases change, but they are useful as reference points.

The possibility of mapping known facts about a protein onto its three-dimensional structural model can be of greater help for unraveling its functional attributes.[22] Clustering of residues conserved in the protein family in the same region of space[23] or the presence of exposed cavities can give important hints and help prioritize experiments in an effective way.

5.3.3 *Redesigning Proteins*

In many cases, the experimental study of multi-domain proteins is difficult and it might be useful, or essential, to sub-clone its functional domains to perform well controlled experiments.[24–27]

The three-dimensional structure of a protein is usually enough to define the boundaries of its domains, and any model of reasonable accuracy — almost certainly any comparative model — can be used for the same purpose. A word of caution is needed here: the correct identification of the structural boundaries of a domain does not necessarily imply that trimming the sequence according to this information will give rise to a protein that can fold correctly or crystallize, because other factors can affect the outcome.

A different, but related, problem is the design of chimeric proteins that can be useful, for example, to study their localization or for biotechnological or medical purposes,[28] and the question arises of whether it is more likely that the protein will accept the insertion at its N- or C-terminus. A related problem is the identification of positions that can be more safely mutated for introducing different or modified amino acids.[29]

The accessibility of the N- and C-terminus as well as of the side chains of its amino acids can usually be reliably evaluated using any comparative or fold recognition model that does not contain gross errors. Since our understanding of the folding process is, to be optimistic, incomplete, predicting whether the latter will be affected by the mutation is not equally straightforward.

A special case is represented by antibodies[30,31] as discussed in Chapter 16 of this book. The relationship between the sequence and structure of the functional site of this class of molecules is rather well understood, and this has led to the development of accurate knowledge-based procedures for antibody modeling. Information gained from the analysis of antibody structures has been successfully exploited to engineer antibody-like molecules endowed with prescribed properties, such as increased stability or different specificity, many of which have a broad spectrum of applications both in therapy and in research.[32–36]

Antibodies or immunoglobulins are multi-chain proteins, consisting of two pairs of light chains (either κ or λ isotype) and two pairs of heavy chains. Both chains are composed of multiple variants of a basic domain of about 100 residues in length. One domain in each chain is variable in sequence and corresponds to the antigen binding region. Their modular nature makes antibody molecules particularly suitable candidates for protein engineering.

The antigen-binding sites of most antibodies are formed primarily by six loops, three from the VL domain (L1, L2, L3) and three from the VH domain (H1, H2, H3). The regions of the variable domains outside these loops are called the framework. In known immunoglobulins, the framework regions are highly conserved in both sequence and main-chain conformation, and they can be accurately predicted using standard homology modeling techniques.

The six loops of the antigen-binding site are even more variable in sequence than the rest of the variable domains. In spite of their high sequence variability, five of the six loops of the antigen-binding site can assume just a small repertoire of main-chain conformations, called "canonical structures".[37–40] These conformations are determined by the length of the loops and by the presence of key residues at specific positions in the antibody sequence (either within the loops or in the framework regions) that determine the conformation of the loops through their packing, hydrogen bonding, or the ability to assume unusual main-chain conformations. The other loop residues are free to vary to modify the topography and physicochemical properties of the antigen-binding site. The identification of the structural determinants of the antigen-binding loops is a necessary requirement for the successful engineering of antibodies with prescribed specificity. In any design involving modifications and/or transplant of the antigen-binding site loops (aimed, for example, at varying the antibody affinity or specificity toward the antigen, at introducing metal-binding sites, or at generating large repertoires of antibody molecules through the use of libraries), it is indeed necessary to keep into account that mutations of residues at most positions of the hypervariable loops will determine only local variations of the antigen binding site, without affecting its main-chain conformation. On the other hand, mutations at key sites will, in most cases, also affect the main-chain conformation of the antigen-binding site loops and are likely to have a larger impact on the affinity toward the antigen.[36]

5.3.4 *Modifying the Biochemical Properties of Proteins*

Biotechnology often requires redesigning proteins having higher stability or solubility than their wild type counterpart.[41] There are several examples of successful cases of such designs in the literature. It is probably reasonable to say that a good comparative model, built using a template with 60–70% sequence identity is as effective as an experimentally determined structure. The computation of the stability of a protein is a difficult task, therefore the previous sentence has to be interpreted in the following way: if we have a hypothesis about which features of a given protein might increase its stability, we can design mutants for testing the hypothesis also on the basis of a good model.

5.3.5 *Docking and Inhibitor Design*

Can we design an inhibitor for an enzyme of known or modeled structure? Undoubtedly, the most desirable use of a protein structure is for the docking of small molecules and designing of inhibitors (see Chapter 18). Can we design an inhibitor for an enzyme of known or modeled structure? Can we identify the mode of binding of a known ligand?

There are many hurdles in designing an inhibitor. Ligands and inhibitors bind to exposed regions of proteins that can be flexible, and therefore, the apo structure, even experimentally determined, cannot necessarily be effectively used as the target of the inhibitor in all cases.[42,43] Next, the conformation of the inhibitor can change upon binding, and again, this is difficult to compute.[44,45] These factors, and a few others, make the design of inhibitors a very challenging task even when a high resolution structure is available. Nevertheless, a very accurate model of a protein can be useful to the docking of small molecules and the design of new ones.[46–48]

In order to be useful, models need to be rather accurate as shown by an experiment performed by Moult and co-workers:[49] they took all protein targets in CASP for which a bound molecule was present in the experimentally determined structure and tried to verify whether the ligands would fit into any of the models submitted to the experiment. The conclusion was that in most cases the side chains of the binding site were not predicted sufficiently correctly to allow the ligand to be positioned. Sequence alignment errors between models and templates were shown to be the most deleterious.[49]

5.4 Future Outlook

We now stand at a point where we can produce models of proteins with respectable accuracy, and not only when homology is exploited.[50] Most modeling methods can be automated and run in high-throughput mode, providing an impressive amount of information.[51] The obvious next step is to transform information into biological knowledge.

There are many hurdles in the path, but the protein structure modeling community is devoting a large effort in overcoming them. Solving them will have a direct impact on the reliability of the biological conclusions that we can derive, on the exploration of new applications, and on an increased usage of the tools within the biological community.

First of all, it is becoming apparent that we need effective methods to evaluate a model and its reliability.[8] Only once this issue has been solved can we attempt to precisely define the fields of application as a function of the expected quality of the model. The recent experience with the CASP quality assessment experiment has highlighted a rather embarrassing situation, and there is no doubt that this will prompt scientists to tackle the issue with more energy. It is reasonable to expect that, in the near future, every model will come with some estimate of its reliability. This will imply that the existing protein model databases[52–54] will be more and more useful to the biological community.

CASP has highlighted another area where more effort should be focused: refinement of models. We only briefly mentioned it here because there are no methods at present that can consistently improve upon an initial model. This is a very crucial issue especially in comparative modeling. Unless we can tell more about a protein than what is implied by its homology with the template, there is no chance that we can use the models to address the issue of specificity of both substrates and interactions. Indeed, blind predictions of protein function as assessed by the CASP experiment are not producing very exciting results.[55–57] Although this is partially due to the difficulty of assessing the quality of the predictions, the results of the experiment have made clear that this is still an open issue that needs much more attention than it has received so far.

The pace at which new sequence data are produced[58] requires effective tools to understand their biological meaning. There is no way we can experimentally analyze all the proteins of known sequence in a reasonable time frame, therefore, we need fast, reliable, and effective computational tools to exploit them, and since function is by and large determined by structure, three-dimensional modeling is called upon to fulfill a crucial role in the process by providing the framework for understanding the biological function of the gene products. We are not there yet, but the speed at which methods are progressing makes it likely that this problem will be at least partially solved in a matter of a few years.

Acknowledgments

This work is supported by the European Commission within its FP6 Program, under the thematic area, "Life sciences, genomics and biotechnology for health," contract number LSHG-CT-2003-503265. The support of the Istituto Pasteur Fondazione Cenci Bolognetti is gratefully acknowledged.

References

1. Tramontano A. (2006) *Protein Structure Prediction: Concepts and Applications*, 1st ed. Wiley VCH, Weinheim.
2. Chothia C, Lesk A. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**: 823–826.
3. Moult J, Pedersen J, Judson R, Fidelis K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**: ii–v.
4. Cozzetto D, Tramontano A. (2005) Relationship between multiple sequence alignments and quality of protein comparative models. *Proteins* **58**: 151–157.
5. Thompson JD, Higgins DG, Gibson TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weigh matrix choice. *Nucl Acids Res* **22**: 4673–4680.
6. Notredame C, Higgins DG, Heringa J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205–217.
7. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. (2007) Critical assessment of methods of protein structure prediction (CASP)-round VI. *Proteins*, **in press**.
8. Cozzetto D, Kryshtafovich A, Ceriani M, Tramontano A. (2007) Assessment of predictions in the model quality assessment category. *Proteins*, **in press**.

9. Wallner B, Fang H, Elofsson A. (2003) Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins* **53**(6): 534–541.
10. Raimondo D, Giorgetti A, Miele AE, Tramontano A. (2005) Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics* **21**: 72–76.
11. Raimondo D, Giorgetti A, Bosi S, Tramontano A. (2006) Automatic procedure for using models of proteins in molecular replacement. *Proteins* **66**: 689–696.
12. GeneOntologyConsortium. (2004) The Gene Ontology (GO) database and informatics resource. *Nucl Acids Res* **32**: D258–D261.
13. Barker JA, Thornton JM. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **19**: 1644–1649.
14. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. (2002) Analysis of catalytic residues in enzyme active sites. *J Mol Biol* **324**: 105–121.
15. Laskowski RA, Thornton JM, Humblet C, Singh J. (1996) X-SITE: use of empirically derived atomic packing preferences to identify favorable interaction regions in the binding sites of proteins. *J Mol Biol* **259**: 175–201.
16. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. (2003) Automatic prediction of protein function. *Cell Mol Life Sci* **60**: 2637–2650.
17. Devos D, Valencia A. (2000) Practical limits of function prediction. *Proteins* **41**: 98–107.
18. Pizzi E, Tramontano A, Tomei L, La Monica N, Failla C, Sardana M, Wood T, De Francesco R. (1994) Molecular model of the specificity pocket of the hepatitis C virus protease: implications for substrate recognition. *Proc Natl Acad Sci USA* **91**: 888–892.
19. Ausiello G, Via A, Helmer-Citterich M. (2005) Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinform* **6**(4): S5.
20. Ausiello G, Zanzoni A, Peluso D, Via A, Helmer-Citterich M. (2005) pdbFun: mass selection and fast comparison of annotated PDB residues. *Nucl Acids Res* **33**: W133–W137.
21. Via A, Peluso D, Gherardini PF, de Rinaldis E, Colombo T, Ausiello G, Helmer-Citterich M. (2007) 3dLOGO: a web server for the identification, analysis and use of conserved protein substructures. *Nucl Acids Res* **35**: W416–W419.
22. Peitsch MC. (2002) About the use of protein models. *Bioinformatics* **18**: 934–938.
23. Aloy P, Querol E, Aviles FX, Sternberg MJ. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol* **311**: 395–408.

24. Chen Y, Qiu S, Luan CH, Luo M. (2007) Domain selection combined with improved cloning strategy for high throughput expression of higher eukaryotic proteins. *BMC Biotechnol* 7: 45.
25. Liu J, Rost B. (2004) Sequence-based prediction of protein domains. *Nucl Acids Res* **32**: 3522–3530.
26. Nimrod G, Glaser F, Steinberg D, Ben-Tal N, Pupko T. (2005) In silico identification of functional regions in proteins. *Bioinformatics* **21**(1): i328–i337.
27. Quevillon-Cheruel S, Leulliot N, Gentils L, van Tilbeurgh H, Poupon A. (2007) Production and crystallization of protein domains: how useful are disorder predictions? *Curr Protein Pept Sci* **8**: 151–160.
28. Arun KH, Kaul CL, Ramarao P. (2005) Green fluorescent proteins in receptor research: An emerging tool for drug discovery. *J Pharmacol Toxicol Meth* **51**: 1–23.
29. Kochendoerfer GG. (2005) Site-specific polymer modification of therapeutic proteins. *Curr Opin Chem Biol* **9**: 555–560.
30. Morea V, Lesk AM, Tramontano A. (2000) Antibody modeling: implications for engineering and design. *Methods* (San Diego, CA) **20**: 267–279.
31. Morea V, Tramontano A, Rustici M, Chothia C, Lesk A. (1997) Antibody structure, prediction and redesign. *Biophys Chem* **68**: 9–16.
32. Kim SJ, Park Y, Hong HJ. (2005) Antibody engineering for the development of therapeutic antibodies. *Mol Cells* **20**: 17–29.
33. Teillaud JL. (2005) Engineering of monoclonal antibodies and antibody-based fusion proteins: successes and challenges. *Expert Opin Biol Ther* **5**(1): S15–S27.
34. Wang HW, Cole D, Jiang WZ, *et al.* (2005) Engineering and functional evaluation of a single-chain antibody against HIV-1 external glycoprotein gp120. *Clin Exp Immunol* **141**: 72–80.
35. Sanz L, Cuesta AM, Compte M, Alvarez-Vallina L. (2005) Antibody engineering: Facing new challenges in cancer therapy. *Acta Pharmacol Sin* **26**: 641–648.
36. Donini M, Morea V, Desiderio A, *et al.* (2003) Engineering stable cytoplasmic intrabodies with designed specificity. *J Mol Biol* **330**: 323–332.
37. Chothia C, Lesk A. (1987) Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* **196**: 901–917.
38. Chothia C, Lesk A, Tramontano A, *et al.* (1989) Conformations of immunoglobulin hypervariable regions. *Nature* **342**: 877–883.
39. Morea V, Tramontano A, Rustici M, Chothia C, Lesk A. (1998) Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol* **275**: 269–294.
40. Tramontano A, Chothia C, Lesk AM. (1990) Framework residue 71 is a major determinant of the position and conformation of the second hypervariable region in the VH domains of immunoglobulins. *J Mol Biol* **215**: 175–182.

41. Li H, Cocco MJ, Steitz TA, Engelman DM. (2001) Conversion of phospholamban into a soluble pentameric helical bundle. *Biochem* **40**: 6636–6645.
42. Rosenfeld R, Vajda S, DeLisi C. (1995) Flexible docking and design. *Ann Rev Biophys Biomol Struct* **24**: 677–700.
43. Waszkowycz B. (2002) Structure-based approaches to drug design and virtual screening. *Curr Opin Drug Discov Devel* **5**: 407–413.
44. Goodsell DS, Morris GM, Olson AJ. (1996) Automated docking of flexible ligands: applications of AutoDock. *J Mol Recogn* **9**: 1–5.
45. Ma B, Shatsky M, Wolfson HJ, Nussinov R. (2002) Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci* **11**: 184–197.
46. Vangrevelinghe E, Zimmermann K, Schoepfer J, *et al.* (2003) Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J Med Chem* **46**: 2656–2662.
47. Ragno R, Simeoni S, Castellano S, *et al.* (2007) Small molecule inhibitors of histone arginine methyltransferases: homology modeling, molecular docking, binding mode analysis, and biological evaluations. *J Med Chem* **50**: 1241–1253.
48. Ooms F. (2000) Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry. *Curr Med Chem* 7: 141–158.
49. DeWeese-Scott C, Moult J. (2004) Molecular modeling of protein function regions. *Proteins* **55**: 942–961.
50. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. (2005) Progress over the first decade of CASP experiments. *Proteins* **61**(7): 225–236.
51. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. (2005) Critical assessment of methods of protein structure prediction (CASP) — round 6. *Proteins* **61**(7): 3–7.
52. Kopp J, Schwede T. (2006) The SWISS-MODEL Repository: new features and functionalities. *Nucl Acids Res* **34**: D315–D318.
53. Castrignano T, De Meo PD, Cozzetto D, Talamo IG, Tramontano A. (2006) The PMDB Protein Model Database. *Nucl Acids Res* **34**: D306–D309.
54. Pieper U, Eswar N, Davis FP, *et al.* (2006) MODBASE: A database of annotated comparative protein structure models and associated resources. *Nucl Acids Res* **34**: D291–D295.
55. Lopez G, Rojas A, Tress M, Valencia A. (2007) Assessment of predictions submitted for the CASP7 function prediction category. *Proteins.*
56. Pellegrini-Calace M, Soro S, Tramontano A. (2006) Revisiting the prediction of protein function at CASP6. *FEBS J* **273**: 2977–2983.
57. Soro S, Tramontano A. (2005) The prediction of protein function at CASP6. *Proteins* **61**(7): 201–213.
58. Consortium IHGS. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

This page intentionally left blank

Section II

From Structure to Function to Design

This page intentionally left blank

Chapter 6

Evolution of Protein Folds

A. N. Lupas* and K. K. Koretke†

6.1 Introduction

Given an estimated 100 million species on earth and several thousand protein-coding genes per species, the total complement of the world's proteome is approximately a trillion. This represents an insignificant proportion of all proteins that are possible. Even at a defined chain length of 100 residues, the number of possible polypeptide chains (20^{100}) vastly exceeds the number of particles in the known universe. Moreover, this trillion is not a random sample of the polypeptide space; instead, many proteins share recognizable similarity in sequence and structure, since they evolved from a basic complement of autonomously folding units, referred to as domains. How did this limited set emerge from among the nearly endless possibilities? In this chapter, we will discuss scenarios for the origin of folded proteins and mechanisms for their differentiation into the families observed today.

*Corresponding author.
Department of Protein Evolution, Max-Planck-Institute for Developmental Biology 72076 Tübingen, Germany. Email: andrei.lupas@tuebingen.mpg.de.
†Computational Chemistry Group, GlaxoSmithKline Pharmaceuticals Collegeville, PA 19426-0989, USA.

6.2 Protein Folding

Proteins need to fold to exert their activity. This process does not simply entail an approximate spatial arrangement, a state generally referred to as a molten globule, but actually requires the polypeptide chain to assume a specific structure to within fractions of an Angstrom in a reproducible fashion. While molten globules do represent folding intermediates, they usually lack biological activity. Natively unstructured proteins only apparently contradict the folding requirement, as they are dependent on folded scaffolds for their activity, in the context of which they also assume a defined, reproducible structure.

Given the importance of folding for the biological activity of proteins, it is surprising to find that this is a complicated, easily derailed process. For example, in healthy humans, only about one third of synthesized cystic fibrosis transmembrane conductance regulators reach the membrane in a folded state; the rest is degraded due to misfolding.[1] Thus even under normal conditions, cells allocate substantial resources to systems ensuring the folding, quality control, and turnover of proteins. Under stress conditions, these systems may come to dominate the cell, such as in the archaeon *Pyrodictium occultum*, where the major folding factor rises over tenfold to 73% of the total soluble protein content when the cells are shifted from their normal growth temperature of 90°C to 108°C.[2] How problematic protein folding can be is illustrated by the fact that in humans, many degenerative diseases are protein misfolding diseases (see Chapter 17), such as cystic fibrosis, Alzheimer's, Parkinson's, and Huntington's diseases.[3]

Despite the frequently encountered problems with folding, natural proteins nevertheless represent a best-case group, as most random polypeptide chains do not fold at all. Estimating the actual proportion of folding polypeptides is practically impossible at our current state of knowledge, but one may arrive at upper boundaries based on a few observations: (i) Although many point mutations appear neutral, combining several such mutations usually results in substantially impaired folding. Thus, even though naturally observed proteins can fold, most of their closely related variants can not. (ii) Despite our best efforts, protein design projects frequently do not yield more than molten globules or amyloid-like aggregates. (iii) Attempts to isolate folded proteins from random sequence libraries have produced only a handful of successes to date,[4] even when the libraries are biased for hydrophobic patterns typical of secondary structures.[5,6] This suggests a success rate of no more than $1:10^{10}$. (iv) An attempt to rescue half of a folded protein by fusion to random sequences from the *Escherichia coli* genome yielded only a small number of folded exemplars, possibly just one.[7] Thus, even in a situation far from random, the success rate was only about $1:10^9$.

At less than one in a billion — and possibly much less — it is fair to say that to an exceedingly good first approximation, polypeptide chains do not fold. It is amazing that life would be built on such a difficult property. The question of how folded proteins evolved is therefore entirely non-trivial.

6.3 Homology and the Reconstruction of Evolutionary Events

Before setting out on evolutionary arguments, we need to address a few fundamental points about homology, analogy, and the reconstruction of evolutionary events: evolution happened once; its path can therefore not be proven by the standards of experimental science. As an added difficulty, proteins do not fossilize, so any direct observation of intermediate forms is impossible. In retracing the path of molecular evolution, we must thus study its mechanism, i.e. the possibility and likelihood of certain types of events, and use it to extrapolate from traits observed today, guided by the principles of parsimony ("Occam's razor") and likelihood. In the process, we face several problems: (i) Similar traits in different proteins may be of homologous or analogous origin, and we cannot prove rigorously by scientific standards, which of the two possibilities is true. Instead, our criteria for what constitutes evidence of homology keep evolving as we find that extrapolation from more and more distant connections allows for useful structural and functional predictions. (ii) The difficulty in obtaining data (as well as the pressures of publication) often leads us to propose general principles from a few observed cases — sometimes as little as one. (iii) Because of the large and unknowable body of missing data (extinct intermediate forms), we streamline evolutionary scenarios under the twin constraints of parsimony and likelihood, even though the path actually taken by nature will almost certainly have been more tortuous. This is a connect-the-dots problem; when most dots are missing, the picture traced will only be a simplified and sometimes erroneous sketch of the underlying figure. (iv) Finally, probabilistic approaches fail us when single events trigger further developments by contingency, since the likelihood of the event itself becomes irrelevant.

Contrasting with this long list of problems is the extraordinary usefulness of extrapolation by homology in modern biology. As the number of structural solutions available to proteins for many tasks is limited — witness the convergent emergence of Ser-His-Asp catalytic triads in many hydrolytic enzymes,[8] but there are nearly endless possibilities for reaching these solutions in the linear space of sequence — sequence similarity is considered the primary marker of homology.[9] Such sequence "homology", deduced from sequence comparisons with programs like BLAST, has developed into one of the most powerful tools in molecular biology. Although ultimately not provable, its results have turned out to be very robust, and the use of protein sequences as documents of evolutionary events has yielded a detailed and coherent picture of molecular evolution reaching back more than 3.5 billion years ago to the time before the last universal common ancestor.

6.4 Stability of Folds Across Time

The main reason for our ability to extrapolate so far back in time is the high evolutionary permanence of protein domains. These are autonomously folding elements that act as the units of structure in modern proteins. Although of considerable diversity, most domains show similarities in sequence and structure, which reflect their origin from a basic complement of ancestral forms and allow us to group them into a hierarchy of families, superfamilies, and folds, as described in Chapter 9. Multiple efforts have been made to evaluate the size and age of this basic complement:[10–12] it would seem that there were only about 10^3 ancestral folds and that these were largely established at the time of the last universal common ancestor, with a few successful domains probably arising later in bacteria, archaea, or eukaryotes, and spreading into the other kingdoms by lateral transfer and endosymbiosis. We note, though, that estimating the age of domains is fraught with problems and probably not reliable at present. Current estimates, which are based on phylogenetic spectrum, depend on our ability to recognize the presence of domains in individual organisms from sequence data and are not robust against lateral transfer (which would make domains seem older) and gene loss (which would make them seem younger). Gene loss may be the greatest source of problems, since even major lineages arose from comparatively small founder populations and the accidental loss of individual genes in these populations would have represented the kind of singularities with large contingent effects mentioned in the previous section.

The stability of domains across evolutionary time is sufficient to allow the assignment of more than a third of all residues encoded in present-day genomes to one of 2500 domain families of known structure; this number rises to two thirds, if only soluble proteins with homologs in other organisms are considered.[13] Occasionally, this evolutionary stability can take impressive forms. Thus, the core complement of ribosomal proteins, which was definitely present at the time of the last common ancestor, is still more than 40% identical in sequence between all organisms on earth. In the face of such persistence, it is probably fair to call the ribosome a living fossil, a molecule so central to cellular processes that its modification has become nearly impossible. Other molecules frozen in time include ubiquitin, which we have chosen to illustrate a basic point on the preservation of sequence and structure in domains (Fig. 6.1). Human and yeast ubiquitin are 96% identical in sequence, despite having diverged more than 2 billion years ago; this level of conservation reflects ubiquitin's role as the pivotal molecule in eukaryotic protein degradation. Unsurprisingly, the structure of the two proteins is also nearly identical. If we now

(a)

```
human UBI  MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG 76
96% id     |||||||||||||||||| |||| ||| ||||||||||||||||||||||||||||||||||||||||||||||||
yeast UBI  MQIFVKTLTGKTITLEVESSDTIDNVKSKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG 76
13% id         |       |                     |   |  |    |       |  |                ||
yeast SUMO INLKVSD-GSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGG 86
```

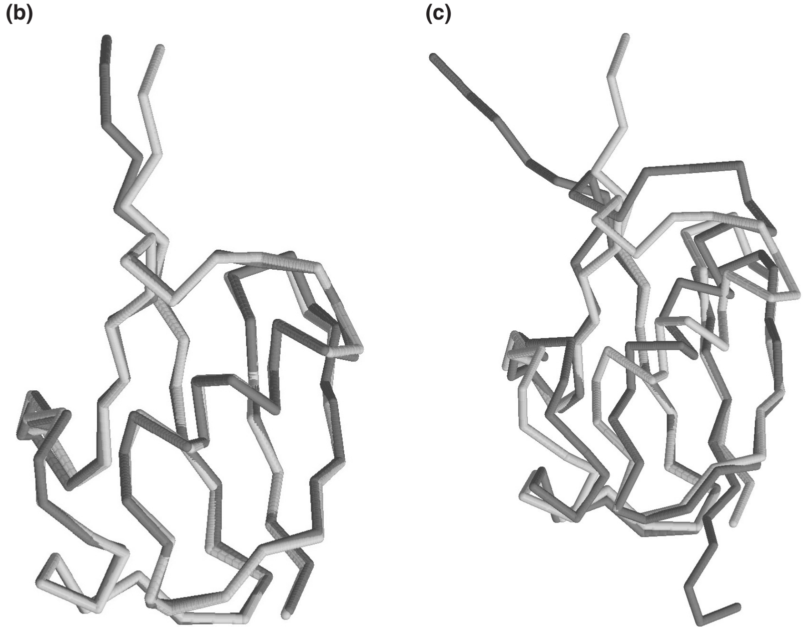


Fig. 6.1 Evolutionary permanence of the ubiquitin fold. **(a)** Sequence alignment of human ubiquitin, yeast ubiquitin and yeast SUMO; identical positions are marked by vertical bars. **(b)** Superposition of ubiquitin from yeast (light) and human (dark). **(c)** Superposition of yeast ubiquitin (light) and yeast SUMO (dark).

turn to a comparison between ubiquitin and a paralog, SUMO, we find that despite only a remnant of 13% sequence identity in yeast (about the level expected by chance between two unrelated proteins), their folds are preserved to an astonishing degree, being closely superimposable. It is noteworthy that this structural conservation is accompanied by the preservation of key sequence properties, since profile-based search methods, such as PSI-Blast, readily identify the homologous relationship between the two proteins despite their apparent sequence dissimilarity.

Why are protein structures so highly preserved, even after considerable divergence of their sequences? We attribute this to a discrete nature of fold space.[14–16] To use an analogy, we view folded proteins as located on islands of stability (folds), which are scattered in a vast ocean of unfolded states. On their respective islands, proteins can pace about by mutation, gradually diverging through adaptive changes and neutral drift. They can, however, not leave the island, as it is separated from other islands by large expanses of unfavorable conformations, which cannot be bridged by point mutation. Occasionally, large and rare events, which will be discussed in the next section, allow proteins to cross to another island of stability, but mostly, proteins will be forced to maintain their folds over billions of years of evolution, even as their sequences diverge. Divergence will not occur randomly, as proteins have to maintain compatibility with their respective folds. This means that they will retain common features, which can be abstracted into characteristic sequence profiles and used to identify the correct fold for newly sequenced proteins. In agreement with this view, methods designed to exploit ever more distant sequence similarities based on profiles and profile Hidden Markov Models are among our most powerful tools for structure prediction.[17–19]

A competing view maintains that fold space is in fact continuous.[20] In this view, protein folds are connected seamlessly by intermediate states that specify structural ensembles, providing for a smooth evolutionary transition between forms. This view is supported by the observation that fold change is possible by point mutation in proteins such as coiled coils,[21] helical bundles,[22] cysteine knots,[23] and the Arc repressor.[24] We find that these observations are unlikely to be informative for the general shape of the folding landscape for the following reasons: (i) All the proteins for which such fold changes have been described are around 50 residues long, a length quite untypical for most domains, which on average have 185 residues (as determined for fold classes a-d of SCOP 1.71); short proteins have disproportionately fewer fold determining residues. (ii) Most of the domains are unusual, as they are known to have multiple, closely spaced energy minima due to the interchangeability of their fold-determining interactions (knobs-into-holes packing in coiled coils and helical bundles, disulfide bonds in cysteine knots). (iii) The extent to which fold change has occurred is debatable; in the Arc repressor it merely involves a strand-to-3_{10}-helix transition in a terminal secondary structure element, and in coiled coils it is unclear whether the changes reflect structural diversity within one fold or true fold diversity. The most impressive instance of fold change by point mutation is the recent report of a 56-residue polypeptide, in which seven point mutations are sufficient to switch the structure from a three-helix bundle to an open-faced β-sandwich.[25] However, the authors do not draw evolutionary conclusions, and it is indeed unclear whether the sequence of this polypeptide could emerge under natural (i.e. functional) selection and whether the final seven required mutations would include any intermediate forms stable enough to allow for the change to happen.

We think that the continuous view of fold space is an instance of a general principle deduced from a small number of examples. In our opinion, the proteins described above are better understood as cases where islands of stability are close together, forming archipelagos in which transitions from one island to the next can occur through minor changes. If such archipelagos were frequent, sequence search programs would routinely connect proteins of different folds. They do not, as can already be gathered from the fact that examples to the contrary are worth publishing,[26] and the extent to which they do not is quite amazing. In fact, even extended chains of consecutive profile searches from low-significance sequence matches rarely connect proteins of different folds, as long as a loose consistency measure is applied.[17] Correspondingly, the growth of sequence databases and the improvement of sequence comparison methods have provided most of the progress in protein structure prediction over the last decade.[27] We would contend that, since sequence searches are among the most extensively used tools in molecular biology and cover the entire known complement of proteins, their output provides a much more robust view of fold space than a few individual examples of proteins with closely spaced energy minima of different folds.

6.5 Fold Change in Evolution

The high evolutionary permanence of folds emphasized above should not detract from the fact that changes can occur over time by a variety of genetic events. Before engaging in a discussion of these events, we would like to briefly touch on what constitutes a fold and how one might judge when two folds are different, issues that will be presented in detail in Chapter 9. In essence, we consider a fold a conserved, topologically distinct arrangement of secondary structures in a domain; extensions and insertions not present in homologs are considered decorations. A fold change occurs when one or more secondary structure elements within the fold alter their nature and/or their topology. The difficulties associated with defining domains and determining when topological differences are sufficiently pronounced to warrant separating two folds are the main reasons for the substantial differences between structure classification systems.[28] Discrepancies also arise because of the uneasy coexistence of homologous and analogous traits used to generate the classifications. Nevertheless, it is possible to arrive at consensus representations for most folds, which have been termed metafolds.[29]

As outlined in the previous section, the accumulation of point mutations may be sufficient to trigger a change in fold in some proteins [Fig. 6.2(a)]; similarly, insertions and deletions (indels) may have this effect [Fig. 6.2(b)]. The main causes of fold change, however, are rare occurrences involving topological substitutions, circular permutations, strand swaps, strand and hairpin invasions, and 3D domain swaps. Most have been described in detail in a review article by Grishin, which we warmly recommend to our readers.[14] We have listed examples for such events, as deduced from the comparison of present-day homologous proteins, in Fig. 6.2. In analyzing these examples, it becomes clear that most rely on preceding point mutations and indel events. Thus, the 3D domain swap that led to the emergence of the histone fold from the C-domains of AAA+ ATPases was most likely preceded by a deletion of the loop connecting helices 2 and 3;[30] this prevented the helices from folding back onto each other and required the antiparallel association of two monomers in

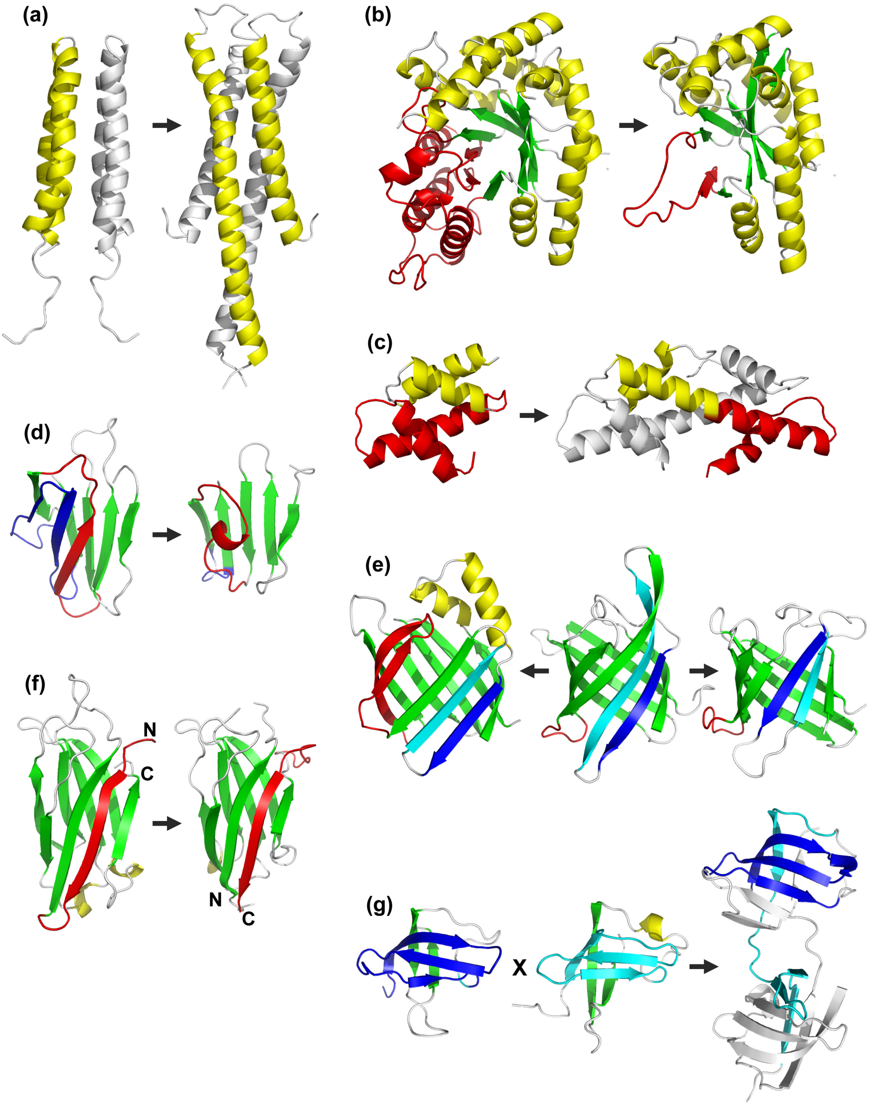


Fig. 6.2 Mechanisms of fold change. Helices are in yellow and strands in green, unless otherwise noted; in homo-dimers, one monomer is shown in grey. **(a)** Transition in the handedness of a four-helix bundle between the DHp domains of EnvZ (1JOY) and *Thermotoga* TM0853 (2C2A) by the cumulated effect of point mutations. **(b)** Transition from an eight-stranded, all-parallel TIM barrel (bacterial luciferase, 1LUC) to a seven-stranded barrel containing one antiparallel strand (nonfluorescent flavoprotein, 1NFP) by deletion and helix-to-strand transition; the affected

order to cover the exposed hydrophobic core [Fig. 6.2(c)]. Similarly, the topological substitution in the C-terminal domain of *Pseudomonas* G4-amylase relative to other amylases was most likely preceded by the deletion of an adjacent hairpin, which required a broader secondary structure element to cover at least part of the exposed hydrophobic core [Fig. 6.2(d)]. Conversely, the hairpin invasion that led to the origin of retinoic acid binding protein from the basic lipocalin fold was probably preceded by a large insertion between strands 4 and 5, which folded back to expand the barrel [Fig. 6.2(e)]. Compared to these events, circular permutation [Fig. 6.2(f)] seems to be considerably more complex, as the simplest mechanism proposed for it requires gene duplication, fusion, and at least two deletions.[31] In light of this, it seems surprising that circular permutations are so frequent, being detectable in about half of all known folds.[32]

As an aside, the extent to which fold change is open to debate is illustrated by the fact that, even though we would consider all cases in Fig. 6.2 examples of fold change, both the SCOP and CATH classifications only agree with this estimate for the 3D domain swap in Fig. 6.2(c). They consider all other examples to still have the same fold, except for Fig. 6.2(g), which they do not classify. We interpret this as a symptom of the tensions arising from the use of both homologous and analogous similarities for structural classification. All examples in Fig. 6.2 involve homologous proteins; analogous proteins with

region is colored red. **(c)** Dimerization induced by 3D domain swapping between the C-domain of a AAA+ ATPase (RuvB, 1INs) and an archaeal histone (1B67); the swapped helical hairpin is shown in red. **(d)** β-Hairpin deletion (blue) and strand-to-helix transition (red) between the C-terminal domains of a canonical amylase (1BPL) and of a variant form (2AMG). **(e)** Variations on the lipocalin fold: β-hairpin invasion between retinol binding protein (1HBQ, center) and retinoic acid binding protein (1CBS, left), and strand swapping between retinol binding protein and thrombin inhibitor (1AVG, right); the invading hairpin is shown in red and the swapped strands in cyan and blue. **(f)** Circular permutation between the C2 domains of synaptogamin I (1RSY) and phospholipase C (1QAS); the permuted strand is colored red. **(g)** Fragment fusion of two β-meanders from different OB folds (cold shock protein, 2MEF, right, blue; and S1 RNA binding domain, 1SRO, center, cyan) yields a domain-swapped dimer (2BH8).

the same topological differences would in most cases be classified into separate folds.

All panels in Fig. 6.2 show transitions between two forms, typically a highly populated core fold and a variant, but this may be a simplified account of the actual evolutionary process. As we mentioned before, in cases where most dots are missing, simplification is unavoidable. A case we have been investigating involves a group of homologous β-barrels with at least three distinct topologies, which we have grouped together into the cradle-loop metafold, based on the peculiar shape of their ligand binding sites (Fig. 6.3). Originally, we surmised that two of the folds, whose homologous relationship we could recognize, were related by circular permutation.[33] As we explored these proteins further, we found that they were related by a strand swap and a strand invasion via an intermediate third fold.[34] Each of the three folds has produced further topological variants by deletion, circular permutation, and/or strand invasion, generating a network of related folds. Given the dearth of similar studies on other proteins, it is impossible to judge how prevalent such fold networks are in nature.

A noteworthy member of the cradle-loop network is the B3 domain, which may have arisen by a mechanism we have not addressed yet, namely the fusion of two different half-barrels. Although the homologous relationship of this domain to other cradle-loop proteins is still unclear and we are not aware of any other good example from naturally occurring proteins, it seems entirely plausible that new folds may arise through the fusion of fragments from existing folds, for example, by illegitimate recombination. In an experiment we alluded to earlier, Riechmann and Winter tried to rescue the N-terminal half of the OB-fold protein CspA by fusion to random segments of the *E. coli* genome.[7] Several fusions were sufficiently protease-resistant to survive the phage-display selection process, and one, which involved fusion to the N-terminal half of another OB-fold protein gave a clearly folded chimera, whose structure could be solved by crystallography.[35] Since the two halves of the chimera were homologous β-meanders, one might well have anticipated a pseudo-symmetrical structure, but in fact the structure showed a homodimeric

OB-fold extended to one side by a β-hairpin; the β-hairpin and the C-terminal strand originated by 3D domain swapping from another monomer [Fig. 6.2(g)]. The extent to which this chimera represents a new fold will certainly be a matter of debate, since several domain-swapped OB-folds, albeit lacking the β-hairpin extension, are currently classified into the same fold as monomeric OB-folds.

A striking aspect of the chimera is the fact that, despite the homology of the two meanders, it was not a "marriage of equals". The N-terminal, CspA-derived meander proved completely dominant in bringing about a fold close to that of its parent; it maintained its structure to within less than 2.5 Å root-mean-square deviation and forced the C-terminal meander to change its topology and the secondary structure state of more than half of its residues. It seems reasonable to assume that such fragments with a strong folding propensity could be complemented not only by other protein fragments, but also by out-of-frame or antisense fragments, which have not been selected for their folding propensity, but are clearly non-random.[36] At the frequency of frame shifts, no part of a genome is permanently out-of-frame.

With so many mechanisms capable of yielding fold changes, how can folds be so stable in evolution? Given the small number of genes per cell, the huge populations and short generation times, illegitimate recombination alone must allow a bacterial species such as *E. coli* to produce hundreds of new folded proteins per year. Where are they? We would like to offer several thoughts on this apparent contradiction. (i) Folding does not imply function; it is just a prerequisite for it. The large majority of the "hopeful monsters" resulting from such events would probably be out-competed rapidly by their established siblings, who have been optimized for all the main biological functions over eons of evolution. (ii) A very small number would survive to contribute to species- or genus-specific, niche functions and die out when their hosts become extinct. Some of the singletons that we encounter every time we sequence a new genome may fall into this category. These proteins would account for the bulk of the so-called unifolds, which are folds that occur in only one protein. Whereas 80% of proteins fall into one of only about 400 common folds, the remaining 20% are estimated to form more than 10^4 unifolds.[37] (iii) Most

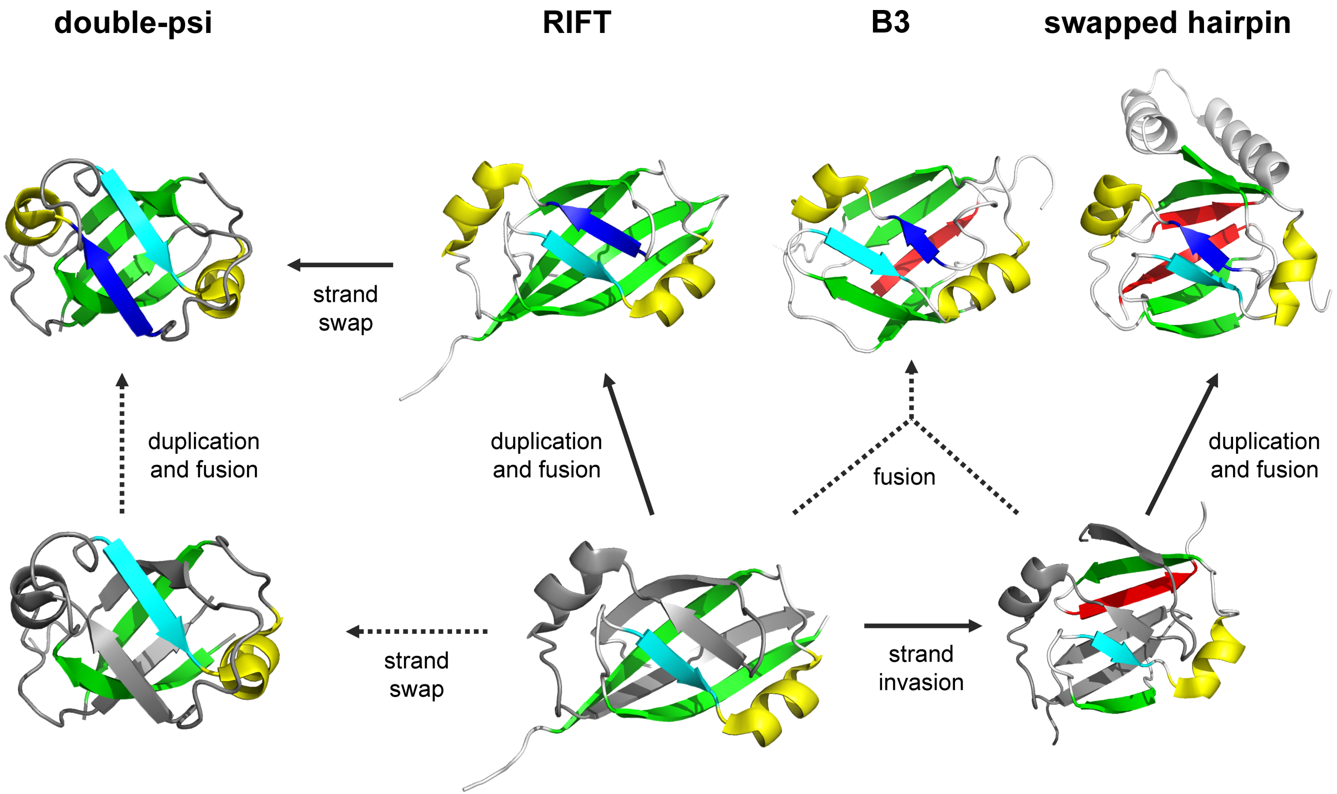


Fig. 6.3. (See caption on next page.)

survivors, however, would either recapitulate a previously successful fold change or resolve back to their original fold, making it difficult in either case to recognize that anything had happened. For example, half of all chains of events required for circular permutation would resolve back to the *status quo ante*. The OB-fold chimera discussed above might well also resolve back to a canonical OB-fold under evolutionary pressure, as this would entail a single deletion event removing the domain-swapped β-hairpin. We conclude that the dearth in starting material is not the primary limitation in establishing new folds; it is the lack of opportunity against entrenched and highly optimized competitors.

6.6 Origin of Folds

In modern proteins, domains act as the unit of protein structure. New proteins arise by the combinatorial shuffling of existing domain types, which adapt to new functional requirements by point mutations and indels, while preserving their basic fold. The origin of folded domains, however, remains substantially unknown.

One possibility is an origin *de novo*, by random concatenation of amino acids, followed by selection. Domains would seem to have too

Fig. 6.3 Evolution of the cradle-loop barrel metafold. An ancestral homo-dimeric RIFT barrel (bottom row, center; modeled) gave rise to swapped-hairpin barrels (e.g. AbrB, 1YFB) by strand swapping. Both folds yielded single-chain barrels by duplication and fusion. For RIFT barrels, this is now the dominant form (e.g. PhS018, 2GLW); for swapped-hairpin barrels, the homo-dimeric form remained dominant, giving rise to single-chain versions by independent events in several lineages (e.g. MraZ, 1N0G). The double-psi barrels originated from a RIFT barrel by strand swapping (e.g. VAT-Nn, 1CZ4). Since all known double-psi barrels are single-chain, we assume that they originated from a single-chain version of the RIFT barrel; we have however made a homo-dimeric double-psi barrel in the laboratory, showing that a pathway in which the strand swapping event preceeded the duplication and fusion cannot be excluded (dotted arrows). Finally, the B3 barrel (e.g. RAV1, 1WID) may have arisen through the fusion of a homo-dimeric RIFT barrel with a swapped-hairpin barrel, although the homology is not clear in this case (dotted arrows). Coloring is as in Fig. 6.2; the invading strand is shown in red and the swapped strands in blue and cyan.

high a sequence complexity and too low a folding yield for random assembly, but we are aware of one example, where a novel, folded protein of 80 residues capable of binding ATP was selected from a random peptide library by *in vitro* evolution.[4,38] This demonstrates, we think, that structure and function can evolve, given time and sufficient starting material. It is, however, entirely unclear how this starting material would have come about. No abiotic processes are known that can produce chains of more than five to 10 amino acids, and these processes are very inefficient. There is also the fundamental question of how the sequence information contained in the evolving proteins could be passed on, since this is an absolute prerequisite for any evolution, biotic or abiotic.

Proteins may have originated by the repetition of short peptides, a process that efficiently yields fibrous proteins such as coiled coils and β-helices.[39,40] Repetitive sequences appear to have a higher chance of folding and also more favorable structural properties than non-repetitive sequences.[41,42] The problem of passing on the sequence information, however, remains unsolved. Also, domains seen today do not have fibrous elements at their core; there is a discontinuity in fold complexity between fibers and all other folded domains and fibers are structural, not catalytic elements, whereas the primary role of proteins is catalysis.

We favor a scenario for the origin of proteins by fusion and recombination from an ancestral set of peptides, which emerged in the context of RNA-dependent replication and catalysis (the "RNA world").[15] These peptides, originally short chains of abiotic origin, would have been selected as co-factors of ribozymes, broadening their catalytic spectrum and improving their stability and folding efficiency. As the abiotic pool became depleted, ribozyme-based organisms developed an evolutionary incentive to ligate peptides catalytically, and later also to establish a primitive code so as to increase the yield of useful peptides. The need for improved specificity provided the evolutionary pressure for the emergence of peptides capable of assuming secondary structure on an RNA scaffold. The assembly of longer polypeptide chains from these pre-optimized peptides led to folding as an emergent property, when peptides found that they could now exclude water between themselves (“hydrophobic collapse”) in the absence of an RNA scaffold. The dominant role of recurrent super-secondary structures in the architecture of modern folds[43] may be the result of this process.

Whatever the mechanism, it appears to have ceased a long time ago, since the basic complement of proteins in living beings has not been enriched by new folds for hundreds of millions of years and has probably been essentially stable since the time of the last common ancestor. Why is that? Did nature find most islands of stability available to the 20 natural alpha-amino acids in one burst around 3.8–3.5 billion years ago? Or is it that, once a set of folded and functional proteins was in place, no new exemplars could emerge across the complexity boundary imposed by the twin constraints of structure and function, without being eliminated immediately by established competitors? The issues resemble the questions surrounding animal body-plans. These also emerged in a comparatively short time (the “Cambrian explosion”) and only a very limited number became established. Even though new opportunities arose periodically through large-scale extinction events, none led to the emergence of new body-plans; rather, the openings were filled by survivors with the same or similar body plans as the extinct species.

6.7 Future Outlook

Clearly, an important challenge in the coming years will be to establish why the basic complement of protein domains is so stable. Has nature identified and populated all the main islands of fold stability, or are there new islands that remain to be discovered? So far, efforts to generate new folds by design[44] have stayed very close to already observed topologies and there is little expectation that these folds will indeed turn out to be absent from nature. We therefore consider this question to be essentially unaddressed at present.

A second major challenge will be to explore the hypothesis that folded proteins arose from an ancestral pool of peptides, the antecedent domain segments.[45] If true, one would expect to observe similar fragments in proteins with dissimilar folds as vestigial traces of this process. We note that the success of fold predictors based on fragment libraries, such as Rosetta,[46] may be due to their ability to track this ancient peptide set. Systematic studies should allow its description in the same way in which ancient vocabularies have been reconstructed from the comparative study of modern languages. Peptides from this set should also be of great interest for protein engineering, as we anticipate that they will turn out to be structurally dominant in the sense discussed above for the CspA β-meander. Assembly of new proteins around such stable peptides may be more feasible than the current whole-chain optimization approach.

We see a further challenge at the junction between the two challenges named above: if folded proteins evolved from a set of peptides selected on an RNA scaffold, might a different scaffold have led to a different set of peptides with different secondary structures opening onto a different fold space? Is there a fold space out there accessible with the natural 20 amino acids, which is not built of α-helices and β-sheets? We note that in his seminal paper on the structure of the α-helix, Pauling in fact proposed two helical structures for the polypeptide chain,[47] one of which has never been observed. Is this because it does not represent a stable solution, or is it because its dimensions do not fit the grooves of nucleic acids? While these questions may seem to cross the threshold into science-fiction, the only thing one can expect with any degree of certainty in science is surprise.

References

1. Ward CL, Omura S, Kopito RR. (1995) Degradation of CFTR by the ubiquitin-proteasome pathway. *Cell* **83**: 121–127.
2. Phipps BM, Hoffmann A, Stetter KO, Baumeister W. (1991) A novel ATPase complex selectively accumulated upon heat shock is a major cellular component of thermophilic archaebacteria. *EMBO J* **10**: 1711–1722.
3. Gregersen N, Bross P, Vang S, Christensen JH. (2006) Protein misfolding and human disease. *Ann Rev Genomics Hum Genet* 7: 103–124.
4. Keefe AD, Szostak JW. (2001) Functional proteins from a random-sequence library. *Nature* **410**: 715–718.
5. Wei Y, Kim S, Fela D, Baum J, Hecht MH. (2003) Solution structure of a de novo protein from a designed combinatorial library. *Proc Natl Acad Sci USA* **100**: 13270–13273.

6. Matsuura T, Ernst A, Pluckthun A. (2002) Construction and characterization of protein libraries composed of secondary structure modules. *Protein Sci* **11**: 2631–2643.
7. Riechmann L, Winter G. (2000) Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. *Proc Natl Acad Sci USA* **97**: 10068–10073.
8. Russell RB. (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol* **279**: 1211–1227.
9. Doolittle RF. (1994) Convergent evolution: the need to be explicit. *Trends Biochem Sci* **19**: 15–18.
10. Chothia C. (1992) Proteins. One thousand families for the molecular biologist. *Nature* **357**: 543–544.
11. Wolf YI, Grishin NV, Koonin EV. (2000) Estimating the number of protein folds and families from complete genome data. *J Mol Biol* **299**: 897–905.
12. Orengo CA, Thornton JM. (2005) Protein families and their evolution — a structural perspective. *Ann Rev Biochem* **74**: 867–900.
13. Marsden RL, Lewis TA, Orengo CA. (2007) Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinform* **8**: 86.
14. Grishin NV. (2001) Fold change in evolution of protein structures. *J Struct Biol* **134**: 167–185.
15. Soding J, Lupas AN. (2003) More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays* **25**: 837–846.
16. Xia Y, Levitt M. (2004) Simulating protein evolution in sequence and structure space. *Curr Opin Struct Biol* **14**: 202–207.
17. Koretke KK, Russell RB, Lupas AN. (2002) Fold recognition without folds. *Protein Sci* **11**: 1575–1579.
18. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* **53 Suppl 6**: 491–496.
19. Soding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**: 951–960.
20. James LC, Tawfik DS. (2003) Conformational diversity and protein evolution — a 60-year-old hypothesis revisited. *Trends Biochem Sci* **28**: 361–368.
21. Lupas AN, Gruber M. (2005) The structure of alpha-helical coiled coils. *Adv Protein Chem* **70**: 37–78.
22. Glykos NM, Cesareni G, Kokkinidis M. (1999) Protein plasticity to the extreme: changing the topology of a 4-alpha-helical bundle with a single amino acid substitution. *Structure* **7**: 597–603.
23. Meier S, Jensen PR, David CN, Chapman J, Holstein TW, Grzesiek S, Ozbek S. (2007) Continuous molecular evolution of protein-domain structures by single amino acid changes. *Curr Biol* **17**: 173–178.

24. Cordes MH, Burton RE, Walsh NP, McKnight CJ, Sauer RT. (2000) An evolutionary bridge to a new protein fold. *Nat Struct Biol* 7: 1129–1132.
25. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci USA* **104**: 11963–11968.
26. Krishna SS, Sadreyev RI, Grishin NV. (2006) A tale of two ferredoxins: sequence similarity and structural differences. *BMC Struct Biol* **6**: 8.
27. Venclovas C, Zemla A, Fidelis K, Moult J. (2003) Assessment of progress over the CASP experiments. *Proteins* **53 Suppl 6**: 585–595.
28. Hadley C, Jones DT. (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* 7: 1099–1112.
29. Day R, Beck DA, Armen RS, Daggett V. (2003) A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci* **12**: 2150–2160.
30. Alva V, Ammelburg M, Soding J, Lupas AN. (2007) On the origin of the histone fold. *BMC Struct Biol* 7: 17.
31. Vogel C, Morea V. (2006) Duplication, divergence and formation of novel protein topologies. *Bioessays* **28**: 973–978.
32. Jung J, Lee B. (2001) Circularly permuted proteins in the protein structure database. *Protein Sci* **10**: 1881–1886.
33. Coles M, Diercks T, Liermann J, *et al.* (1999) The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple beta-alpha-beta-beta element. *Curr Biol* **9**: 1158–1168.
34. Coles M, Hulko M, Djuranovic S, *et al.* (2006) Common evolutionary origin of swapped-hairpin and double-psi beta barrels. *Structure* **14**: 1489–1498.
35. de Bono S, Riechmann L, Girard E, Williams RL, Winter G. (2005) A segment of cold shock protein directs the folding of a combinatorial protein. *Proc Natl Acad Sci USA* **102**: 1396–1401.
36. Fischer N, Riechmann L, Winter G. (2004) A native-like artificial protein from antisense DNA. *Protein Eng Des Sel* **17**: 13–20.
37. Coulson AF, Moult J. (2002) A unifold, mesofold, and superfold model of protein fold use. *Proteins* **46**: 61–71.
38. Lo Surdo P, Walsh MA, Sollazzo M. (2004) A novel ADP- and zinc-binding fold from function-directed *in vitro* evolution. *Nat Struct Mol Biol* **11**: 382–383.
39. Chen L, DeVries AL, Cheng CH. (1997) Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc Natl Acad Sci USA* **94**: 3811–3816.
40. Liou YC, Tocilj A, Davies PL, Jia Z. (2000) Mimicry of ice structure by surface hydroxyls and water of a beta-helix antifreeze protein. *Nature* **406**: 322–324.
41. Main ER, Lowe AR, Mochrie SG, Jackson SE, Regan L. (2005) A recurring theme in protein engineering: the design, stability, and folding of repeat proteins. *Curr Opin Struct Biol* **15**: 464–471.

42. Binz HK, Kohl A, Pluckthun A, Grutter MG. (2006) Crystal structure of a consensus-designed ankyrin repeat protein: implications for stability. *Proteins* **65**: 280–284.
43. Salem GM, Hutchinson EG, Orengo CA, Thornton JM. (1999) Correlation of observed fold frequency with the occurrence of local structural motifs. *J Mol Biol* **287**: 969–981.
44. Kuhlman B, Dantas G, Ireton GC, *et al.* (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**: 1364–1368.
45. Lupas AN, Ponting CP, Russell RB. (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol* **134**: 191–203.
46. Rohl CA, Strauss CE, Misura KM, Baker D. (2004) Protein structure prediction using Rosetta. *Meth Enzymol* **383**: 66–93.
47. Pauling L, Corey RB, Branson HR. (1951) The structure of Proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA* **37**: 205–211.

This page intentionally left blank

Chapter 7

Classification of Protein Structures

A. Cuff, O. Redfern* and C. Orengo

7.1 Introduction

Since the determination of the first protein structure (of Myoglobin) in the early 1970s and the establishment of the Protein structure Data Bank (PDB) shortly afterwards in the United States,[1] the number of solved protein structures has continued to rise at an exponential rate, with more than 47 000 entries in the PDB as of July 2007. In order to conveniently organize these data for analysis, resources for classifying the structures into evolutionary families (e.g. CATH[2] and SCOP[3]) arose in the 1990s. To facilitate classification, the majority of resources employed a new generation of sensitive structure comparison methods (see Ref. 4 for review). As domains are thought to be the primary unit of evolution, CATH and SCOP generally first split whole protein chains into their component domains prior to classifying them into families.

In this chapter, we review the computational approaches that have been developed for identifying domains and evolutionary relationships between protein structures. We summarize the most popular methods for detecting whether two protein domains share structural

*Corresponding author.
Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT. Email: ollie@biochem.ucl.ac.uk.

similarities (a common fold) and what additional evidence needs to be considered to infer homology.

In addition, we present the most regular and/or common structural architectures observed in nature and assess the extent to which different fold groups and architectures are predicted to recur in completed genomes. Finally, we consider how the current repertoire of structural data influences our understanding of how new functions emerge in the genomes and discuss how the classification resources should be developed to capture information on structural and functional divergence in superfamilies more effectively.

7.2 Recognizing Domain Boundaries in Multi-domain Structures

At the beginning of 2007, nearly 40% of structures classified in the CATH domain resource were multi-domain proteins. The majority comprise just two domains, although other large multi-domain structures were also observed (see Fig. 7.1). Indeed, calculations on

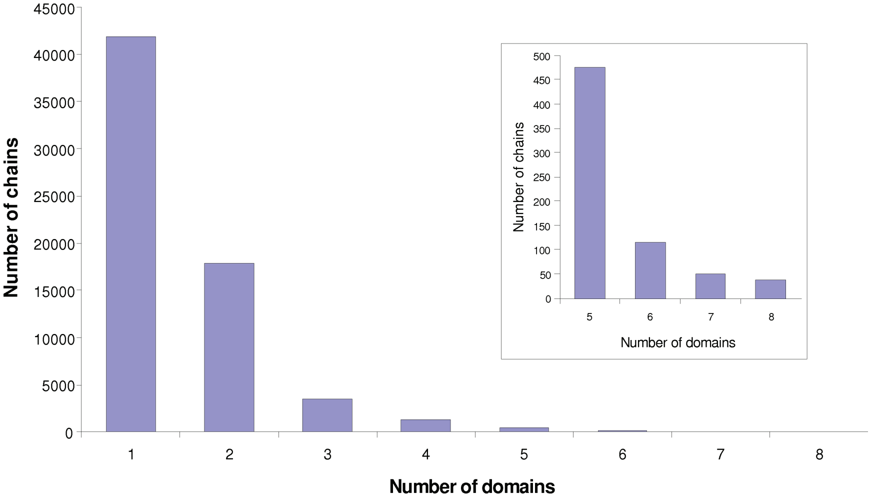


Fig. 7.1 Plot showing the number of multi-chain structures in CATH (version 3.1). The number of chains is on the *Y*-axis and the corresponding number of domains on the *X*-axis. The inset gives a close-up view of the number of chains with five or more domains in the CATH database.

predicted domains in sequences from completed genomes have suggested that at least 60% (prokaryotes) and 80% (eukaryotes) of proteins are likely to be multi-domain.[5] Hence, the proportion of multi-domain structures deposited in the PDB is expected to grow with advances in technologies for structure determination resulting from the structure genomics initiatives (SGIs).

To classify whole protein structures from the PDB, one of the first steps is the recognition of individual domains. However, this often poses a challenge due to the lack of a clear definition of what constitutes a domain. Although various heuristic definitions have arisen over the years (see, for example, in Ref. 6 — principally that domains are compact globular units, that secondary structures are rarely shared between domains, and that there are generally more residue contacts within domains than between domains — none of these concepts has successfully been encoded in an algorithm to recognize domain boundaries with higher than ~70% accuracy.[7] Moreover, even the most successful computational approaches only agree on their domain boundary prediction around 20% of the time,[7] which means that the results of all methods are often manually checked to ensure accuracy. For example, in the CATH classification, three independent algorithms — PUU,[8] DOMAK, DETECTIVE[9] — are combined to inform the manual curation process.

To address the variation in the structural properties of domains, many classification resources [e.g. SCOP,[3] CATH,[2] DALI Domain Database (DDD)] instead exploit the principle of "fold recurrence" to look within a library of previously classified domains for similar domains in the query structure. In fact, this approach is becoming increasingly practical as fewer than 2% of newly-solved domains are found to have new folds, although this percentage is higher for structures solved by structural genomics initiatives.[10,11] As a consequence, recurrence methods can often predict reliable boundaries more frequently than those that apply a more generic domain definition.

For instance, a new method developed for classifying structures in CATH (CATHEDRAL — CATH's Existing Domain Recognition Algorithm[12]) combines two established structure comparison algorithms to recognize existing CATH domain folds in new structures and is able to predict the correct domain boundaries ~80% of the time.

7.3 Recognizing Structural Similarities

Whether structure classification is effected at the domain or whole protein chain level, the primary aim is to group together similar structures. To keep pace with the exponential growth of the PDB, the majority of protein classification resources make use of computational structure comparison methods. For the purposes of identifying evolutionary relationships, there are two main difficulties facing these algorithms. First, distantly related domains are likely to have undergone a significant number of residue substitutions, insertions, and deletions (indels) over evolutionary time. Second, the extent to which these occur varies considerably across different protein families.

7.3.1 *Structural Variation between related Protein Structures*

Recent analysis of structural families in CATH has shown that as few as 40% of residues, usually in the hydrophobic core of the domain, are structurally conserved in most families.[13] In these remote homologues (<30% sequence identity), which are more likely to be paralogues arising from domain duplication events, the structures can change considerably, and frequently, this is accompanied by some change in the function too. However, orthologous domains arising from speciation events are much more likely to have similar structures and functions.

Building on early analyses of the PDB,[14] Reeves *et al.*[13] recently confirmed that quite extensive insertions (frequently >10 residues) can occur between distant relatives in a superfamily. These residue insertions often occur in the loops between secondary structures. Moreover, mutations in the core can produce substantial shifts in the orientations of equivalent secondary structures, although, on average, most pair-wise orientations vary by less than 20°.

Below, we review some of the most widely used methods that are either routinely used for classifying structures into fold groups and evolutionary families (e.g. in CATH or SCOP), or are available through public servers (e.g. attached to the PDB web sites at the EBI or RCSB[1]). Further details on all these methods and a more comprehensive review of structure comparison methods can be found in several reviews published recently.[4,15]

7.3.2 *Rigid Body Superposition and Quantifying Structural Similarity*

One of the earliest methods of structural comparison was developed by Rossmann and Argos using a "rigid body superposition" approach, which minimized the distance between the two proteins by superimposing the equivalent Cα atoms of one protein structure on top of the other. This is achieved by translating both protein structures to a common position in the co-ordinate frame of reference, and then rotating one structure relative to the other until the distances between the superimposed atoms are minimized.

The distances between the equivalent atoms in the two structures are measured using a function referred to as the Root Mean Squared Deviation (RMSD) (see Equation 7.1 below), with a low RMSD value indicating a closer structural similarity. An RMSD of less than 3.5Å is indicative of significant fold similarity and possible structural homology. However, if the aim is to assess global structural similarity between two proteins it is also important to consider the number of residues over which the RMSD has been calculated. Small highly recurrent super-secondary motifs, e.g. recurring $\alpha\beta$ motifs, can result in low RMSD values between two proteins arising from these small common motifs, and these can obscure large differences between the structures, which only become apparent when the whole structures are considered.

Kolodny and co-workers[16] have suggested normalizing the RMSD on the basis of the number of residues aligned using the formula shown in Equation 7.2 below. However, for the purpose of detecting significant fold similarities between domains, it is still important to calculate the number of aligned residues with respect to the larger structure. This may achieved by normalizing the RMSD on the basis of the larger domain (see Equation 7.3), as is done for classifying structures in CATH.

$$\text{RMSD} = \frac{\sum_{i=1}^{N} d_i^2}{N} \tag{7.1}$$

Expression for RMSD, where *d* is the distance between two equivalent residues after superposition and *N* is the number of aligned residues.

$$\text{SAS} = 100\frac{\text{RMSD}}{N} \tag{7.2}$$

Expression for SAS, where *N* is the number of aligned residues

$$\text{SiMax} = \max(l1, l2)\frac{\text{RMSD}}{N} \tag{7.3}$$

Expression for SiMax, where *N* is the number of aligned residues and *l*1 and *l*2 are the lengths of the two superposed structures.

At this point, it is important to note that rigid body superposition requires prior knowledge of equivalent residues and is often applied after two structures have been aligned from their sequences, or by using one of the methods described below.

7.3.3 *Approaches for Comparing Secondary Structures between Proteins*

Since most residue insertion/deletions (indels) occur in the random coil regions connecting secondary structures, a number of structure comparison methods have evolved that disregard these loops, concentrating solely on the secondary structure elements. Many of these approaches rely on graph theory due to its performance and accuracy.

The use of graph theory to compare secondary structures between protein structures was pioneered in 1993 by Artymiuk, Willett, and co-workers.[17] A protein structure can be represented as a simplified two-dimensional map or "graph". Each secondary structure element in the graph is represented by a point or node in the graph and labeled according to whether it is an α-helix or β-strand. Geometric relationships between the secondary structures, for example, distances or angles, are represented by the lines (or edges) that connect the nodes in the graph.

A recent implementation was developed for classifying domains in CATH (GRATH[18]). Graph edges connecting nodes are labeled with mid-point distances and both the tilt and the rotation between secondary structures. The algorithm detects common secondary structure "cliques" to identify equivalent secondary structures in a given pair of proteins. The size of this clique is converted to a statistical score (E-value) to quantify structural similarity.

SSM is another recent graph theory method developed by Krissinel and colleagues[19] and available through the SSM server at the EBI. SSM labels edges between nodes with distances and angles in much the same way as GRATH but places greater emphasis on how similar the secondary structure elements are in terms of size (i.e. number of residues).

The VAST algorithm[20] also focuses on secondary structure relationships. Two proteins are aligned by identifying equivalent "units" of secondary structure elements having similar orientations and sequential connections. An optimal superposition score is calculated across all pairs of equivalent secondary structure elements (SSEs) and the program assesses the probability of this score being observed by chance by superposing random pairs of SSE combinations for the two structures being compared.

One of the major advantages of using a comparison method based on secondary structure matching is the speed of performance, as there are typically an order of magnitude fewer secondary structures than residues within a protein. Secondary structure based approaches are often used for rapidly identifying putative fold matches, which can then be realigned using slower, but generally more accurate, residue-based approaches.

7.3.4 Residue-based Approaches for Comparing Secondary Structures

There are a number of widely used residue-based methods for comparing protein structures. Some of the most popular methods used by the classification resources (e.g. SCOP, CATH) and structural biology community (e.g. PDB) will be reviewed here. DALI[21] and CE[22] overcome the problems of indels between remote homologues by first splitting the structures into fragments and then concatenating equivalent fragments into a global alignment. While in the SSAP,[23] STRUCTAL,[16] and CATHEDRAL[12] algorithms, dynamic programming is exploited to cope with indels. All these algorithms are publicly available.

The DALI algorithm[21] splits the structures of the two proteins being compared into fragments of six residues (hexapeptides) and then compares the contact maps of these fragments. The contact maps capture information on residues in contact with each other within a threshold distance, e.g. 8 Å. Equivalent fragments are identified by looking for similar patterns of distances between residues within a given threshold. These matching pairs are then extended to increase the alignment length by concatenating other equivalent fragment pairs between the two proteins using a Monte Carlo optimization. The RMSD between the two concatenated structures is measured after each concatenation to assess the quality of the alignment as it grows.

The Combinational Extension (CE) algorithm[22] works in a similar manner to DALI by splitting each protein into fragments, identifying equivalent protein fragments, and then combining them to calculate a global alignment. As with DALI, variable loop regions are omitted to improve the quality of the alignment. However, CE splits proteins into octapeptides rather than hexapeptides, and aligns equivalent residues according to local geometry characteristics. Matching fragments are referred to as Aligned Fragment Pairs (AFPs) and are concatenated using a heuristic method with gaps inserted where required. The concatenated AFPs with low RMSD values are then accurately aligned using dynamic programming.

In contrast to DALI and CE, SSAP[23] captures the residue relationships by measuring the vectors between them. Vectors are determined within a common co-ordinate frame based on the local geometry of the Cα atoms, which can help in accommodating shifts in the orientation of equivalent secondary structure elements. Proteins are aligned by performing dynamic programming at two levels, first, to discover putatively equivalent residues by comparing sets of vectors for selected pairs of residues between the two proteins. Alignments from high-scoring residue pairs are then accumulated in a summary score matrix, which is re-analyzed by dynamic programming to obtain the optimal alignment (see Refs. 12 and 23 for more details). As two levels of dynamic programming are used, the algorithm has been coined “double dynamic programming” and has been exploited for other applications (e.g. threading, see Chapter 2). SSAP generates a score between 0 and 100 (for identical protein structures), which is normalized by the size of the largest structure being compared.

Another residue based approach, STRUCTAL,[16] identifies an initial alignment between the structures and uses this to superimpose the structures by rigid body transformation to obtain a minimal RMSD. Subsequently, an optimal alignment is obtained by dynamic programming. Initial alignments are obtained in various ways, for example, by considering the sequence similarity of the proteins or torsional angle similarity. An iterative approach is employed whereby alignments are refined by dynamic programming, and this is followed by further superposition until a local optimum is converged upon. STRUCTAL provides a statistical measure of significance of the final alignment produced in the form of a *p*-value.

Recent benchmarks of several structure comparison algorithms (DALI, SSAP, STRUCTAL, CE, LSQMAN,[16] O. Redfern personal communication) has shown that DALI and SSAP are highly efficient at searching libraries of domain folds to classify a newly determined structure, frequently ranking the correct fold at the top of the list of matches. These algorithms have been specially tuned to perform well in the classification of structures into fold groups and superfamilies, and therefore, it is not surprising that they outperform other methods. Although other methods do not perform as well, they are less computationally expensive, and hence, often better suited to database searching.

7.4 Identifying Evolutionary Relationships

Structural similarity is not always a sufficient criterion for recognizing homologous domain structures. As there are constraints on the manner in which α-helices and β-strands can pack together in 3D, there will clearly be limits on the number of possible folds, and therefore, the recognition of structural similarity between two domain structures could simply denote convergence to energetically favorable arrangements of secondary structures. Therefore, most structural classifications seek additional evidence of homology. This may constitute an unusual sequence pattern (e.g. detectable using Hidden Markov Models (HMMs) or sequence profiles) or via evidence of functional similarity (see Section 7.9).

7.4.1 *Classifying Homologues Using Sequence Profile Methods*

Both of the major structural classifications (SCOP, CATH) perform considerable manual validation to recognize homologues. In CATH, close homologues are validated using pair-wise sequence comparison and detection of 35% or more sequence identity between domains, in addition to significant structural similarity. For remote homologues, HMM-based methods are used, namely HMMer[24] and SAM-T.[25] The more powerful Profile-Profile based approaches of PRC[26] and COMPASS[27] can further aid homologue detection. Indeed, recent benchmarking using a manually validated dataset of CATH homologues[28] has shown that COMPASS recognizes ~20 times more remote homologues in the midnight zone (<20% sequence identity) than BLAST.[29] Further, in some superfamilies, COMPASS and PRC can recognize homologues that are missed by structure comparison as the relatives have diverged significantly so that the folds can no longer be considered similar.[28]

Since structural divergence varies considerably between families (see Ref. 13), individual family-based thresholds on structure similarity are more appropriate for classifying relatives. However, the relative scarcity of structural data in most families makes it hard to do this reliably at present. For classification in CATH, the problem is resolved to some extent by using neural networks to combine information on structural similarity and sequence similarity for detecting homologues, an approach that succeeds ~95% of the time.[30]

7.5 Review of the Major Domain Structure Classifications and Structural Neighborhood Resources

Due to the large amount of expert curation required, there are currently only two manually-validated protein structure classification databases that aim to cover the entire PDB: SCOP and CATH. However, there are a number of other resources that automatically cluster structurally similar structures, to create structural neighborhoods (see Table 7.1).

7.5.1 *The CATH Database*

CATH[2] is an acronym for Class, Architecture, Homologous Superfamily, and Topological motif, the four major levels in the classification hierarchy (Fig. 7.2). Domain structures are classified in this resource according to sequence, structural, and functional similarity using both automated and manual approaches.

Domains are initially assigned to one of four (C)lasses according to their secondary structure content (i.e. containing: mainly α-helical structures, mainly β-sheet structures, mixed $\alpha\beta$ structures, or very little secondary structure content). They are then further classified according to their (A)rchitecture, which refers to the gross arrangement of

Table 7.1 Table of Protein Structure Classification Resources and Neighborhood Resources

		Structure	
Database	**Coverage (July 2007)**	**Comparison Method**	**Type**
CATH CATH is a hierarchical classification of protein domain structures, clustered by Class, Architecture, Topology and Homologous Superfamily.	93 885 domains in 2091 superfamilies.	SSAP, GRATH, CATHEDRAL	Automatic structural and sequence comparison methods are combined with manual validation of superfamily alignments and domain boundaries.
CE A database of structural alignments and similarities between all structures in the PDB.	All chains in the PDB.	CE	Fully automatic, nearest neighbors.
MMDB Contains pre-calculated pairwise structural comparisons and alignments between all structures in the PDB.	All chains in the PDB.	VAST	Fully automatic, nearest neighbors.
HOMSTRAD HOMologous STRuctures Alignment Database. Database of annotated structural alignments for homologous protein families, utilizing SCOP, Pfam and SMART to identify relatives.	3454 structures in over 1000 superfamilies.	MNYFIT, STAMP and COMPARER	Manual classification of close homologues.
SCOP Structural Classification of Proteins. Hierarchical classification by Class, Fold, Superfamily, Family.	75 930 domains in 1589 superfamilies.	None.	Manual classification.

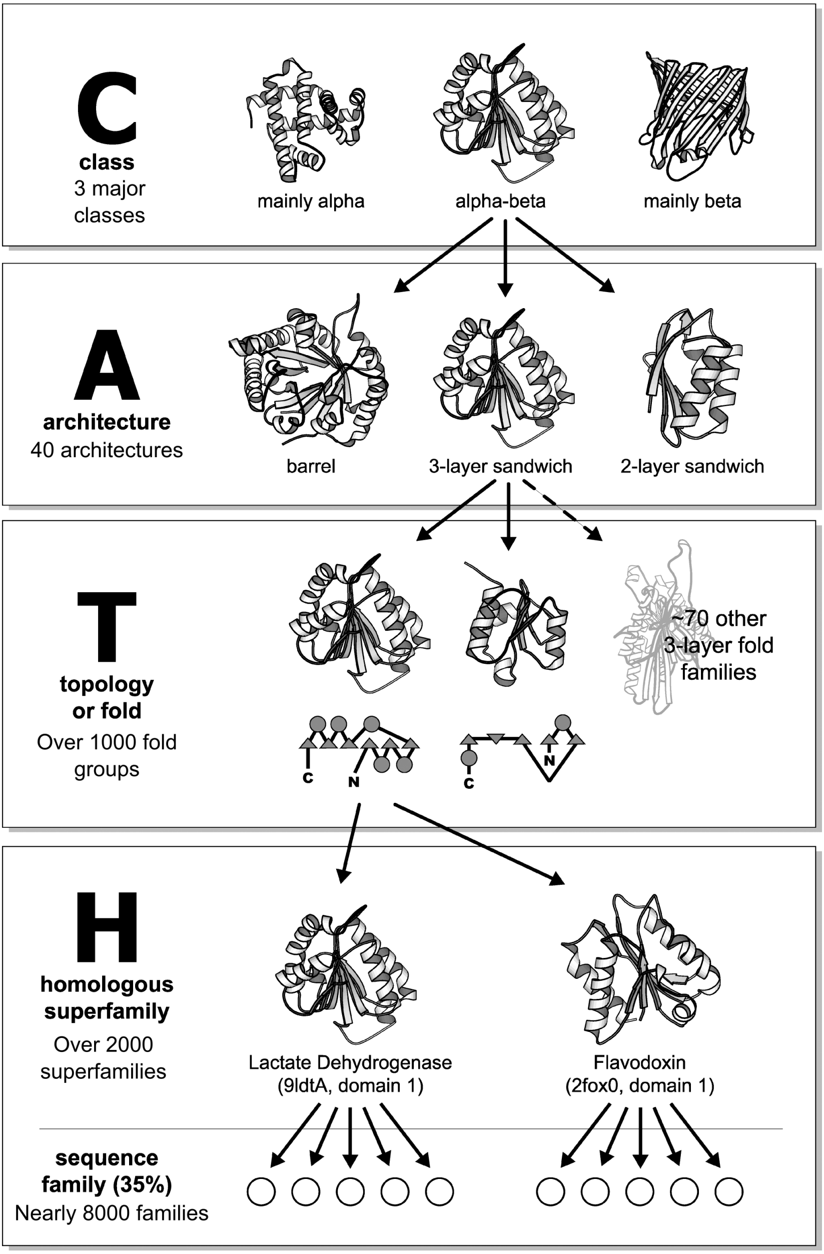


Fig. 7.2 Schematic representation of the class, architecture, topology/fold, and homologous superfamily levels in the CATH database.

secondary structures in 3D space, independent of their connectivity. Next, the (T)opological motif or fold group is determined by both the arrangement of the secondary structures and their connectivity. Finally, domains are clustered into the same (H)omologous superfamily provided there is clear indication of an evolutionary relationship. Domains should share significant structure, sequence and/or functional similarity.

Lower levels in the CATH hierarchy comprise subfamilies of domains clustered according to significant levels of sequence similarity measured by the pair-wise Needleman and Wunsch algorithm.[31] The SOL and I levels refer to groups of domains clustered together because they have at least 35%, 60%, 95%, or 100% sequence identity respectively. A final level, D, can be seen as a "counter" within the I level and is added to ensure that each domain entry in the CATH database has a unique "CATHsolid" identification code. Sub-clustering is performed at the different levels to provide increasing confidence for inheriting structural and functional properties.

As mentioned above, CATH is updated using both automated algorithms and manual curation. For newly determined structures found to be closely related (≥ 80% sequence identity) to structures already in the database, domain boundaries are assigned completely automatically using a sequence-based in-house algorithm (ChopClose[30]). For those without a close relative in CATH, several independent programs are run (see Section 7.2) and manual validation of the boundaries is performed. Information on the results of these programs and the final manual refinement of the boundaries can be viewed via the CATH update web pages(http://www.cathdb.info).

Domains are subsequently classified into fold groups in CATH by assessing their structural similarity to classified domains using the CATHEDRAL and SSAP algorithms. To recognize homologous relationships at least two out of the three following criteria must be met: (i) significant sequence similarity (by Needleman-Wunsch pair-wise method or by SAM-T or PRC HMM based approaches); (ii) significant similarity in structure (by CATHEDRAL or SSAP); and (iii) similarity in function — functional information is extracted from publicly available databases such a GO, EC, COGs, Pfam, and KEGG, and also from the relevant literature and by running function prediction algorithms (see Section 7.9).

For example, programs are run that automatically compare functional annotations between domains (SAWTED[32,33]) and a machine learning approach has been developed for combining information on sequence similarity, structural similarity, and functional similarity to gauge whether domains are likely to be evolutionary related. In cases where automatic homology assignment is not possible manual, curation is employed. The results from all the programs run on each domain can be viewed on the following CATH update pages (http://www.cathdb.info).

7.5.2 The Structural Classification of Proteins (SCOP) Database

The Structural Classification of Proteins (SCOP) was developed by Murzin and collaborators[3] in 1994, and like CATH, each protein is divided into one or more domain structures. However, unlike CATH, domain boundary assignment and classification is almost entirely achieved by manual inspection.

As with CATH classification, SCOP follows a hierarchical organization with the following major levels. In the highest level (Class), protein structures are grouped into different classes according to their secondary structure content. The five major classes in SCOP are: (i) all alpha (for structures almost entirely composed of alpha helices), (ii) all beta (for structures almost entirely composed of beta sheets), (iii) alpha/beta (structures composed of interspersed alpha-helices and beta-strands), (iv) alpha+beta (structures composed of segregated alpha-helices and beta-strands), and (v) multi-domain (structures that are composed of two or more domains that belong to different classes).

Unlike CATH, there is no architecture level in SCOP. The next major level below Class is the Fold. As with CATH, Fold Group describes how secondary structure elements are arranged, and their connectivity (i.e. topology) and the Superfamily level groups domain structures are thought to be evolutionary related. The family level groups' closely related domains are likely to have similar structures and functions.

7.5.3 *Other Structural Classification Resources*

In addition to hierarchical classifications, there are several online resources (e.g. FSSP,[34] MMDB[35]) that provide lists of structural neighbors for a given query. FSSP provides a search tool that exploits the DALI algorithm to find structural relatives. Although a high structural similarity suggests homology, it is up to the user to assess the likelihood of this based on the data provided. The MMDB exploits the vector-based VAST algorithm to automatically find similar structures within the PDB. It provides alignments annotated with automatic domain assignments and graphical structural superposition. The PDB resource itself makes use of the CE[35] program to search for structural neighbors automatically. Again, it is up to the user to further group these into individual protein families. Conversely, the HOMSTRAD database provides manually verified structural alignments for over 1000 families, often where function has been conserved (see Table 7.1).

7.6 Predicting Sequence Relatives in the Genomes and Sequence Databases

Over the last two decades, powerful new profile based sequence comparison methods (e.g. Refs. 26, 27 and 36) have been developed, some of which are capable of recognizing very remote homologues (<20% sequence identity). As described in Section 7.4.1, recent benchmarking of several of these has established safe thresholds for applying them to predict structural domains in the genomes. The most sensitive HMM-HMM-based approaches are currently too slow to use for large-scale structure prediction of genome sequences and most protocols exploit single HMM searches.

For example, structural annotations of all CATH superfamilies (Version 3.1, January 2007) have been predicted for sequences from 527 completed genomes and from Refseq and UniProt[37] (>5 million

Table 7.2 Table of Gene3D-CATH Coverage for Some Selected Model Organisms

Taxon	Number of Proteins	% Proteins with CATH Domain	% Proteins with PFAM Domain	% Proteins with CATH or PFAM Domain
Escherichia coli	4179	49.17	13.11	88.18
Neurospora crassa	9969	31.4	45.78	54.1
Mycoplasma genitalium	809	53.03	50.56	72.56
Dictyostelium discoideum	13 014	36.58	48.61	57.97
Saccharomyces cerevisiae	5586	42.32	62.37	75.58
Arabidopsis thaliana	33 097	40.77	48.9	74.45
Homo sapiens	34 888	44.26	25.49	67.56
Rattus norvegicus	11 872	53.93	35.97	84.65
Drosophilia melanogaster	16 058	42.4	41.14	70.53
Danio rerio	16 289	57.29	54.48	83.9

sequences) using the SAM-T HMM[36] method with conservative thresholds giving less than 1% error rate. The percentage of genome sequences (or residues) that can be assigned to a CATH structural family is shown in Table 7.2 for a selection of organisms from each kingdom. An average coverage of 49% of domain sequences in a genome is currently achieved, with higher levels of annotation in bacterial organisms as sequences from these organisms dominate the databases from which the HMMs are built.

CATH domain structure predictions are presented in a sister resource, Gene3D,[38] which can be accessed on the Web (URL: http://cathwww.biochem.ucl.ac.uk:8080/Gene3D/). In this resource, complete protein sequences have first been clustered into families comprising relatives with similar multi-domain architectures using a powerful new clustering protocol (TRIBE-MCL) developed by Enright and co-workers.[39] Subsequently, CATH structural domain annotations are mapped onto these sequences using HMM technologies. In addition, Pfam[40] annotations are also predicted for any structurally uncharacterized sequences or partial sequences, again using HMMs. This is applied to increase domain coverage and generate more comprehensive information on the multi-domain architecture and domain context for each domain.

In addition to the approximately 50% of domain sequences from UniProt that can be assigned to CATH structural families, a further 25% of domain sequences within a genome (on average) can be assigned to structurally uncharacterized Pfam families. The largest of these families are currently being targeted for structure determination in the Protein Structure Initiative (PSI) structural genomics projects in the US.

Recent analyses of the first 255 structures solved for these families suggests that targeting these families is considerably expanding our knowledge of the fold repertoire. Nearly 40% of them are very remote structural homologues of existing CATH superfamilies, having no close structural relative that can be superimposed within RMSD <5A, and ~4% of them are completely novel domains representing new fold groups. This contrasts with a much lower proportion (on average <1%) of novel folds currently being solved by traditional structural biology (Fig. 7.3).

Other structural genome annotation resources include SUPERFAMILY[41] based on SCOP, which also uses HMMs to assign domain superfamilies to genomes. The 3Dgenomics resource[42] uses PSI-BLAST[29] and HMMer to assign SCOP superfamilies, with PSI-BLAST optimized for genome annotation.[43] Whilst the Genomic Threading Database[44] uses PSI-BLAST and a threading related method to validate matches (see Chapter 2 and Ref. 28 for a review of these approaches).

7.7 Population Statistics from Domain Structure Classifications

As of August 2007, there were 47 251 entries in the protein databank comprising a total of 113 978 chains. The proportion of these chains classified in the CATH and SCOP databases are shown in Table 7.3. It can be seen that both SCOP and CATH now recognize approximately 1000 different fold groups comprising over 1500 evolutionary superfamilies. Recent comparisons of SCOP and CATH[45]

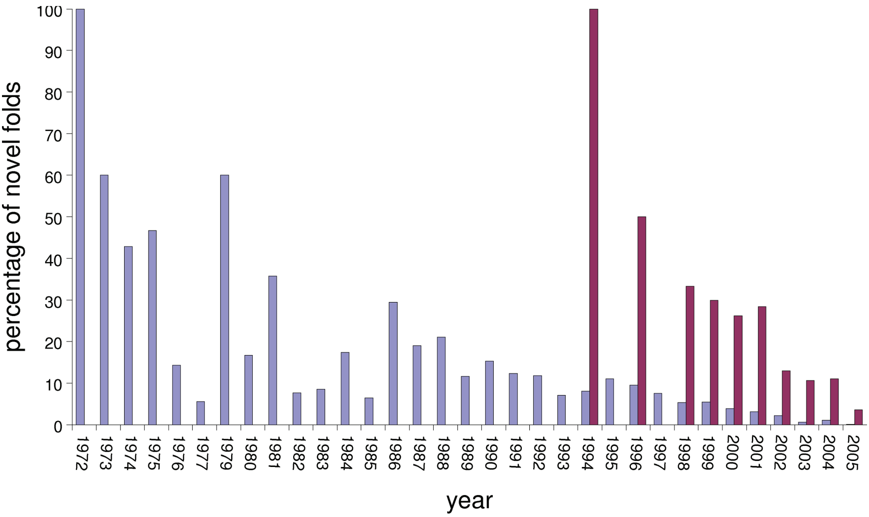


Fig. 7.3 Plot showing the percentage of structures solved each year found to be novel folds. Percentages were calculated for structures solved by either structural biology (blue bars) or structural genomics initiatives (maroon bar) by the total number of classified structures solved by that method released in a particular year.

Table 7.3 Table Showing the Population of the First Four Levels in the CATH and SCOP Hierarchical Protein Structure Databases

CATH Hierarchy	Number of CATH Domain Representatives	Number of SCOP Domain Representatives	SCOP Hierarchy
Class	4	7	Class
Architecture	1084	971	Fold
Topology	2091	1589	Superfamily
Homologous Superfamily	7794	3004	Family

have revealed that there is agreement on about 70% of the homologous superfamilies. This was determined by identifying common CATH/SCOP superfamilies in which 75% or more of relatives matched between the two resources where matching implies that at least

80% of the domain residues (normalized by the larger domain) were identical.

There is less agreement at the level of fold group, which is a more subjective measure of similarity. For example, SCOP uses largely manual identification of the fold group and tends to break the large Rossmann fold group classified in CATH into several smaller fold groups although these all comprise relatives exhibiting the characteristic topological crossover of the classic Rossmann fold and a core structural motif of four β-strands and two α-helices. For some purposes, though, it might actually be more practical to sub-divide this fold group on the basis of the size of the domain as there is significant structural variation across the group.

The skewed population of fold groups, which is evident in both resources, is illustrated for CATH in Fig. 7.4. It can be seen that whilst the majority of fold groups in CATH contain only one

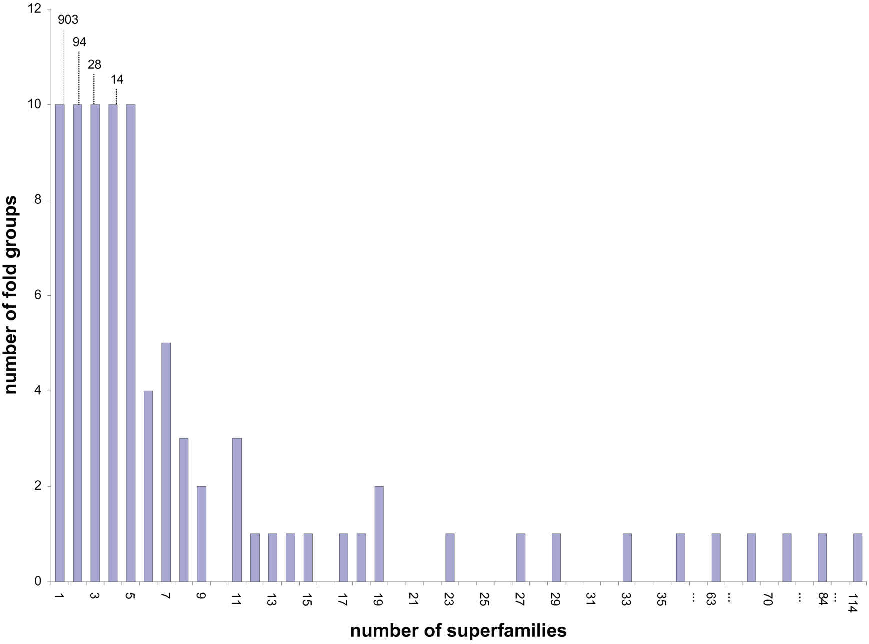


Fig. 7.4 Plots showing the population of each fold group within CATH (version 3.1).

homologous superfamily, a small percentage of fold groups include three or more different homologous superfamilies.

Although, there is no clear correlation of fold group with function,[46] some fold groups are predominantly associated with a particular molecular functional class in the COGs functional database. For example, the arc repressor fold domains are largely involved in binding to DNA and in regulation, whilst the TIM barrels operate mainly as enzymes. However, some fold groups, most notably the Rossmann and $\alpha\beta$-plait folds, are members of multiple functional classes. Indeed, relatives in the Rossmann fold group exhibit more than 200 different GO terms. Frequently, this change in function is acquired through a change in the multi-domain context.[13,47–49] The Rossmann domain is also often present to provide redox equivalents or energy for complexes on diverse biological pathways and processes.

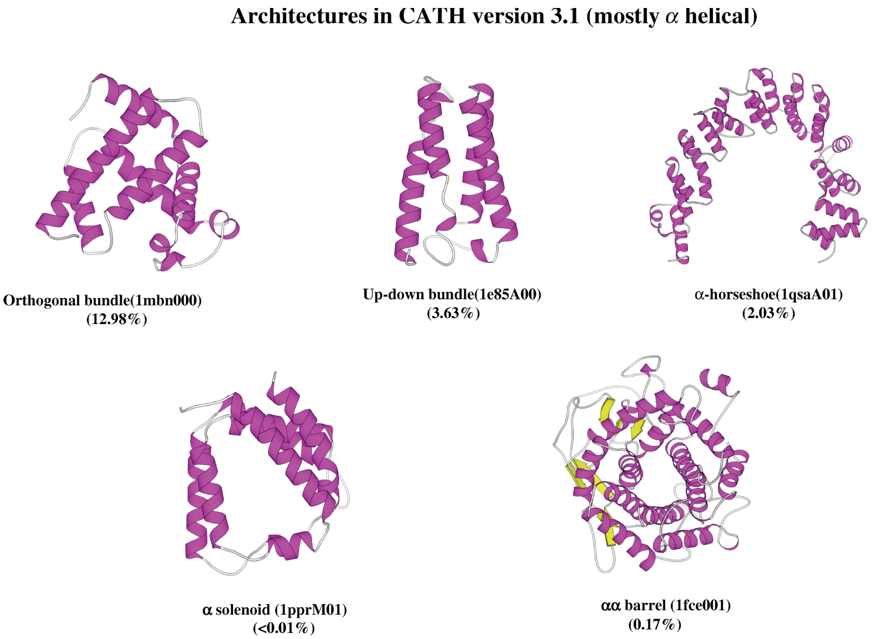
The populations of domain structure superfamilies are similarly skewed with the largest 50 superfamilies accounting for nearly 50% of sequence diverse structures in CATH (i.e. relatives clustered at 35% sequence identity) and 47% of sequence diverse relatives in Gene3D. The 20 largest CATH superfamilies account for 40% of the predicted domain structures in the genomes found in Gene3D.

In addition to the power law bias prevalent in the population of fold groups and superfamilies in CATH, there are also biases in the populations of particular architectures in CATH. Figure 7.5 shows the 27 most well-defined architectures in CATH and gives their populations in the genomes, measured according to the total number of sequences in 527 completed genomes in Gene3D. Four regular layered architectures dominate the genome annotations, namely the two and three layer $\alpha\beta$-sandwiches, the $\alpha\beta$-barrels and the 2-layer β-sandwiches.

7.8 Structural Variation in Domain Superfamilies and Correlation with Functional Modifications

The impact of mutations and indels on the protein fold or function vary according to their position in the structure. Mutations that occur in the core of the protein may cause shifts in orientation of secondary structure elements to maintain the optimal packing required for stability. Early analyses by Lesk and Chothia[49] revealed that secondary structure shifts of up to 30° occurred between relatives in α-globin and β-immunoglobulin superfamilies. More recent analyses of 294 well-populated CATH superfamilies[13] revealed that 61% of superfamilies exhibit a mean deviation in pair-wise secondary structure orientations of 8–16°, 15% of families have mean deviations less than 8°, and 14% are much more variable tolerating a mean deviation of between 16–25°. Only 5% of families exhibit mean deviations greater than 25°, although in these families some relatives show deviations as high as 80° in their secondary structure orientations.

Structural changes that arise from indels tend to occur within the loop regions of a protein's structure and thereby minimize any



Architectures in CATH version 3.1 (mostly α helical)

Orthogonal bundle(1mbn000)
(12.98%)

Up-down bundle(1e85A00)
(3.63%)

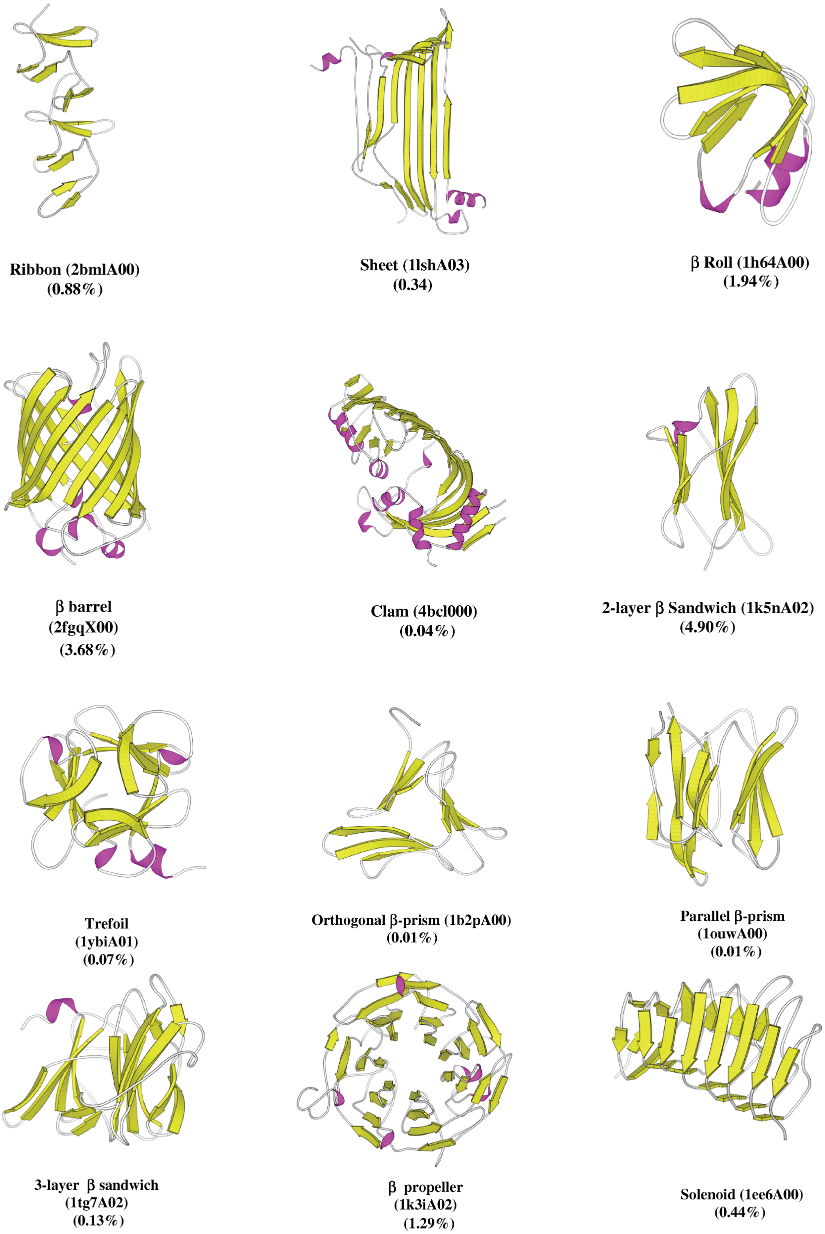
α-horseshoe(1qsaA01)
(2.03%)

α solenoid (1pprM01)
(<0.01%)

αα barrel (1fce001)
(0.17%)

Fig. 7.5 Molscript representations of the major architectures in the CATH hierarchy. The population of each architecture in the genomes is calculated as a percentage and is displayed in brackets beneath each one.

Architectures in CATH version 3.1 (mostly β sheet)



Ribbon (2bmlA00)
(0.88%)

Sheet (1lshA03)
(0.34)

β Roll (1h64A00)
(1.94%)

β barrel
(2fgqX00)
(3.68%)

Clam (4bcl000)
(0.04%)

2-layer β Sandwich (1k5nA02)
(4.90%)

Trefoil
(1ybiA01)
(0.07%)

Orthogonal β-prism (1b2pA00)
(0.01%)

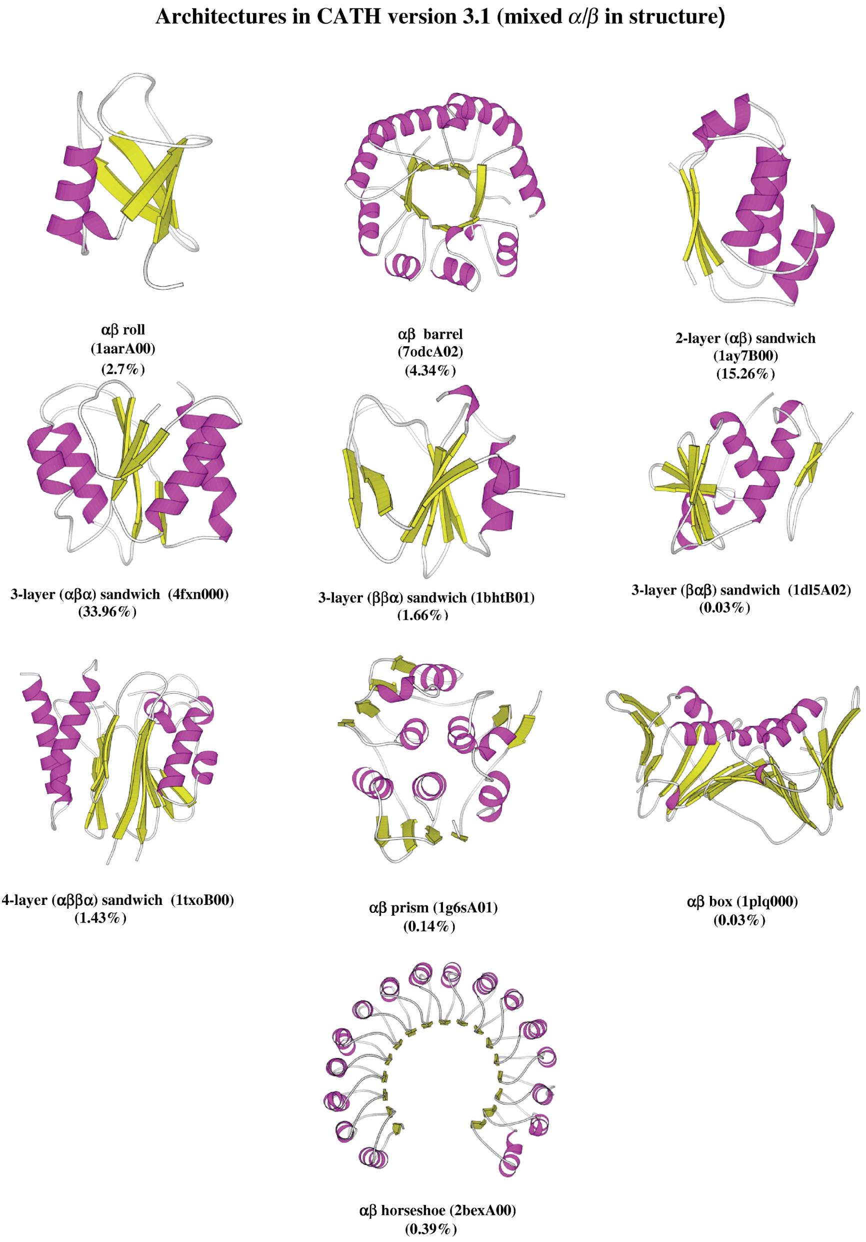
Parallel β-prism
(1ouwA00)
(0.01%)

3-layer β sandwich
(1tg7A02)
(0.13%)

β propeller
(1k3iA02)
(1.29%)

Solenoid (1ee6A00)
(0.44%)

Fig. 7.5 (*Continued*)



Architectures in CATH version 3.1 (mixed α/β in structure)

αβ roll (1aarA00) (2.7%)

αβ barrel (7odcA02) (4.34%)

2-layer (αβ) sandwich (1ay7B00) (15.26%)

3-layer (αβα) sandwich (4fxn000) (33.96%)

3-layer (ββα) sandwich (1bhtB01) (1.66%)

3-layer (βαβ) sandwich (1dl5A02) (0.03%)

4-layer (αββα) sandwich (1txoB00) (1.43%)

αβ prism (1g6sA01) (0.14%)

αβ box (1plq000) (0.03%)

αβ horseshoe (2bexA00) (0.39%)

Fig. 7.5 (*Continued*)

deleterious effects on the overall structure and stability of the fold. As initially reported by Pascarella and Argos,[14] and more recently revisited by Reeves *et al.*,[13] relatives with a high average sequence identity (from 40% to 95%) tend to have indels of no more than two residues, whilst at lower levels of sequence identity (0–10%), indels as large as 12 residues are frequently observed.

The three-layer sandwich architectures ($\beta\beta\alpha - 3.50$ and $\alpha\beta\alpha - 3.40$) are more tolerant to larger indels at all sequence identities. In these folds, insertions are often tolerated because they occur as extra strands at the edges of the β-sheets or as additional α-helices between strands, which do not disturb the overall layered architecture of the protein. In the α-horseshoe folds (1.25) up to 30% of indels are greater than 10 residues long. These often comprise a series of adjacent helices, forming super-secondary motifs, rather than single secondary structures.

The $\alpha\beta$-barrel architecture (3.20) also shows tolerance to larger insertions and these often occur as additional helices in the loops connecting β-strands in the barrel. It is possible that in both the $\alpha\beta$-sandwiches and $\alpha\beta$-barrels, extensive hydrogen bonding between β-strands in the β-sheets gives rise to a stable central framework, which is able to support greater structural variation in the remainder of the fold.

In many sequence diverse superfamilies, residue insertions give rise to additional secondary structures, which decorate or "embellish" the conserved structural core found in all relatives. Recent analysis of CATH found that for 56% of highly populated superfamilies (>9 sequence diverse relatives), there are two-fold or more increases in the numbers of secondary structures in some relatives. In some families, five-fold increases occur, sometimes modifying the fold of the domain.

Manual inspection of secondary structure insertions or embellishments in 48 particularly variable CATH superfamilies revealed that although these insertions were usually discontiguous in the sequence they were often co-located in 3D resulting in a larger structural motif that often modified the geometry of the active site or the surface conformation, thereby promoting diverse domain partnerships and

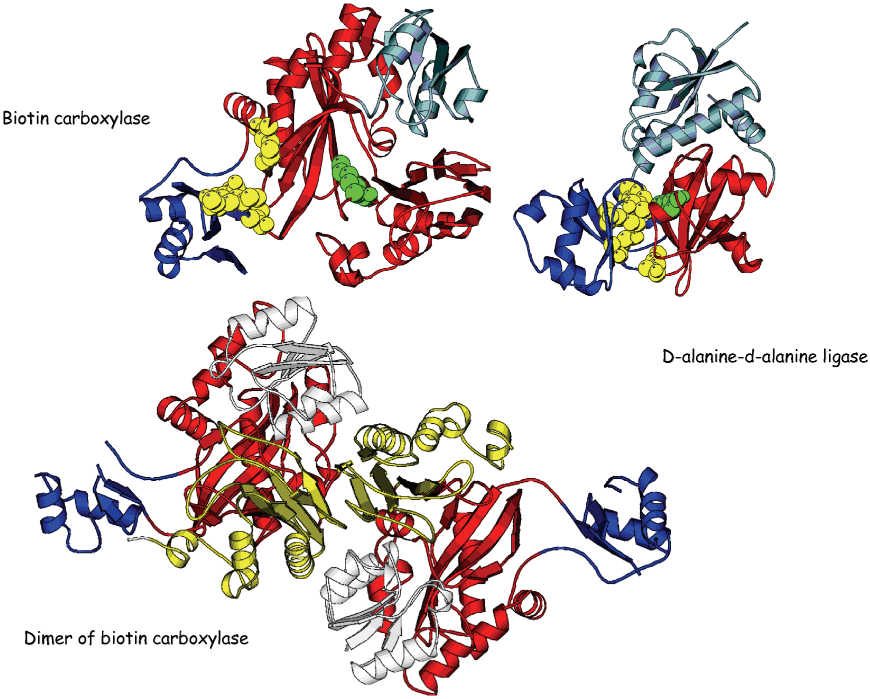


Fig. 7.6 Structurally diverse relatives from the ATP grasp superfamily in the CATH database. In red, the large domain, in blue, the small domain, and in light blue, the B domain. Residues shown in yellow are involved in ATP binding, and residues in green are involved in substrate binding.

protein interactions. These observations, supported by automatic analysis of all well-populated CATH superfamilies and shown to be statistically significant using random models, suggest that accretion of small secondary structure insertions during evolution may provide a simple mechanism for evolving new functions in diverse relatives.

Again, layered domain architectures (e.g. mainly-β and $\alpha\beta$-sandwiches), as adopted by the Rossmann fold and $\alpha\beta$-plait superfamilies, which recur highly in the genomes, more frequently exploit these types of embellishments to modify function. Because secondary structure insertions often occur at the edges, top, or bottom of the β-sheets, this gives just a few sites on the protein surface where insertions can aggregate to give much larger structural motif, impacting on functional sites

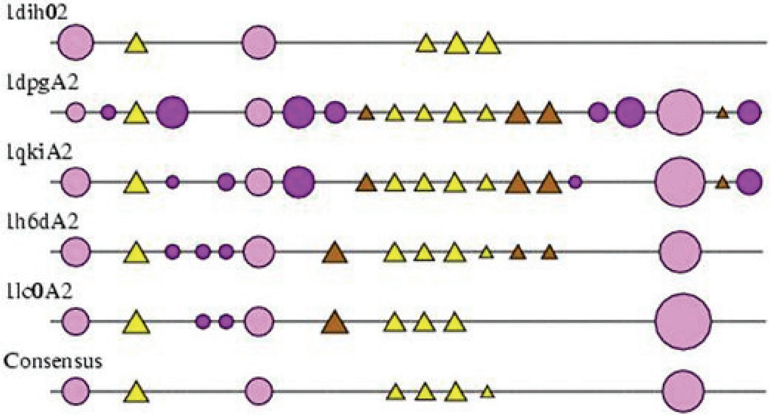


Fig. 7.7 2DSEC plot showing the embellishments present in the oligomerization domain in the NADP oxidoreductase superfamily. Alpha-helices are represented by circles and beta sheets are represented by squares.

(e.g. active sites or protein interaction surfaces). Information on structural variability across domain superfamilies has been made available through the CATH Dictionary of Homologous Structures (DHS) (URL: http://www.biochem.ucl.ac.uk/bsm/dhs/). For each superfamily and structural subfamily within it, multiple structural alignments are provided, as well as plots showing the alignment of equivalent secondary structures (Fig. 7.7). These can be used to assess the extent of structural divergence between remote homologues and structural embellishments to the domain core.

There are some superfamilies that, despite having a significant number of sequence diverse relatives (<35% sequence identity), have a high degree of structural conservation. Relatives in these superfamilies generally have highly similar functions,[13] and it is likely that the structural conservation is largely due to functional constraints.

7.9 Identifying Functional Relationships in Homologous Superfamilies

As discussed above, structural changes between homologous proteins often correlate with a divergence of function. However, in

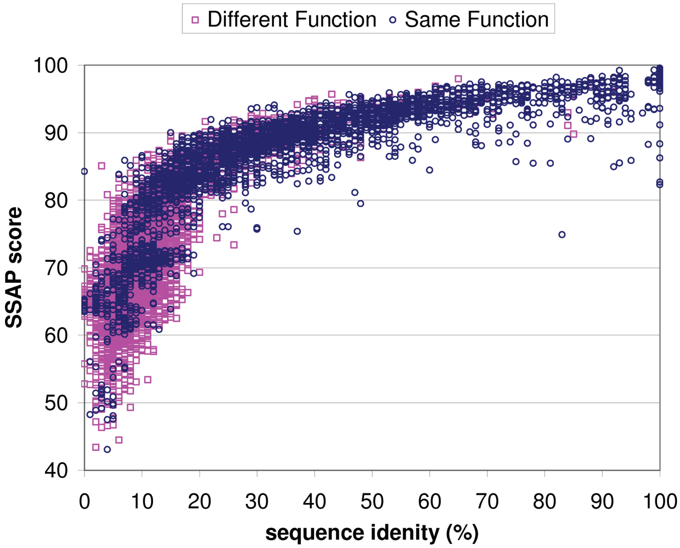


Fig. 7.8 Scatter plot showing the relationship between sequence, structure, and function of all homologues in enzyme superfamilies. Relatives having the same EC classification number are shown in blue. Those with different EC numbers are shown in pink.

highly variable superfamilies, function can still be conserved between relatives where global structural similarity is relatively low.

Figure 7.8 shows the relationship between sequence and structural similarity with respect to the conservation of function in enzyme domain superfamilies. It can be seen that once structural similarity dips below a SSAP score of 85, the majority of domains are involved in different functions. However, it is also important to note that even at structural similarities above 90, there are examples of functional divergence.

Given the complex relationship between sequence, structure, and function, further classification of structural domains into functional sub-families can prove problematic. However, there are a large number of automated methods that aim to identify domains with similar functions. We will summarize some popular approaches, focusing on current work being undertaken in the CATH group to construct functional families within homologous superfamilies.

Although in practice functional information from the literature is often available to assess the similarity of two protein functions, structured functional annotation data provides a means to computationally assess the similarity between the functions of two proteins, and hence, operate on larger data sets. For example, the Gene Ontology (GO)[50] provides annotation in three categories: molecular function, biological process, and cellular localization. Annotation terms are organized in a directed graph, which facilitates automatic comparison by methods such as GOSim.[33] The Enzyme Classification (EC) database[51] provides a similar resource for classifying enzymatic functions. Where two protein structures have GO or EC annotations, these can be compared to generate a measure of functional similarity.

Where global structure comparison fails to identify significant similarity between two proteins, detection of local motifs can prove useful. To retain a specific function through evolution, the local environment of a functional site(s) must be preserved, even if other portions of the fold have become altered, producing a relatively low global structural similarity. Indeed, enzymatic catalysis is performed by a limited set of residues that comprise the active site and the specificity of DNA-binding proteins is often conferred by relatively small regions of positive charge on the surface of the protein structure. Consequently, there are a number of methods that focus on comparing smaller structural motifs associated with a specific function.

The Catalytic Site Atlas[52] held at the European Bioinformatics Institute (EBI) is a database of protein structures whose catalytic residues (up to six per protein) have been manually annotated from the literature. Structural templates constructed from the catalytic residues of the proteins in the database and a fast search algorithm[53] can be used to identify similar enzymatic activity. Similarly, PDBSiteScan[54] is able to compare functional (SITE) records contained in the PDB structure files. Again, similarities detected between known functional sites can be indicative of a common evolutionary origin. However, it is important to bear in mind that some catalytic triads (e.g. the serine proteases) are known to exist in unrelated folds.[55]

As curated functional templates are not available for the majority of proteins, many groups have developed methods to automatically detect local structural motifs associated with function. The Reverse Template Method (RTM)[56] splits two structures into tri-peptide fragments and looks for equivalent fragments. In contrast to methods such as DALI, SSAP, and CE, RTM does not seek a global alignment but instead looks for fragments that exist in sequence-similar structural environments. This exploits the principle that residues in the functional site are often well-conserved with respect to their sequence compared to the rest of the structure.[57]

By utilizing EC and GO annotation data, a novel method (FLORA) has been developed for CATH to capture structural motifs associated with specific enzymatic functions. FLORA constructs multiple structure alignments of structurally similar domains with the same EC/GO annotations to determine regions of the structures that are significantly conserved. These motifs can then be compared to other domains within the superfamily to merge groups of domains with similar functions into families.

As no method of function prediction is able to detect similarities with greater than 90% accurate over all superfamilies, the CATH database aims to combine several lines of evidence to group domains into functional families.

7.10 The End of the Fold? Is There Evidence for a Structural Continuum?

There has been considerable speculation in the literature[58] on the existence of a structural continuum that links domains in different fold groups and challenges the value of hierarchical fold-based classifications. As discussed above, CATH analyses of well populated superfamilies have clearly revealed the phenomenon of significant structural drift in some highly divergent superfamilies. Although the number of superfamilies exhibiting structural drift is relatively few, these tend to be the most highly populated structural families in the genomes containing paralogous structures that have diverged considerably in both structure and function. For these superfamilies it is perhaps not meaningful to adopt a single fold characterization for relatives.

However, as Chothia and Lesk[59] first pointed out over two decades ago, there is still considerable structural conservation in the domain core (~40% of residues).[13] Furthermore, this topological core motif is structurally distinct from core motifs found in other superfamilies. In this sense, the hierarchical classifications of resources such as SCOP and CATH, are still valuable as the fold group (SCOP) or (T)opology level (CATH) can be thought of as grouping domains sharing conserved core topologies.

Furthermore, recent exploration of structural overlaps between fold groups and structural families based on pair-wise structure comparisons between all sequence diverse CATH domains (<35% sequence identity) has revealed very little structural overlap between superfamilies and fold groups when meaningful global similarity is sought (i.e. at least 60% of residues in the larger domain aligned against the smaller domain).

When this criteria is relaxed allowing a smaller proportion of residues to align between domains, frequent overlap is observed between superfamilies in some architectures (e.g. $\alpha\beta$-sandwiches, β-sandwiches) largely due to small structural motifs common to many different folds in nature. For example, the β-hairpin, $\alpha\beta$-motif, split $\alpha\beta$-motif. This has been characterized by Harrison *et al.*[60] who identified "gregarious" folds comprising a high proportion of these common motifs. Superfamilies with no overlap at all have very distinctive folds comprising rather unusual motifs or unusual combinations of common motifs.

7.11 Future Outlook

The structural genomics initiatives have been successful in revolutionizing structural biology, giving more sophisticated technologies that allow a higher throughput approach to structure determination. This is leading to significant increases in the number of structures solved annually (2631 in 2000, 4931 as of September 2007) By carefully targeting sequence space and protein families to reduce the redundancy in structures solved and increase the repertoire of structurally diverse representatives of fold space, the proportion of hard-to-classify structures has increased significantly over the last few years.

Fortunately, new and much more powerful methods of sequence comparison (e.g. HMM-HMM[28]) have been introduced at the same time, as well as rapid methods for recognizing domain boundaries in solved structures (e.g. CATHEDRAL[12]). These new approaches and machine learning systems that integrate results from several complimentary algorithms will help in managing the data. As the genome initiatives progress, improved integration of the sequence data from completed genomes into structural classifications will lead to further insights into protein family evolution and help us to understand how differences in protein distributions and evolution of new functions influence phenotypic variations between organisms.

The prediction of protein functions from sequence and structure remains a considerable challenge. The ambitions of structural genomics to deliberately target sequence families, which are likely to have novel structures and functions, will enrich the databases with information, which can illuminate mechanisms by which functions diverge and enhance the prediction methods. We also clearly need to improve the integration of data on protein interactions, associations and functional networks. As the structure databases expand, it will become possible to extend characterizations within structural families and to cluster relatives more reliably according to functional properties.

References

1. Berman HM, Westbrook J, Feng Z, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**: 235–242.
2. Orengo CA, Michie AD, Jones S, *et al.* (1997) CATH — a hierarchic classification of protein domain structures. *Structure* **5**: 1093–1108.
3. Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**: 536–540.

4. Redfern O, Grant A, Maibaum M, Orengo C. (2005) Survey of current protein family databases and their application in comparative, structural, and functional genomics. *J Chromatogr B Analyt Technol Biomed Life Sci* **815**: 97–107.
5. Apic G, Gough J, Teichmann SA. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* **310**: 311–325.
6. Richardson JS. (1981) The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34**: 167–339.
7. Jones S, Stewart M, Michie A, *et al.* (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci* **7**: 233–242.
8. Holm L, Sander C. (1994) Parser for protein folding units. *Proteins* **19**: 256–268.
9. Swindells MB. (1995) A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci* **4**: 93–102.
10. Todd AE, Marsden RL, Thornton JM, Orengo CA. (2005) Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol* **348**: 1235–1260.
11. Chandonia JM, Brenner SE. (2006) The impact of structural genomics: expectations and outcomes. *Science* **311**: 347–351.
12. Greene LH, Lewis TE, Addou S, *et al.* (2007) The CATH domain structure database: New protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucl Acids Res* **35**: D291–D297.
13. Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA. (2006) Structural diversity of domain superfamilies in the CATH database. *J Mol Biol* **360**: 725–741.
14. Pascarella S, Argos P. (1992) Analysis of insertions/deletions in protein structures. *J Mol Biol* **224**: 461–471.
15. Koehl P. (2001) Protein structure similarities. *Curr Opin Struct Biol* **11**: 348–353.
16. Kolodny R, Koehl P, Levitt M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol* **346**: 1173–1188.
17. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J Mol Biol* **243**: 327–344.
18. Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J, Orengo C. (2003) Recognizing the fold of a protein structure. *Bioinformatics* **19**: 1748–1759.
19. Krissinel E, Henrick K. (2004) Secondary-structure matching. (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* **60**: 2256–2268.

20. Madej T, Gibrat JF, Bryant SH. (1995) Threading a database of protein cores. *Proteins* **23**: 356–369.
21. Holm L, Sander C. (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol* **233**: 123–138.
22. Shindyalov IN, Bourne PE. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**: 739–747.
23. Taylor WR, Orengo CA. (1989) Protein structure alignment. *J Mol Biol* **208**: 1–22.
24. Eddy SR. (1996) Hidden Markov models. *Curr Opin Struct Biol* **6**: 361–365.
25. Karplus K, Barrett C, Hughey R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**: 846–856.
26. Madera M. (2006) PRC — the Profile Comparer.
27. Sadreyev R, Grishin N. (2003) COMPASS: A tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* **326**: 317–336.
28. Reid AJ, Yeats C, Orengo CA. (2007) Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. *Bioinformatics* **23**: 2353–2360.
29. Altschul SF, Madden TL, Schaffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res* **25**: 3389–3402.
30. Greene LH, Lewis TE, Addou S, *et al.* (2006) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucl Acids Res* **35**: D291–D297.
31. Needleman SB, Wunsch CD. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443–453.
32. MacCallum RM, Kelley LA, Sternberg MJ. (2000) SAWTED: Structure assignment with text description — enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics* **16**: 125–129.
33. Lord PW, Stevens RD, Brass A, Goble CA. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**: 1275–1283.
34. Holm L, Sander C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucl Acids Res* **25**: 231–234.
35. Marchler-Bauer A, Addess KJ, Chappey C, *et al.* (1999) MMDB: Entrez's 3D structure database. *Nucl Acids Res* **27**: 240–243.
36. Wistrand M, Sonnhammer EL. (2005) Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. BMC *Bioinformatics* **6**: 99.

37. Apweiler R, Bairoch A, Wu CH, *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucl Acids Res* **32**: D115–D119.
38. Yeats C, Maibaum M, Marsden R, *et al.* (2006) Gene3D: modeling protein structure, function, and evolution. *Nucl Acids Res* **34**: D281–D284.
39. Enright AJ, Van DS, Ouzounis CA. (2002) An efficient algorithm for large-scale detection of protein families. *Nucl Acids Res* **30**: 1575–1584.
40. Bateman A, Birney E, Cerruti L, *et al.* (2002) The Pfam protein families database. *Nucl Acids Res* **30**: 276–280.
41. Wilson D, Madera M, Vogel C, Chothia C, Gough J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucl Acids Res* **35**: D308–D313.
42. Fleming K, Muller A, MacCallum RM, Sternberg MJ. (2004) 3D-GENOMICS: A database to compare structural and functional annotations of proteins between sequenced genomes. *Nucl Acids Res* **32**: D245–D250.
43. Muller A, MacCallum RM, Sternberg MJ. (1999) Benchmarking PSI-BLAST in genome annotation. *J Mol Biol* **293**: 1257–1271.
44. McGuffin LJ, Street SA, Bryson K, Sorensen SA, Jones DT. (2004) The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucl Acids Res* **32**: D196–D199.
45. Hadley C, Jones DT. (1999) A systematic comparison of protein structure classifications: SCOP, CATH, and FSSP. *Structure* **7**: 1099–1112.
46. Rison SC, Teichmann SA, Thornton JM. (2002) Homology, pathway distance, and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*. *J Mol Biol* **318**: 911–932.
47. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. (2004) Structure, function, and evolution of multi-domain proteins. *Curr Opin Struct Biol* **14**: 208–216.
48. Todd AE, Orengo CA, Thornton JM. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307**: 1113–1143.
49. Lesk AM, Chothia C. (1982) Evolution of proteins formed by beta-sheets. II. The core of the immunoglobulin domains. *J Mol Biol* **160**: 325–342.
50. Ashburner M, Ball CA, Blake JA, *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–29.
51. Bairoch A. (2000) The ENZYME database in 2000. *Nucl Acids Res* **28**: 304–305.
52. Porter CT, Bartlett GJ, Thornton JM. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl Acids Res* **32**: D129–D133.
53. Barker JA, Thornton JM. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **19**: 1644–1649.

54. Ivanisenko VA, Debelov VA, Pintus SS, *et al.* (2002) PDBSiteScan: a tool for search for the best-matching superposition in the database PDBSite. *Third Int Conf Bioinform Genome Regul Struct* **3**: 149–152.
55. Wallace AC, Borkakoti N, Thornton JM. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci* **6**: 2308–2323.
56. Laskowski RA, Watson JD, Thornton JM. (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucl Acids Res* **33**: W89–W93.
57. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. (2002) Analysis of catalytic residues in enzyme active sites. *J Mol Biol* **324**: 105–121.
58. Kolodny R, Petrey D, Honig B. (2006) Protein structure comparison: implications for the nature of "fold space", and structure and function prediction. *Curr Opin Struct Biol* **16**: 393–398.
59. Chothia C, Lesk AM. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**: 823–826.
60. Harrison A, Pearl F, Mott R, Thornton J, Orengo C. (2002) Quantifying the similarities within fold space. *J Mol Biol* **323**: 909–926.

Chapter 8

Methods to Characterize the Structure of Enzyme Binding Sites

A. Kahraman* and J. M. Thornton

8.1 Introduction

Enzyme binding sites are regions on the surface of an enzyme specially designed to interact with other molecules. An enzyme can have different sorts of binding sites that differ in their functions and the molecules they bind. Amongst these, the most important is the active site, which consists of two or three parts. The first part is the catalytic site, which contains the catalytic machinery of the enzyme in the form of usually two to six amino acids that perform the catalytic reaction. The second part is the substrate binding site, which has the task of specifically recognizing the molecule upon which the enzyme acts. Besides the specificity, the substrate binding site also provides binding energy to keep the substrate bound on the active site for the time the catalytic reaction progresses. Enzymes can act on a huge variety of substrates, from small molecules, like hormones and sugar, and moderate sized molecules, like polypeptides and oligosaccharides, to macromolecules, like DNA and other proteins. Figure 8.1 shows an exemplary substrate binding site for an asparagine in the structure of the *Escherichia coli* asparagine synthetase [see also Fig. 8.2(a)].

*Corresponding author.

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB4 1BY, UK. Email: abdullah@ebi.ac.uk.

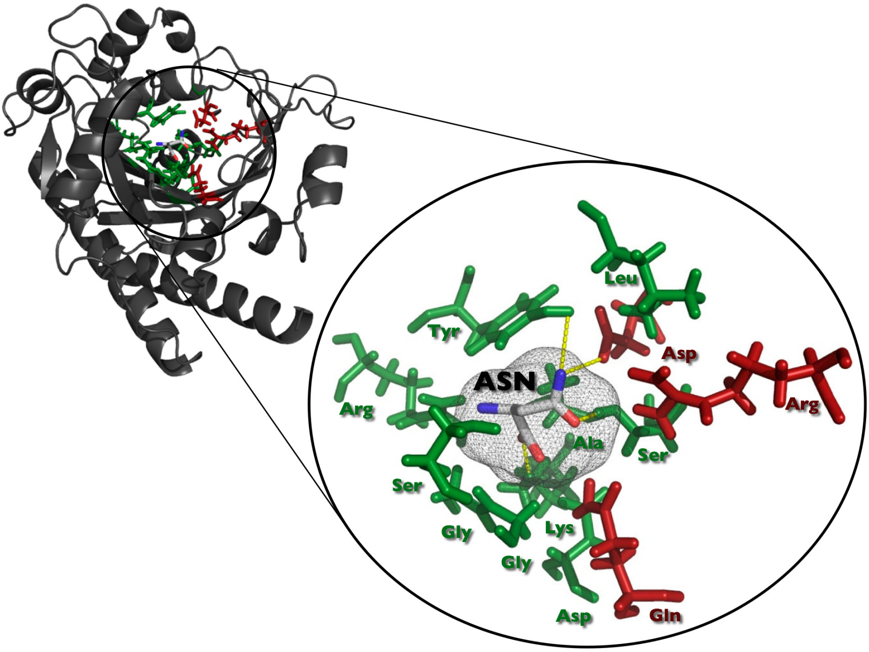


Fig. 8.1 Structure of the *Escherichia coli* Asparagine Synthetase (PDB Id: 12as) with a zoom-in into the binding site of the substrate asparagines. Binding site residues as determined by HBPLUS (see Section 8.2.9) are colored in green; catalytic active residues extracted from CSA (see Section 8.3.1) are coloured in red, and the substrate is varicolored. Hydrogen bonds between binding site residues and substrate are indicated by yellow dashed lines. The binding site shape is shown as a grey mash as approximated with spherical harmonic functions (see Section 8.2.3).

As enzymes are proteins they usually consist of 20 amino acids with either a hydrophobic or polar, charged or uncharged side chain. For some catalytic reactions the chemical properties of these amino acids may be sufficient, but for the majority of reactions such as redox reactions or chemical group transfers, enzymes require the assistance of additional molecules that bind on the third part of an active site. These molecules are defined as either cofactors, which are tightly bound to the enzyme throughout the catalytic reaction or coenzymes, which are released during the reaction. Cofactors distinguish themselves from coenzymes by being not consumed in the catalytic reaction.

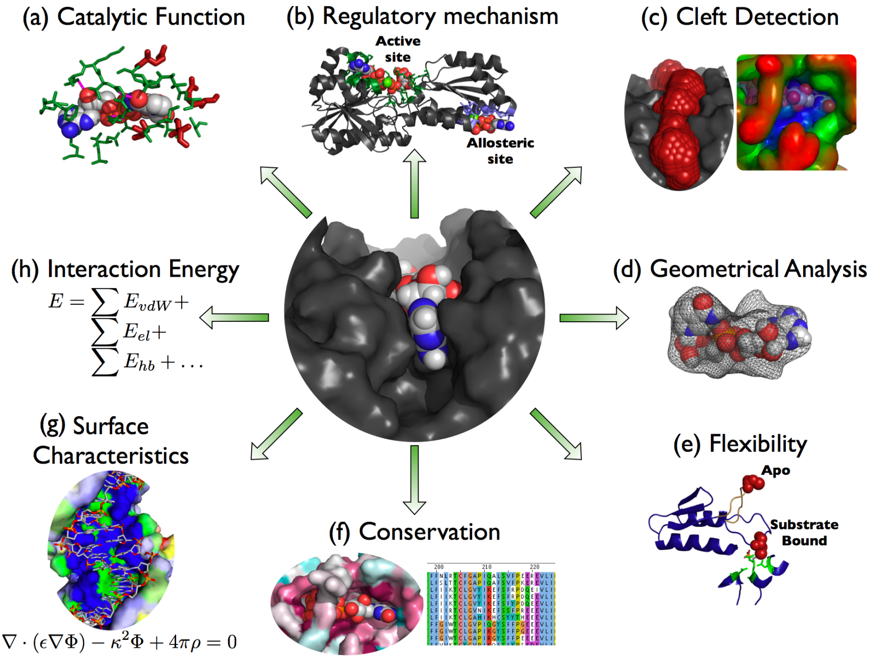


Fig. 8.2 Characteristics of enzyme binding sites. (**a**) The active site is a specific binding site in an enzyme that contains catalytic residues to perform the enzymatic reaction on a substrate. (**b**) The activity of an enzyme can be regulated, for example, by allosteric regulator molecules that bind to a remote binding site. (**c**) In most enzymes the active site is found in the largest or deepest cleft of the enzyme, (**d**) and encloses at least partially the ligand with amino acids, resulting in similar geometrical shapes for the binding site and ligand. (**e**) Binding sites can undergo major conformational changes upon subtrate binding, especially when some parts of the site are located in flexible loops. (**f**) As binding sites are essential for the function of a protein, their residues are often amongst the most highly conserved residues. (**g**) The binding affinity of a ligand is influenced by the physicochemical properties on the binding site surface like complementary electrostatic potentials or perturbed pK_a values (**h**), which can be exploited to calculate estimated binding energies between ligand and binding sites.

Though they get altered while the catalysis takes place, they are recovered again in the same catalytic process. In contrast, coenzymes support the enzyme reaction by providing chemical groups to the substrate, and subsequently, detaching from the enzyme to start a recovery process outside the enzyme. Typical cofactors are the inorganic metals and sulphate ions or the organic flavin and heme groups. Examples of coenzymes are vitamins or the cellular energy carrier, ATP.

Some enzymes, especially ones assembled by several domains or several chains, can have allosteric sites in addition to the substrate and cofactor/coenzyme binding sites [see Fig. 8.2(b)]. These allosteric sites play an important role in the regulation of enzymes as they induce, upon binding a regulator molecule, conformational changes on the whole enzyme structure, which can affect also the atomic constellation of the active sites. Depending on whether the regulator molecule is an effector or an inhibitor, the changes on the active site can either enhance or hamper the enzymatic catalysis.

The underlying principles of allosteric regulation, as well as the atomic interactions of any binding process between an enzyme and a molecule, have only been elucidated since high-resolution data of the three-dimensional (3D) coordinates of enzyme-molecule complexes were determined. Two main approaches are used for the determination of such high-resolution data for biomolecules, namely "X-ray crystallography" and "Nuclear Magnetic Resonance (NMR) spectroscopy" (see Chapters 22 and 24 for an in-depth description of these methods). The first enzyme structure discovered in 1965 was the X-ray structure of lysozyme, an enzyme found in tears or egg white that digest bacterial cell walls. Since then, many enzyme structures have been determined and their functions analyzed, and the resulting information has been stored in databases. See Table 8.1 for the number of enzymes in some structure-based databases.

The most important among them is the Protein Data Bank (PDB)[1] (http://www.pdb.org) and the Enzyme Commission (EC) number for enzyme reaction.[2] The first is important as it is the worldwide depository for 3D coordinates of enzymes and any other macromolecules like other proteins, nucleic acids, or carbohydrates (see Chapter 26 for further information on the PDB). Structures in the PDB are assigned a unique four alphanumeric PDB Identifier (Id). The importance of the EC number is that it provides a classification scheme for all enzyme reactions and allows their comparison.

Table 8.1 The Extent of Enzyme Data in Some Structural Databases as on 21 July 2007

Number of	Quantity
Known enzyme reactions (unique EC numbers)	~4040
Enzymes in UniProt/Swiss-Prot (56)	~107 400
Enzymes in PDB	~19 600
EC Reactions in PDB	~1390
Enzymes with catalytic residues in CSA	880
Enzymes with catalytic mechanisms in MACiE (57)	202
Enzymes as specified by EC number in PDB with the largest number of structures	
1. Lysozyme, EC 3.2.1.17	~930
2. Non-specific serine/threonine protein kinases, EC 2.7.1.37	~580
3. Trypsin, EC 3.4.21.4	~430
Most enzymes in PDB originate from	
1. Human	~10 700
2. *Escherichia coli*	~4200
3. House mouse	~2100
4. Cow	~1550
5. Baker's yeast	~1300
No of organisms that have one or more enzyme structures in PDB	~1128

The EC number consists of four digits separated by full stops. The first number (class) indicates the reaction type, the second number (sub-class) together with the third number (sub-subclass) represents the occurring chemistry, and the last number gives the substrate specificity.

From the three-dimensional structures of enzymes, it became evident that substrates and secondary molecules like cofactors and coenzymes do not bind randomly on the enzyme surface. The same molecule always binds at the same site within the same enzyme structure. This has led to the assumption that binding sites must have unique features that distinguish them from other areas on the enzyme surface, and in addition, allow the binding site to recognize its associated molecule from the thousands that exist in a living cell. Two models were suggested to explain in particular the specificity of active sites. First, the Lock and Key model by Fischer,[3] and second, the Induced Fit model by Koshland.[4] The Lock and Key model assumed that a ligand is geometrically complementary to its active site and that both shapes fit exactly into one another. The more recent model of Induced Fit was a modification to the Lock and Key model and incorporated the flexibility of enzymes and substrates. The model suggests an "open" state for an enzyme when the substrate binds, followed by a "closed" state where the enzyme encloses the bound substrate and performs its catalysis. In the process of converting from the open state to the closed state, the active site adjusts its shape to the transition state that is the conformation of the ligand at the highest reaction energy (see Chapter 10), and allows the catalytic reaction to take place.

This chapter addresses different aspects or features of protein binding sites (see Fig. 8.2). It will give some background information to each feature and describe one exemplary methodology to calculate it. A more comprehensive list of computational methods can be found at the end of the next section. All tools and programs introduced in this chapter are not just important to visualize the features in an enzyme but also to try to predict the function for an enzyme. The latter becomes more and more important as more and more enzyme structures are deposited in the PDB without any functional annotation. Many of these structures were targets of global structural genomics initiatives, which aim to develop high-throughput methods for the rapid determination of protein structures. One goal of these initiatives is to determine the structures of all existing protein folds in nature.[5] The high-throughput principle is advantageous for determining many structures in a short time but does not address the functional annotation of proteins, which usually involves many different wet lab experiments and thus is a time-consuming procedure. In order to obtain hints about the function of these unannotated structures, one can extract the features described in this chapter and search for similar features in annotated enzymes. For this purpose, the third part of this chapter will be devoted to algorithms for the comparison of binding site features.

Before we start with the binding site characteristics we would like to note that in this chapter we will refer to any small molecule that is bound by an enzyme as a ligand whether it is a substrate, product, or allosteric effector.

8.2 Enzyme Binding Sites and Their Unique Features

8.2.1 *Active Sites are in Largest Cleft*

Enzyme active sites tend to be within sizeable depressions on the protein's surface, which are known as clefts or pockets. In 70–85% of enzymes, the largest of these clefts is where the substrate and relevant cofactors or coenzymes bind.[6,7] The average volume of a binding site depends on the ligand it binds, and ranges mostly from 400 to 2000 Å^3.[8]

SURFNET[9] is an elegant approach to identify and visualize clefts in proteins. It detects gap regions within the protein by fitting spheres of certain range of sizes between protein atoms. The spheres are not allowed to clash with any neighboring protein atoms. Overlapping SURFNET spheres are clustered and regarded as protein clefts [see Fig. 8.3 and Fig. 8.2(c)]. Placing a grid on the cleft and determining the number of grid cells occupied by a sphere enables the calculation of the volume for each cleft.

8.2.2 *Active Sites are in Deepest Cleft*

The enclosure of a ligand within large and deep clefts helps the enzyme to maximize the number of interactions with its ligand.[10] In particular, active sites are often found in the deepest cleft of an enzyme. The average depth of a cleft that contains a binding site depends on the protein size and can be up to 30 Å.[11]

The algorithm of travel depth[11] is an elegant way to visualize and measure the depth of clefts relative to the convex hull of the enzyme's molecular surface [see Fig. 8.2(c)]. The convex hull is defined for a simplified two-dimensional molecule as the region that is enclosed by

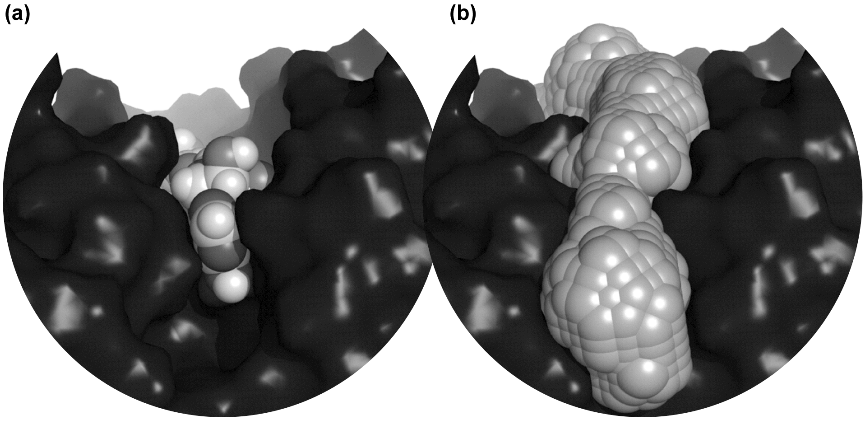


Fig. 8.3 **(a)** Spherical section of the protein structure of ribosyl-transferase (PDB Id: 1og3) colored in black, with bound coenzyme NAD in the active site. **(b)** Largest cleft, as determined by SURFNET, contains the active site. SURFNET spheres are represented by light grey spheres.

a rubber band that is stretched around the whole molecule. The travel depth algorithm finds for a probe sphere on the protein surface the minimum distance to reach the convex hull. It works by placing the protein into a grid and assigning to all grid cells outside the convex hull a depth of zero. For grid cells inside the convex hull, the algorithm scans recursively through the grid and adds to the size of each grid cell the minimum depth of its neighboring cells.

8.2.3 *Binding Site Shapes are Complementary to Ligand Shapes*

It is a common assumption that the shapes of protein binding pockets are complementary to the shapes of the ligands they bind. This assumption became manifest in the Lock and Key model and Induced Fit model for molecular binding (see Section 8.1). A recent study however showed that exact shape complementarity between a binding site and its bound ligand is rarely achieved, and that more often, some free space can be found between the binding site and its ligand[8] (see Figs. 8.2(d) and 8.4).

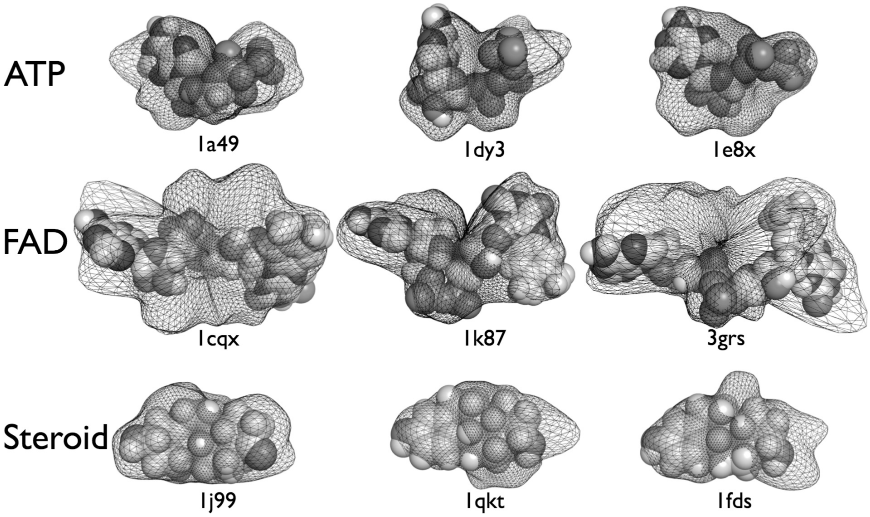


Fig. 8.4 Binding site shapes are not truly complementary to the ligand shape and often show some empty space between the ligand and the binding site like a "buffer zone." The PDB identifier of each associated protein structure is given below each binding site.

For the analysis and visualization of binding sites and ligand shapes, one can apply an elegant approach, which utilizes mathematical functions called spherical harmonics. The computational description of shapes can be simplified by a radius function, which returns for any point on the shape surface its distance to the center of the shape. The common way of obtaining the function is by selecting a number of points on the shape surface and exploiting their radii to approximate the radius function. The approximation can be done by summing up spherical harmonic functions in an equivalent way to the Fourier series, where sine or cosine functions are summed up to obtain a periodic one-dimensional function. While the summation progresses, each spherical harmonic function contributes, with a different weight, to the radius function. The contribution weights are usually referred to as coefficients. Once the approximation finishes, a vector of all coefficients is retrieved and used to reconstruct the shape of the binding site (see Fig. 8.5).

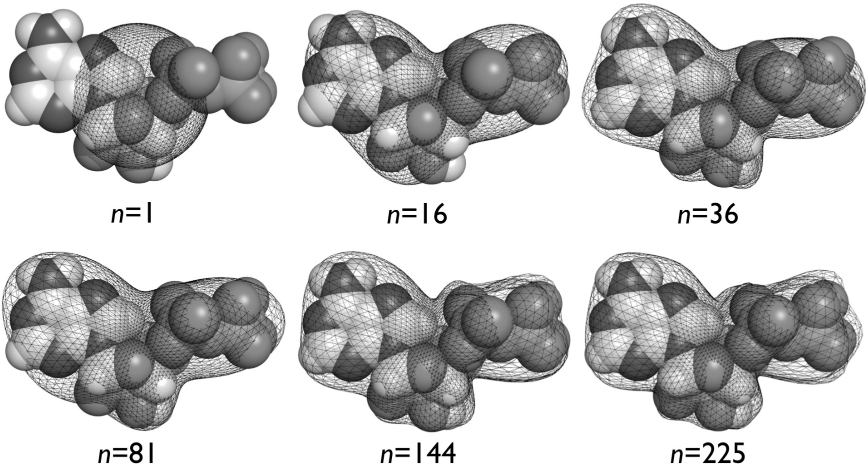


Fig. 8.5 Various approximations of the molecular surface of an ATP (PDB ID: 1e8x) with increasing number *n* of spherical harmonic functions.

8.2.4 *Binding of the Ligand Induces Conformational Changes in the Binding Site*

The Induced Fit model for molecular binding states that enzymes undergo conformational changes upon substrate binding. For a small fraction of enzymes these changes are large, in particular, if they include a flexible loop region that closes/opens the entrance to the active site, preventing/allowing the binding of a ligand [see Fig. 8.2(e)]. However, for the majority of enzymes, the changes are small. The average RMSD (see below) upon ligand binding between C_a atoms of binding sites and catalytic residues is less than 1 Å.[12] Similar values are observed for the side chain atoms. It is interesting to note that residues in active sites are on average more flexible than other residues in the protein structure. This can be traced back to the geometrical adjustments of the active site to generate the transition state of the ligand (see Section 8.1). But there are also enzymes, like prothymosin-α, that are intrinsically disordered in their native state.[13] Neither the Lock and Key nor the Induced Fit model can describe their functionality. A third model, the "New View" model, has recently been introduced, and it states that a protein exists in an ensemble of pre-existing conformations with discrete and similar free energies. Among them is also the structure of the bound conformation. The actual binding of the ligand induces a shift in the equilibrium of existing conformations towards the bound conformation and causes the protein to appear well structured in an X-ray crystal.[14]

The standard method for measuring the flexibility of enzymes binding sites is to calculate the Root Mean Square Deviation (RMSD) between different conformations of the binding site. The RMSD is calculated between the Cartesian coordinates of all atom pairs between both proteins using the following formula:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=0}^{N}\left[(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2\right]}{N}} \tag{8.1}$$

where *x, y, z* are the Cartesian coordinates of the protein atoms and *N* is the number of compared atoms. Depending on the scientific question or on the available data, one can calculate the RMSD of all atoms, of all residue side chain atoms, or of only the backbone/C_a atoms between two structures. For the qualitative analysis of flexibility, one can use the web server of STRuster.[15] STRuster analyzes an ensemble of different conformations of a protein by first calculating the Euclidean distances between all residues within a conformation, and next, comparing the distances to the distances in the second conformation. The compared distances are summed up and plotted in an "all-conformation versus all-conformation" distance matrix. The distance matrix is utilized to cluster each conformation according to its level of flexibility and group similar conformations.

8.2.5 *Binding Site Residues are Highly Conserved*

Another characteristic of enzyme binding sites is that the residues forming the sites tend to be strongly conserved within the protein family. That is, all members of a protein family tend to have the same residues in the same position in both their sequences and their 3D structures. The reason for this is that they all have evolved from a common ancestor and have the same function but are found in different organisms. Each family member is however subject to natural variation and selection with mutation and duplication events throughout their protein sequences. However, mutations are not tolerated at all positions in the protein sequence. While those residues that had no functional role in the protein could mutate freely, substitutions of functionally important residues (i.e. residues that are involved in ligand binding or in keeping the structural integrity) are restricted, as these mutations could have led to the loss of protein function. Residues found in binding sites and especially catalytic residues in active sites are amongst the most important residues in an enzyme structure, and consequently, particularly highly conserved. Most often, these residues are either polar or charged (up to 70% of residues are Arg, Asp, Cys, Glu, His, and Lys).[16]

ConSurf[17] calculates the conservation of each amino acid in a protein sequence using the evolutionary trace method.[18] This method first runs a multiple sequence alignment on a set of homologous sequences, i.e. sequences that have a common ancestor. In the second step, the method uses the alignment to compute a phylogenetic tree, which represents the evolutionary relationship of the homologous sequences. In the third step, the homologous sequences are divided into groups and subgroups based on the branches of the tree. In the fourth step, the residue positions in all sequences in each group and subgroups are analyzed for the frequency of residue changes. If at a particular subgroup a residue is invariant throughout all sequences in the subgroup, it becomes assigned a rank, which states how many times the tree was required to be divided to yield the ranked residue. The same procedure is applied to all residues until every residue gets assigned an evolutionary rank. According to the ranks, ConSurf groups the residues of the query sequence into nine classes, with “1” being the least conserved and “9” being the most conserved residues, and the conservation scores are mapped onto the protein structure [see Fig. 8.6 and Fig. 8.2(f)]. A visual inspection of the protein structure can identify clusters of highly conserved residues on the protein surface.

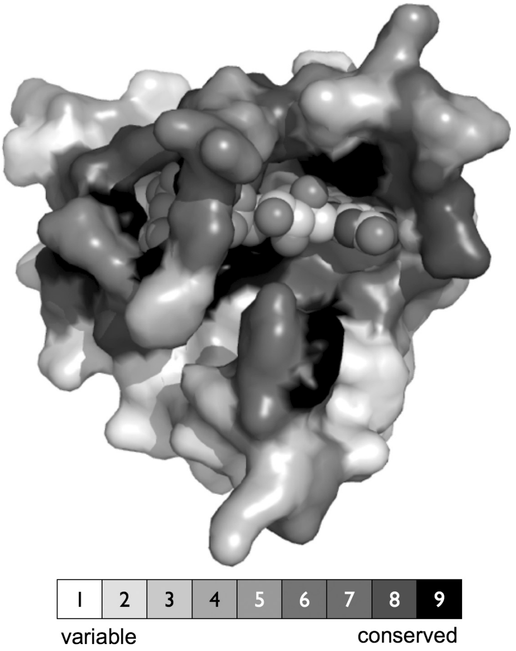


Fig. 8.6 ConSurf conservation scores mapped on PDB structure 1p4m. Note the higher conservation in and around the binding sites.

8.2.6 *Complementary Electrostatic Potentials Between Binding Sites and Ligands*

Electrostatic potentials are long-range potential energies and one of nature's strongest forces at the atomic scale. All energies between atoms and molecules are electrostatic in origin, whether they are transient dipole-dipole interactions as in the case of van der Waals interactions, charge-charge interactions, or hydrogen bond interactions. They differ in the rate of decreasing interaction energy with increasing atomic distance.[19]

One theory about electrostatic complementarity between binding sites and ligands suggests that electrostatic potentials are strong enough to attract the ligand from the solvent into the active site. This assumption has been derived from enzymes that have catalysis rates approaching the diffusion limit, like the copper-zinc-superoxide-dismutase

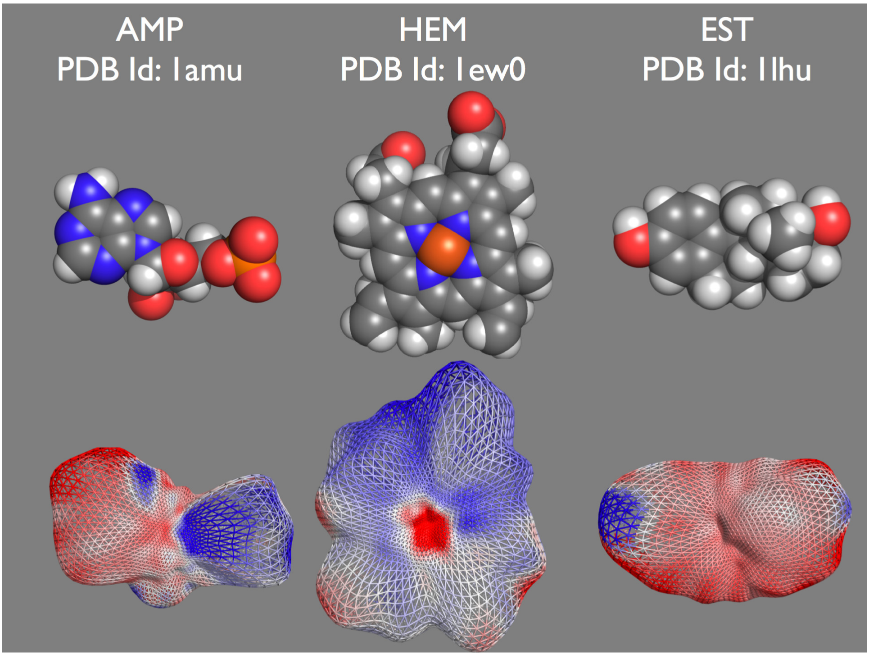


Fig. 8.7 Electrostatic potential of three proteins mapped on the molecular surface of their ligands as represented by spherical harmonics (see Section 8.2.3): AMP, heme, and Estradiol. Negative potentials are colored red, neutral potentials are colored white, and positive potentials are colored blue.

protein family. This protein family exerts a positive electric field over the active site, which attracts negatively charged oxygen radicals towards the active site copper ion.[20] The visualization of the electrostatic potentials mapped on the structure surface, also referred to as potential surfaces, is particular useful for identifying DNA binding sites. Many DNA binding proteins possess a large patch of positively charged amino acids on their surface to electrostatically attract their negatively charged binding partner[21] [see Fig. 8.2(g)]. Figure 8.7 visualizes the electrostatic potentials by showing the potentials of three proteins on the molecular surface of their ligands.

The eF-site[22] database contains pre-calculated potential surfaces for all PDB structures. Auxiliary servers to the eF-site database allow the calculation of the electrostatic potential for any user-provided structure and the search for similar surface potentials in the eF-site database. The electrostatic potentials in eF-site are calculated by a standard procedure applied by many electrostatic methodologies, among them APBS and Delphi (see Table 8.2).

The methodology simplifies the representation of the protein and the solvent by ignoring the molecular details of the solvent molecules and treating all solvent molecules as a single continuum. The simplification is necessary as the explicit calculation of all interactions between water molecules to each other and to the protein is computational demanding and most often not feasible. In combination with the simplification, the electrostatic potential of a protein is calculated by solving the Linear Poisson-Boltzmann differential Equation (LPBE).[23] As every protein has an arbitrary shape, the LPBE is solved numerically by discretizing the space occupied by the protein with a grid and calculating iteratively the electrostatic potential for each grid cell using the finite difference technique.[24]

8.2.7 *Catalytic Residues Destabilize the Enzyme Structure and Have Perturbed pK$_a$-values*

The ability to calculate the electrostatic potential for a protein structure facilitated the computational analysis of two further phenomena in active sites. Both phenomena are unique properties of ionizable catalytic residues (all Lys, Arg, Asp, Glu, His, Tyr, Cys, N-terminus, C-terminus) and distinguish them from the remaining residues in the enzyme structure. One of these properties is their capacity to destabilize the integrity of enzyme structures, especially when they are found in active sites that exert repulsive electrostatic forces towards the ionizable catalytic residues. Experiments have shown that the replacement of the affected residues with neutral or oppositely charged residues tended to stabilize the protein structure.[25]

Another of these properties is a perturbed pK$_a$-value for ionizable catalytic residues. The pK$_a$ is defined as the pH for which the average protonation state of an ionizable molecular group is 0.5. It can be measured by titration curves that plot the solvent's pH against the net charge of the ionizable group. For non-catalytic residues, in general, these

Table 8.2 Programs and Web Servers to Analyze Different Aspects of Enzyme Binding Sites

Method	Program/Server	URL	Notes
Size	SURFNET	http://www.biochem.ucl.ac.uk/~roman/surfnet/surfnet.html	Active sites are most likely in the largest protein cleft.
	CASTp	http://sts.bioengr.uic.edu/castp/	
	VOIDOO	http://xray.bmc.uu.se/usf/voidoo.html	
Depth	TravelDepth	http://crystal.med.upenn.edu/travel_depth.tar.gz	Binding sites are often found in deep protein clefts.
	PocketPicker	http://gecco.org.chemie.uni-frankfurt.de/pocketpicker/index.html	
Flexibility	STRuster	http://struster.bioinf.mpi-inf.mpg.de/	Protein structures can undergo conformational changes upon ligand binding.
	MolMovDB	http://www.molmovdb.org/	
Conservation	ConSurf	http://consurf.tau.ac.il/index.html	Binding sites are among the most conserved regions on the protein.
	Evolutionary Trace	http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html	
	JevTrace	http://www.cmpharm.ucsf.edu/~marcinj/JEvTrace/	
3D templates	CSA	http://www.ebi.ac.uk/thornton-srv/databases/CSA/	Catalytic residues are often found to be highly conserved in their spatial disposition.
	PINTS	http://www.russell.embl-heidelberg.de/pints/	
	Rigor	http://xray.bmc.uu.se/usf/rigor_man.html	

(*Continued*)

Table 8.2 (*Continued*)

Method	Program/Server	URL	Notes
Electrostatic potential	APBS	http://apbs.sourceforge.net/	DNA binding proteins have often large, positively charged binding sites.
	DELPHI	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:DelPhi	
	PCE-Pot	http://bioserv.rpbs.jussieu.fr/cgi-bin/PCE-Pot	
	eF-site	http://ef-site.hgc.jp/eF-site/	
pK_a-values	PROPKA	http://propka.ki.ku.dk/	Catalytic residues have often perturbed titration curves.
	WHAT IF pKa	http://enzyme.ucd.ie/Science/pKa/Software	
	PCE-pKa	http://bioserv.rpbs.jussieu.fr/cgi-bin/PCE-pKa	
Hydrophobicity	GRASP	http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:GRASP	Hydrophobic binding sites often bind hydrophobic ligands.
Hydrogen bond	HBPLUS	http://www.biochem.ucl.ac.uk/bsm/hbplus/home.html	Hydrogen bonds provide specificity for ligand binding.
	LIGPLOT	http://www.biochem.ucl.ac.uk/bsm/ligplot/ligplot.html	

(*Continued*)

Table 8.2 (*Continued*)

Method	Program/Server	URL	Notes
Potential function	Q-SiteFinder	http://www.bioinformatics.leeds.ac.uk/qsitefinder/	Binding sites can often develop high interaction energies that can be assessed by potential functions.
	Grid	http://www.moldiscovery.com/soft_grid.php	
	MCSS	http://www.accelrys.com/insight/mcss.html	
Biological Unit	PQS	http://pqs.ebi.ac.uk/pqs-quick.html	PDB structures often represent not the biological active conformation of the protein.
	Pita	http://www.ebi.ac.uk/thornton-srv/databases/pita/	
	PISA	http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html	
Cognate Ligand	PROCOGNATE	http://www.ebi.ac.uk/thornton-srv/databases/procognate/	Not all bound ligands to a protein structure are functionally related.
Enzyme mechanism	MACiE	http://www.ebi.ac.uk/thornton-srv/databases/MACiE/	Enzyme reactions consist of various catalytic steps.

curves adopt a specific shape in which the net charge decreases with increasing pH and with a sharp decline around the pK_a-value. For catalytic residues, however, these curves can be perturbed, generating regions of constant protonation state or shifts in the pK_a-value.[26]

Theoretical microscopic titration curves (THEMATICS)[26] can be computed for every ionizable residue in a protein using electrostatic potential calculations. The superposition of titration curves obtained for all residues of the same type within the protein identifies perturbed curves and may indicate ionizable catalytic residues.

8.2.8 *Hydrophobic Interactions are Essential for Binding*

In a study where organic solvent molecules were computationally mapped on the protein surface to predict potential binding sites of ligands, it was found that hydrophobic patches are also important within binding sites, inducing organic solvents to cluster therein.[27] The results are in agreement with earlier experiments that showed that binding affinities of ligands can increase by promoting hydrophobic interactions between binding sites and ligands.[28] Our own calculations confirmed that hydrophobic ligands like heme and steroids are often bound by binding sites that expose mainly hydrophobic residues.

Computationally, the hydrophobicity of amino acids can be calculated by exploiting the fact that hydrophobic amino acids are usually surrounded by other amino acids in the protein's core and not accessible to solvent molecules. Calculating the mean fractional area loss upon protein folding of a residue provides an estimate on the residue's hydrophobicity. The area loss is obtained by relating the solvent accessible surface area (SASA) of an amino acid in a fully extended conformation to the mean SASA of the amino acid in the protein structure. The SASA can be calculated by rolling a probe sphere over the atomic van der Walls surfaces and placing a fixed number of dots per unit area on the roll track of the probe sphere. The number of dots multiplied by the area that a dot occupies gives the accessible surface area. The ASA of the extended conformation is usually given as the surface area of the residue within the extended tripeptide Gly-X-Gly.[29]

8.2.9 *Hydrogen Bonds Provide Binding Specificity*

Unlike other chemical interactions, hydrogen bonds require directionality between the hydrogen-bond acceptor and donor. This directionality provides the enzyme's specificity for its ligand. Only ligand atoms that have a specific orientation towards a particular binding site can form hydrogen bonds. Ligands that do not have the right atoms at the right place cannot form hydrogen bonds and must rely on other forms of interaction to achieve binding.[30] Most hydrogen bonds in binding sites are formed among the atoms of the binding site in order to stabilize the positions of the catalytic residues. Only a small portion (10–20%) are formed with ligand atoms.[16] In protein-ligand complexes there are on average 10 bonds, of which two-thirds are hydrogen bond acceptors and a third hydrogen bond donors.[31]

The program LIGPLOT[32] uses the application HBPLUS[33] to extract and plot hydrogen bonds between the binding site and ligand atoms. The algorithm of HBPLUS begins with placing hydrogen atoms in the protein structure. This is necessary, as most X-ray crystal structures do not include hydrogen atoms except for NMR or very high-resolution X-ray structures. Once the hydrogen atoms are generated, the hydrogen bonds are determined by applying purely geometrical criteria[32] to the protein-ligand complex. In addition to hydrogen bonds, HBPLUS also calculates non-covalent bond interactions by applying a simple cut-off of 3.9 Å to atomic distances between the binding site and ligand. Finally, LIGPLOT draws a schematic two-dimensional diagram of the binding site ligand complex and highlights the calculated hydrogen bonds and non-covalent bond interactions (see Fig. 8.8).

8.2.10 *Potential Functions for Estimating Binding Energy*

The process of molecular binding requires in the first instance shape complementarity to allow ligand atoms to approach the binding site atoms. The proximity between both binding partners is important as

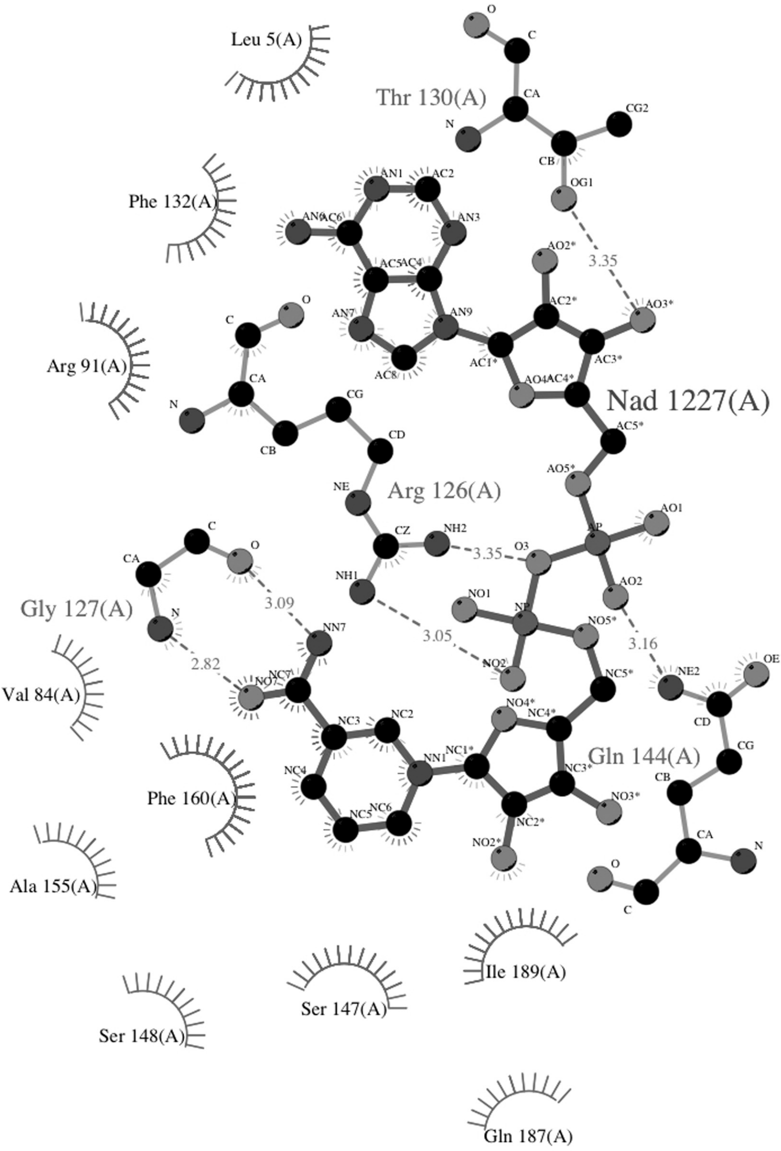


Fig. 8.8 Schematic diagram of the non-covalent interactions between NAD and its binding site in PDB structure 1p4m. Thick lines belong to the ligand and thin lines to the hydrogen-bonded residues in the binding site. Hydrogen bonds are indicated with dashed lines. Non-covalent bond interactions are shown as spoked arcs pointing towards the ligand.

their binding energy depends very much on the distances between their atoms. Since ligand molecules do not bind at random sites on a protein structure, their binding sites should feature particular high binding energies towards the ligand.

Q-SiteFinder[34] calculates the potential binding energies on a protein surface and detects energetically favorable surface patches that may present ligand binding sites. The favorable patches are found by placing the protein in a grid and rolling a probe sphere along the grid points over the molecular surface. At each grid point an energy function, which incorporates van der Waals potential, electrostatic potential, and hydrogen bond potential, is applied to the probe sphere [see Equation in Fig. 8.2(h)]. Grid points that exceed a predefined energy threshold are clustered if they are below a certain separation. For each cluster, the single interaction energies of the grid points are summed up and ranked according to their total interaction energy. The cluster with the most favorable interaction energy is identified and is considered as a potential binding site.

8.2.11 *Unusual Amino Acids*

There are 20 standard amino acids used by nature to build up proteins, however, under certain circumstances some amino acids can be catalytically altered, giving rise to a 21st amino acid. One such change occurs in active sides of copper amine oxidases (PDB Id: 1pu4), which increases the catalytic activity of the enzyme. The change occurs at the catalytic active tyrosine, which becomes autocatalytically oxidized to tri-hydoxy-phenyalalanine (Topa) in the presence of a copper ion.[35] Another example is the phosphomannose isomerase, which when expressed in *E. coli* has a di-hydoxy-phenylalanine (Dopa) substituting for a tyrosine.[36]

8.2.12 *Precautions with PDB Structures*

Structures deposited as single chains in the PDB are often actually dimers or tetramers or sometimes vice versa. When analyzing binding

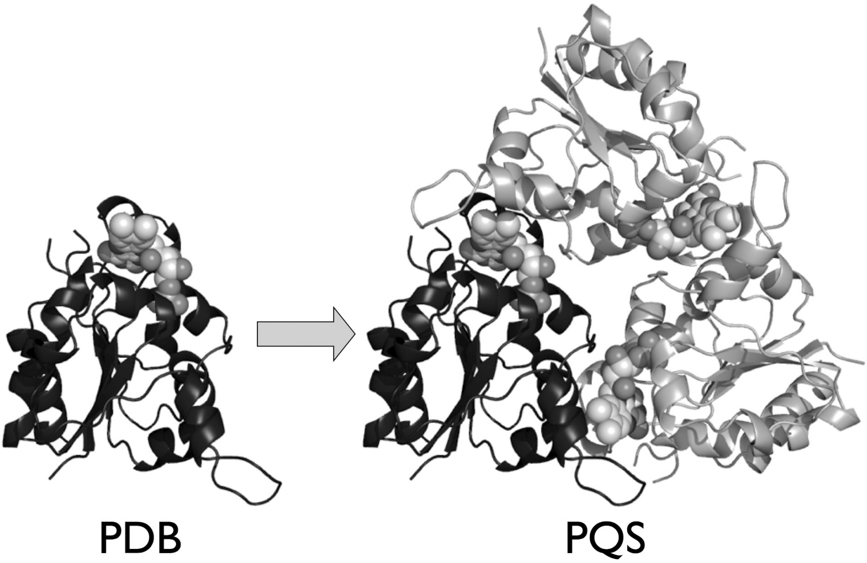


Fig. 8.9 The PDB structure of the decarboxylase 1mvl shows only a monomer with the FMN being exposed to the solvent. However, the biological relevant conformation is a trimer as calculated by PQS, with a FMN binding site shared between two subunits.

sites, one has to bear in mind this obstacle, especially as many binding sites in dimers are found at the interface of the two monomers (see Fig. 8.9). The PQS (Protein Quaternary Structure)[37] file server is a depository of estimated quaternary structures of all PDB structures. We would encourage the reader to use in any of their protein studies these assembles for their proteins from the PQS database, since although not perfect, they are much more reliable than using the single chain.

Ligands that are found attached to an enzyme in a crystal structure may not always be the native substrate or cofactor, etc., of an enzyme. Many such ligands found in the active site are substrate analogues or enzyme inhibitors that compete with the substrate for binding into the active site. In addition, some ligands can be artifacts of the crystallization buffer, which is a mixture of different solvents to promote the crystallization process of a protein. In general, all ligands that are not required for the enzyme function are called non-cognate, whereas ligands that are functionally related to an enzyme are designated as cognate. The PROCOGNATE[38] database has been established to address this problem and contains information about cognate ligands in enzymes and provides similarity scores for non-cognate ligands that allow their structural comparison to the cognate ones.

8.3 Methods and Tools for Comparing Enyzme Binding Sites

Tools to assess the similarity between binding sites compare either atomic coordinates or surface properties. In the field of computer vision, many methods exists for comparing three-dimensional coordinates, features, or surfaces. See Ref. 39 for a review on the existing methods. However, only a few of them have been realized in structural biology. Among these, the most important ones are the kd-tree search, graph matching, geometrical hashing, and coefficient comparison of spherical harmonic functions. A detailed description of each follows.

8.3.1 *Comparing Catalytic Templates*

As mentioned in the introduction, two to six catalytic residues within an active site perform the catalytic reaction of an enzyme (see red colored residues in Fig. 8.1). Usually, the spatial conformation of these residues is highly conserved for the same enzymatic reaction and can be recovered in evolutionary unrelated enzymes, as in the case of serine proteases and their Ser-His-Asp catalytic triad. The Catalytic Site Atlas (CSA)[40] stores a catalogue of catalytic residues as templates and provides the motif finder JESS[41] to search for the existence of the templates in a query protein structure. The JESS algorithm works by extracting constraint conditions from the template, which include the type of residues that are allowed to participate in a catalytic site and the allowed separations between these residues. The aim of JESS is to find residues in the protein structure that fulfill these constraints.

8.3.2 *Comparing Atomic Coordinates*

According to graph theory, a binding site can be regarded as a graph, with atoms being the nodes and the distance vectors between the atoms being the edges. A new association graph can be inferred using all atoms in both binding sites that are similar with respect to their physicochemical property and spatial location. Given such an association graph, the task is to find the maximum clique, i.e. the largest subset of nodes that are all connected with each other in a pair-wise manner. This problem is computationally demanding since every additional node increases the computation time with N^2. To reduce the complexity, the program IsoCleft[42] uses exclusively C-alpha atoms as a pre-filtering step, and only in a second stage, runs a more demanding all atom comparison. Another method, CavBase[43] (http://www.ccdc.cam.ac.uk/products/life_sciences/relibase/) uses a small set of pseudospheres to represent the location and physicochemical property of a binding site residue. In any case, once the maximum clique is found, the similarity between two binding site is assessed by relating the size of the maximum clique to the smaller binding sites.

The second approach to structural atomic comparisons is geometric hashing, which consists firstly of a preliminary preprocessing stage that runs offline only once and is followed by a recognition stage. In the preliminary stage, a database is created with a hash table for each binding site following four steps:

1. Three atoms being non-collinear to each other are picked out from a binding site. The triplet represents a plane in space from which an orthonormal reference frame can be built. The reference frame will help to describe the geometrical positions of the remaining binding site atoms independent from their original Cartesian coordinates.
2. Each remaining atom in the binding site is located within the triplet reference frame.
3. Representative information on the triangle together with location information of the fourth atom (quadruplet) is stored in a hash. If required, any other properties of the atoms can be added.
4. Repeat steps 1–3 for all other triplet combinations in the binding site.

Once the database of hash tables is built, the recognition stage can begin by applying the same approach as above to a query binding site. However, instead of storing the quadruplets in a hash table, they are checked for their existence in the database. Hash tables that exceed a user-defined minimum match value are considered as similar and are further analyzed for atom clusters.

8.3.3 *Comparing Binding Surfaces*

The molecular surface is crucial in intermolecular interactions as it is the interface through which the molecules interact. Different surface models exist for molecules with the two most important ones being the molecular surface and the solvent accessible surface. Both surfaces can be obtained by rolling a probe sphere over the van der Waals atom shells of a molecule. Whilst the inward-facing surface of the probe sphere produces the molecular surface, the solvent accessible surface is built by tracing the centre of the probe sphere. The radius of the probe sphere influences the appearance of both surfaces. A smaller probe sphere will show the surfaces in greater detail, whereas a larger probe sphere will reveal only major surface characteristics. Usually, the radius of a water molecule with 1.4 Å is used as the probe sphere radius.

Different representations exist for molecular surface models. The piecewise-quartic representation splits up the molecular surface into concave spherical triangles, saddle shape rectangles, and convex spherical regions. The Connolly dot representation spreads over the surface dots that allow a transparent view of the molecular surface. Another transparent representation is gained by tessellating the surface into linked empty triangles.

Although the visualization of molecular surfaces is well established, their comparison is just the opposite. Only a few attempts have been made to compare molecular surfaces. Their methodology is based mainly on the comparison techniques mentioned above, in which points on the molecular surface are compared using geometric hashing[44] or, as in the case of the publicly accessible eF-seek webserver (http://ef-site.hgc.jp/eF-seek/), using graph matching.

An elegant approach for surface comparison is to compare the coefficient vectors of binding site shapes that are approximated with spherical harmonics functions (see Section 8.2.3). The comparison between two shapes reduces to a Euclidean distance calculation between two coefficient vectors, with smaller distances for similar shapes.[8] Note that this approach is not fully comparable to the graph matching and geometric hashing methods mentioned above, as it compares the shape and not the molecular surface of the binding site. The volumetric shape represents not directly the molecular surface but the negative imprint of a binding site that is occupied by the ligand in the binding process.

8.3.4 *Other Comparison Methodologies*

Instead of describing the binding site properties specifically, FFF[45] (Fuzzy Functional Forms) explore to what extend properties can be relaxed and still allow a recognition of a binding site in a database scan.

pvSoar[46] (http://pvsoar.bioengr.uic.edu/) compares local sequence and geometric similarities of binding sites. It extracts the residues building up the wall of clefts from the CASTp[47] database and runs a sequence alignment to detect any highly conserved sequence patterns. In a second step, the geometric positions of the conserved residues are compared using a simple RMSD (see Equation 8.1) calculation.

8.4 Future Outlook

Even with the wide variety of identified binding site features and the methods described above, it remains difficult to correctly predict potential interactions between proteins and ligands. Drug discovery programs report some promising results on the prediction of interaction energies with *in silico* docking programs. However, in general, the prediction successes of docking and mapping applications remain rather moderate. The reasons are mainly the oversimplification of the physical conditions in the interaction process as well as persisting problems in recognizing the fundamental processes of molecular recognition[48,49] (see Chapter 17 for an in-depth discussion on *in silico* docking).

The current concept of molecular recognition states that molecular binding occurs primarily due to complementary physicochemical properties between a binding site and its ligand. This hypothesis might require amendments, as more and more examples arise that show binding despite non-complementarity. The most striking examples occur in phosphate receptors (PDB Id: 1pbp), sulphate binding proteins (PDB Id: 1sbp), flavodoxin structures (PDB Id: 2fox), and DNase I structures (PDB Id: 2dnj).[50] All binding sites in these structures exert a negative electrostatic field over their binding sites despite binding a highly negative substrate. The question remains open whether in their evolutionary past these enzymes were binding ligands with complementary electrostatic potentials. Enzyme promiscuity might play a decisive role to answer this question. The current view on proteins, which is mainly governed by their specificity towards their functionality, is likely to change towards functional promiscuity, which states that a protein can exert different functions with the same active site. More and more enzymes are discovered that, despite their specificity, promiscuously catalyze other sometimes very different and unrelated reactions.[51] The increasing amount of data coming from growing 3D structure databases, annotations of catalytic mechanisms and in-depth binding site analyses will provide useful information to reveal the fundamental process in molecular recognition.

Structural information about enzymes have been derived mainly from X-ray crystal structures, which provide atom coordinates of unparallel high resolution. X-ray crystallography has one major drawback, which is that it provides only a static picture of an otherwise flexible protein. Protein dynamics and motions in crystals are usually only visible as a lack of "clarity" caused by the averaging process over many molecules. Molecular dynamics simulations attempted to overcome this obstacle by simulating motions in proteins using the X-ray structure as the starting point for their calculation. The steadily growing computer power, new developments of faster algorithms and better physicochemical parameterizations in recent years have improved dynamic simulations. Soon, larger molecular dynamic simulations will be possible and hopefully allow a deeper investigation of the importance of protein dynamics in molecular binding.[52]

But most likely the explicit simulation of water molecules in and around proteins will have the biggest impact on our comprehension about molecular binding. Molecules are solvated in water and their interaction occurs in water. For many years, water was necessarily omitted in molecular docking and mapping applications as their *in silico* simulation was computational expensive. It was hoped that, in general, shape and physicochemical complementarity would be sufficient to drive molecular interactions. But many crystal structures of proteins show conserved water molecules at binding interfaces or next to binding sites and suggest an active role of water molecules in the protein-ligand complex.[53] Especially for molecular parts that interact via hydrophobic interactions by decreasing the entropy of the water molecule network, water acts as a "molecular glue" and induces the approach of protein and ligand molecules. The first methodologies that simulated hydration effects on protein structures considered water as a continuum, but had, in general, limited success. A second generation of simulation software treated water molecules explicitly but did not reached the expected accuracy especially due to the immense computational cost that dynamic simulations require. The growing computer power will eventually also help in this field to provide simulations of hydration effects under physical conditions.[53]

Once we achieve a comprehensive understanding of the fundamental processes in molecular binding, the *de novo* design of enzymes, i.e. the alternation of the enzymatic function, will be within reach. Other than inorganic catalysts, enzymes catalyze their reactions under mild conditions with high specificity and rate enhancements. This unique property makes enzymes attractive for many industrial processes although often they do not catalyze the required chemical reactions. Methods like rational-design and directed evolution in protein engineering have shown to be very useful in producing desired functionality in enzymes. As the factors for protein integrity namely, hydrogen bonds and hydrophobic effects, are well understood, many enzymes have been successfully altered to stabilize the structural integrity against harmful chemicals, or extreme temperature and pH conditions. Comparable results could not be obtained for altering the catalytic machinery of enzymes.[54] Only few enzymes so far have been successfully altered, like the modification of an inert ribose-binding protein into a highly active triose-phosphate-isomerase.[55] As long as the general mechanisms of molecular binding and catalysis are little understood, such successful examples will remain rare. Improved understanding of the mechanisms for molecular binding will have an impact not only to function prediction in structural biology, but will also have effects within the fields of medicine and biotechnology.

References

1. Berman HM, Westbrook J, Feng Z, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**: 235–242.
2. International Union of Biochemistry and Molecular Biology. Nomenclature Committee and Webb, E.C. (1992) *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Academic Press, San Diego, London.
3. Fischer E. (1894) Einfluss der configuration auf die wirkung der enzyme. *Ber Dtsch Chem Ges* **27**: 2985–2993.
4. Koshland DE. (1958) Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci USA* **44**: 98–104.
5. Brenner SE. (2001) A tour of structural genomics. *Nature Rev* **2**: 801–809.
6. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. (1996) Protein clefts in molecular recognition and function. *Protein Sci* **5**: 2438–2452.
7. Nayal M, Honig B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* **63**: 892–906.
8. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. (2007) Shape variation in protein binding pockets and their ligands. *J Mol Biol* **368**: 283–301.
9. Laskowski RA. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* **13**: 323–330.
10. Kraut DA, Sigala PA, Pybus B, *et al.* (2006) Testing electrostatic complementarity in enzyme catalysis: hydrogen bonding in the ketosteroid isomerase oxyanion hole. *Plos Biol* **4**: 501–519.
11. Coleman RG, Sharp KA. (2006) Travel depth, a new shape descriptor for macromolecules: application to ligand binding. *J Mol Biol* **362**: 441–458.

12. Gutteridge A, Thornton J. (2005) Conformational changes observed in enzyme crystal structures upon substrate binding. *J Mol Biol* **346**: 21–28.
13. Uversky VN, Gillespie JR, Millett IS, *et al.* (2000) Zn(2+)-mediated structure formation and compaction of the "natively unfolded" human prothymosin alpha. *Biocheml Biophys Res Commun* **267**: 663–668.
14. James LC, Tawfik DS. (2003) Conformational diversity and protein evolution — a 60-year-old hypothesis revisited. *Trends Biochem Sci* **28**: 361–368.
15. Domingues FS, Rahnenfuhrer J, Lengauer T. (2004) Automated clustering of ensembles of alternative models in protein structure databases. *Protein Eng Des Sel* **17**: 537–543.
16. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. (2002) Analysis of catalytic residues in enzyme active sites. *J Mol Biol* **324**: 105–121.
17. Glaser F, Pupko T, Paz I, *et al.* (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19**: 163–164.
18. Lichtarge O, Bourne HR, Cohen FE. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257**: 342–358.
19. Morris GM, Goodsell DS, Halliday RS, *et al.* (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**: 1639–1662.
20. Livesay DR, Jambeck, P, Rojnuckarin, A, Subramaniam, S. (2003) Conservation of electrostatic properties within enzyme families and superfamilies. *Biochemistry* **42**: 3464–3473.
21. Tsuchiya Y, Kinoshita K, Nakamura H. (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins* **55**: 885–894.
22. Kinoshita K, Furui, J, Nakamura H. (2002) Identification of protein functions from a molecular surface database, eF-site. *J Struct Funct Genomics* **2**: 9–22.
23. Honig B, Nicholls A. (1995) Classical electrostatics in biology and chemistry. *Science* **268**: 1144.
24. Klapper I, Hagstrom R, Fine R, Sharp K, Honig B. (1986) Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. *Proteins* **1**: 47–59.
25. Elcock AH. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J Mol Biol* **312**: 885–896.
26. Ondrechen MJ, Clifton JG, Ringe D. (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci USA* **98**: 12473–12478.
27. Silberstein M, Dennis S, Brown L, Kortvelyesi T, Clodfelter K, Vajda S. (2003) Identification of substrate binding sites in enzymes by computational solvent mapping. *J Mol Biol* **332**: 1095–1113.

28. Davis AM, Teague SJ. (1999) Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew Chem Int Ed* **38**: 736–749.
29. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. (1985) Hydrophobicity of amino acid residues in globular proteins. *Science* **229**: 834–838.
30. Kortemme T, Morozov AV, Baker D. (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* **326**: 1239–1259.
31. Panigrahi SK, Desiraju GR. (2007) Strong and weak hydrogen bonds in the protein-ligand interface. *Proteins* **67**: 128–141.
32. Wallace AC, Laskowski RA, Thornton JM. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng* **8**: 127–134.
33. McDonald IK, Thornton JM. (1994) Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**: 777–793.
34. Laurie AT, Jackson RM. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **21**: 1908–1916.
35. Matsuzaki R, Fukui T, Sato H, Ozaki Y, Tanizawa K. (1994) Generation of the topa quinone cofactor in bacterial monoamine oxidase by cupric ion-dependent autooxidation of a specific tyrosyl residue. *FEBS Lett* **351**: 360–364.
36. Smith JJ, Thomson AJ, Proudfoot AE, Wells TNC. (1997) Identification of an Fe(III)-dihydroxyphenylalanine site in recombinant phosphomannose isomerase from *Candida albicans. Eur J Biochem/FEBS* **244**: 325–333.
37. Henrick K, Thornton JM. (1998) PQS: a protein quaternary structure file server. *Trends Biochem* Sci **23**: 358–361.
38. Bashton M, Nobeli I, Thornton JM. (2006) Cognate ligand domain mapping for enzymes. *J Mol Biol* **364**: 836–852.
39. Iyer N, Jayanti S, Lou K, Kalyanaraman Y, Ramani K. (2005) Three-dimensional shape searching: state-of-the-art review and future trends. *Comput-Aided Des* **37**: 509–530.
40. Porter CT, Bartlett GJ, Thornton JM. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl Acids Res* **32**: D129–D133.
41. Barker JA, Thornton JM. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **19**: 1644–1649.
42. Najmanovich RJ, Allali-Hassani A, Morris RJ, *et al.* (2007) Analysis of binding site similarity, small-molecule similarity and experimental binding profiles in the human cytosolic sulfotransferase family. *Bioinformatics* **23**: e104–e109.
43. Schmitt S, Kuhn D, Klebe G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol* **323**: 387–406.

44. Rosen M, Lin SL, Wolfson H, Nussinov R. (1998) Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng* **11**: 263–277.
45. Fetrow JS, Skolnick J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J Mol Biol* **281**: 949–968.
46. Binkowski TA, Adamian L, Liang J. (2003) Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J Mol Biol* **332**: 505–526.
47. Binkowski TA, Naghibzadeh S, Liang J. (2003) CASTp: computed Atlas of Surface Topography of proteins. *Nucl Acids Res* **31**: 3352–3355.
48. Jain AN. (2006) Scoring functions for protein-ligand docking. *Curr Protein Peptide Sci* **7**: 407–420.
49. Sotriffer C, Klebe G. (2002) Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Farmacology* **57**: 243–251.
50. Ledvina PS, Yao N, Choudhary A, Quiocho FA. (1996) Negative electrostatic surface potential of protein sites specific for anionic ligands. *Proc Natl Acad Sci USA* **93**: 6786–6791.
51. Khersonsky O, Roodveldt C, Tawfik DS. (2006) Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol* **10**: 498–508.
52. Karplus M, McCammon JA. (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **9**: 646–652.
53. Levy Y, Onuchic JN. (2006) Water mediation in protein folding and molecular recognition. *Ann Rev Biophys and Biomol Struct* **35**: 389–415.
54. Bolon DN, Voigt CA, Mayo SL. (2002) *De novo* design of biocatalysts. *Curr Opin Chem Biol* **6**: 125–129.
55. Dwyer MA, Looger LL, Hellinga HW. (2004) Computational design of a biologically active enzyme. *Science* **304**: 1967–1971.
56. Apweiler R, Bairoch A, Wu CH, *et al.* (2004) UniProt: Rhe universal protein knowledgebase. *Nucl Acids Res* **32**: 115.
57. Holliday GL, Bartlett GJ, Almonacid DE, *et al.* (2005) MACiE: a database of enzyme reaction mechanisms. *Bioinformatics* **21**: 4315–4316.

This page intentionally left blank

Chapter 9

Atomistic Simulations of Reactions and Transition States

M. Meuwly*

9.1 Introduction

Much of the exciting progress in and insights from atomistic simulations is intimately related to the concept of a simplified representation of the intermolecular interactions in complex systems. This dates back to the late 1960s when such models — called force fields — were first constructed and used in the refinement of crystal structures.[1] The first molecular dynamics (MD) simulation of the bovine pancreatic trypsin inhibitor (BPTI) in 1977, together with the realization that B-factors can be related to the thermal motion of the protein atoms, were instrumental in replacing the view of rigid proteins.[2,3] Since then, the role of flexibility in structural biology is undisputed.[4–6] Much of this progress is strongly influenced and driven by atomistic simulations and continued experimental developments.

Over the past decade, the application of force field-based methods has demonstrated that they are useful in understanding and describing various processes including protein folding, protein-ligand interactions and protein dynamics, at least at a qualitative level. Recent developments in capturing finer details of the intermolecular interactions, which should pave the way for more quantitative studies, are broadly

*Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland. Email: m.meuwly@unibas.ch.

summarized here. One of the most challenging problems in structural biology and biophysics is to follow a system in executing its function. This is often accompanied by switching between a reactant and a product state connected by transition states and metastable states.

Atomistic simulations are unique in that they provide details about the motion of every atom for the property of interest. One example is the diffusion of small molecules in protein cavities, such as in Myoglobin. It is, however, essential that results from such simulations are critically validated in view of experimental data. This has become possible through dramatically increased computer power, which allows simulations to be carried out on time scales relevant to experiment. Conversely, the development of new experimental techniques such as ultra fast spectroscopy[7] or two-dimensional infrared spectroscopy[8] also provides valuable benchmarks for a critical assessment of the numerical results. With continuous improvement of existing and the development of new numerical methods, it is also possible to compare calculations with more traditionally available experimental data, such as equilibrium rate constants. For this, however, it is necessary to investigate the paths by which the system under investigation evolves from the reactant to the product state. This is a challenging problem due to the high dimensionality of the phase space involved.

In the following, the concept of atomistic potential energy functions and recent developments in this area are briefly summarized. Several excellent and recent reviews exist on this topic, and the reader is encouraged to consult these for more in-depth information.[9,10] In a subsequent chapter, computational strategies to identify and sample transition states in high-dimensional systems are presented. This is followed by examples from different fields relevant to structural biology and biophysics. Finally, the outlook discusses future directions and applications of atomistic simulations.

9.2 Potential Energy Functions for Biomolecular Simulations

Every simulation of a macromolecular system requires prescriptions (or models) according to which the total energy for a given configuration can be calculated. The shorter the length scale that should be resolved (e.g. atomistic vs. coarse grained), the more detailed the representation of the interaction potential needs to be. For atomistic simulations, the configurations of interest are given by the coordinates $\vec{x}$ of all atoms.

Given the $3N$ coordinates $\vec{x}$ of a macromolecular system, a potential energy function provides the total interaction potential $V(\vec{x})$. For systems containing many atoms (typically $N > 10^3$), models have to be developed that allow $V(\vec{x})$ to be evaluated. One such class of models are empirical force fields. They decompose the total energy into internal (or bonded) and external (or nonbonded) contributions: V_{bonded} and V_{nb}, respectively.[1,11,12] The bonded terms are related to the covalent interactions and are further separated into contributions from the chemical bonds (distance r), the valence angles (θ), and dihedral angles (ϕ) described by the following equations:

$$\begin{aligned} V_{bond} &= \sum K_b (r - r_e)^2 \\ V_{valence} &= \sum K_\theta (\theta - \theta_e)^2 \\ V_{dihe} &= \sum \sum_n K_\phi (1 + \cos(n\phi - \delta)) \end{aligned} \tag{9.1}$$

Here, K represents the force constants associated with the particular type of interaction; r_e and θ_e are equilibrium values, n the periodicity of the dihedral, and δ the phase that determines the location of the maximum. The summations are carried out over all respective terms.

Non-bonded interactions include electrostatic and van der Waals terms, which are:

$$\begin{aligned} V_{elstat} &= \sum k \frac{q_i q_j}{r_{ij}} \\ V_{vdw} &= \sum \varepsilon_{ij} \left[\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\min,ij}}{r_{ij}} \right)^{6} \right] \end{aligned} \tag{9.2}$$

where the summations include all non-bonded atom pairs. q_i and q_j are the partial charges of the atoms i and j involved in the electrostatic interaction, and k is the Coulomb force constant, which depends on the units used. In textbooks, $k = 1/4\pi\varepsilon_0$, where ε_0 is the vacuum dielectric constant. If q_i and q_j are measured in units of elementary charge e and distances in Angstroms (which is customary in most force fields), $k \approx 331.843$ Åe^2 kcal/mol gives the energy in kcal/mol. For the van der Waals terms, the potential energy is expressed as a Lennard-Jones potential with well depth $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$ and range $R_{\mathrm{min},ij} = (R_{\mathrm{min},i} + R_{\mathrm{min},j})/2$ at the Lennard-Jones minimum. This interaction captures long-range dispersion ($\propto r^{-6}$) and exchange repulsion ($\propto r^{-12}$), where the power of the latter is chosen for convenience. The combination of Eqs. 9.1 and 9.2 constitute a minimal model for a force field that might be extended by using explicit terms for hydrogen bonds or for metal-containing systems.

The merits and shortcomings of a particular force field are encoded in the actual parameter values (K_b, K_θ, K_ϕ, q_i, etc). The step from a specific mathematical representation of the inter- and intramolecular interactions to a force field consists of determining a set of parameter values, which are achieved by fitting the parameters to particular target data.[11–13] In this context, it is important to mention that parallel to the mathematical form of a force field, specific properties of the atoms are captured by the atom types. For example, C–C single bonds are reflected in a larger value for r_e and a smaller force constant K_b compared to a C=C double bond. As another example, hydrogen atoms bonded to an oxygen generally have a larger partial charge than those bonded to carbon atoms. A typical force field has of the order of 100 different atom types (with associated q_i, ε_i, $R_{\mathrm{min},i}$), which leads to several hundred parameters to be determined.

For biomolecular simulations, widely used force fields include CHARMM (Chemistry at HARvard Molecular Mechanics), AMBER (Assisted Model Building with Energy Refinement), OPLS (Optimized Potentials for Liquid Simulations), and GROMOS (GROningen MOlecular Simulation). This list is by no means exhaustive and a more comprehensive compilation can be found in Ref. 9.

As briefly mentioned above, the parameters of particular force fields differ largely through the target data they are fitted to. For example, partial charges in the early CHARMM (Param 19)[14] and AMBER[12] parametrizations were determined by using a distance-dependent dielectric constant to represent the solvent, whereas OPLS[13] uses an explicit representation of the aqueous environment. This also highlights that it is an important decision by the user which force field to employ for a particular application. Finally, it must be emphasized that parameters between different empirical force fields are in general not transferable. This already follows from the above remark that force field parameters are determined vis-a-vis particular training data. Thus, if a molecule is not available in the force field chosen, the missing parameters should be determined along similar lines as used for the existing parameters. Often, a large number of parameters can be determined by analogy with existing fragments. However, additional *ab initio* calculations may be required to maintain the balance in the energy evaluations. Further details on parameter determination can be found in Ref. 10.

The empirical force fields discussed so far are also known as “class I force fields”, which contain harmonic terms for bonds and valence angles and do not include explicit coupling between different degrees of freedom (such as stretch-stretch or stretch-bend interactions). A further development represent “class II force fields” where anharmonic terms and couplings between bond, angle, dihedral, and out-of-plane internal coordinates are included.[15]

Recently, particular attention in force field development has been paid to go beyond fixed atomic point charge electrostatics. Fixed point charges centered on the atoms do not allow the correct description of: i) the particular features of the molecular charge distribution, and ii) the response of the charge distribution to changes in the environment. An early example where higher multipoles were used to better describe the electrostatic field around a molecule is carbon monoxide (CO) in Myoglobin (Mb). Straub and Karplus included an additional charge site at the CO-center of mass to capture the quadrupole moment of CO. More recently, this model has been refined by allowing the partial charges to fluctuate as a function of the

CO bond length *r*.[16] This describes the reversal of the direction of the dipole moment $\mu(r)$ along r and accurately reproduces the quadrupole moment $\Theta(r)$. With this model, it was possible to correctly describe the docking site (B state), the infrared spectrum of CO in this state, and the energetics between the two metastable states Fe···CO and Fe···OC.[16,17]

Atom- or bond-centered multipole expansions truncated after the quadrupole have been shown to reproduce the electrostatic potential calculated directly from the wavefunction to within 0.1%.[18] An efficient formulation of higher multipole moments is provided by the distributed multipole analysis (DMA)[19] or the effective fragment method.[20] However, coherent implementations of such ideas into existing force fields are part of current research efforts.[21]

Including the response of the charge distribution to changes in the environment requires the explicit inclusion of polarization effects, which represents a next significant step in force field development. This can be done in a variety of ways.[22,23] The different methods to incorporate polarization are fluctuating charges, the Drude model, and a method based on induced dipole moments. The polarization energy is given as

$$U_{pol} = \frac{1}{2}\sum_{i}\vec{\mu}_i\vec{E}_i \tag{9.3}$$

where $\vec{\mu}_i$ is the dipole moment of atom i, and $\vec{E}_i$ is the electric field at the position of atom i originating from all surrounding point charges. Here,

$$\vec{\mu}_i = \alpha\left(\vec{E}_i^0 - \sum_{i\neq j}^{N}\frac{-3\vec{r}_{ij}\vec{r}_{ij} + \mathbf{1}r_{ij}^2}{4\pi\varepsilon_0 r_{ij}^5}\vec{\mu}_j\right) \tag{9.4}$$

where $\vec{E}_i^0$ is the electrostatic field due to the N static charges, $\mathbf{1}$ is the unit matrix, and r_{ij} is the distance between two charges q_i and q_j. Usually, the above coupled equations (the second term depends on $\vec{\mu}_j$) are solved iteratively,[24] although noniterative procedures are also available.[25]

A considerable amount of work on polarizable force fields has been concerned with improved water models.[26] Other applications considered the solvation of ions and condensed phase properties of small molecules.[27] Also, the simulation of several small proteins including solvent for times in the nanosecond range have been reported.[28] Further progress in developing and parametrizing polarizable force fields can be expected. Once consistent polarizable force fields for biomolecular simulations are available, the merits of the increased computational effort can be truly assessed.

9.3 Atomistic Simulations of Reactions in Proteins

9.3.1 *Potential Energy Functions*

The mathematical form of the empirical force fields discussed in the previous section is not suitable to describe chemical reactions where chemical bonds are broken and formed. An important step to investigate reactions by simulation methods has been the introduction of mixed quantum mechanical/classical mechanics methods (QM/MM).[29–31] In QM/MM, the total system is divided into a (small) reaction region, for which the energy is calculated quantum mechanically, and an (large) environment, which is treated with a conventional force field. The majority of applications of QM/MM methods to date use semi-empirical (such as AM1, PM3,[32] SCC-DFTB[33,34]) or density functional theory methods[35] on isolated structures. Studies including the nuclear dynamics (QM/MM MD) are still the exception.[36–43] Typically, the QM part contains of the order of several tens of atoms. It should also be noted that studies of reactive processes in the condensed phase often employ energy evaluations along predefined progression coordinates,[32,38] i.e. the system is forced to move along a set of more or less well suited coordinates. Thus, molecular orbital (i.e. DFT or *ab initio*) QM/MM calculations cannot yet be used in fully quantitative studies. One of the main reasons why molecular orbital QM/MM calculations are not yet used routinely in fully quantitative studies is related to the fact that the evaluation of the intermolecular interactions in the QM region is computationally too expensive to allow proper configurational averaging, which is required for reliably estimating essential quantities such as free energy changes. Alternatives to QM/MM calculations have been developed over the past two decades. They include EVB (Empirical Valence Bond),[44] AVB (Approximate Valence Bond),[45] MCMM (multi-configuration molecular mechanics)[46] and RMD (Reactive Molecular Dynamics).[47–50]

9.3.2 *Rate Constants and Transition States*

One of the primary observables from a biophysical or biochemical experiment is the rate of a particular process as a function of external driving coordinates. The "process" of interest may be as diverse as the folding of a protein (configurational reorganization), proton transfer in enzymatic catalysis or rebinding of a ligand after photodissociation. External driving forces include, but are not limited to, temperature T, pressure p, pH of the solvent, or the amino acid composition of the polypeptide chain. All these scenarios are related by the fact that the experiment observes the transition from an initial stable state (reactant R) to a final stable state (product P) — possibly via metastable intermediate states — separated by an energetic barrier, which is large compared to the thermal fluctuations in the system. Crossing this energy barrier is only possible if the system concentrates sufficient energy ($\gg k_B T$) in one or a few degrees of freedom, which promote the transition. This energy flows into the relevant degrees of freedom through thermal fluctuations.

9.3.2.1 *Transition state theory*

In its original formulation, transition state (TST or activated-complex theory) is based on two essential assumptions. First, there is a separation of time scales between the dynamics *within states* R *and* P *respectively*, and the dynamics *between states* R *and* P. Second, every trajectory that reaches the transition state coming from R relaxes to P (no recrossing). Based on these assumptions, the temperature-dependent rate for crossing the transition state between R and P is

$$k(T) = \frac{k_B T}{h} \exp\left(-\frac{\Delta^{\neq} G^0}{RT} \right) \tag{9.5}$$

where ΔG^0 is the Gibbs free energy of activation, h is Planck's constant, k_B is the Boltzmann constant, and R is the gas constant. A different formulation makes this expression more amenable to dynamical simulations. The overall TST rate constant can also be written as the product of the rate at which the system moves across the dividing surface, which separates P from R, and the probability to find the system at the transition state $x^{\ddagger}$:

$$k_{\text{TST}} = \frac{1}{2} \left\langle |v| \, \delta(x - x^{\ddagger}) \right\rangle \tag{9.6}$$

where v is the velocity of the system at the transition state (i.e. $\dot{x}^{\ddagger}$), and the factor of 1/2 takes into account that the system moves from R to P only half the time. The average has to be carried out over the reactant configurations. If the barrier region is approximated as a harmonic oscillator, the expression yields

$$k_{\text{TST}}^{\text{HO}} = \frac{\omega}{2\pi} \exp(-\beta \Delta E) \tag{9.7}$$

which can be interpreted as the attempt frequency ω of oscillations in the reactant minimum multiplied by the probability $\exp(-\beta\Delta E)$ of reaching the transition state. For atomistic simulations, often a form derived from statistical thermodynamics is used:[51]

$$k(T) = \frac{1}{2} \kappa \frac{\langle \dot{x} \rangle \, \rho(x^{\ddagger})}{\int \rho(x) dx} \tag{9.8}$$

Here, κ is the transmission coefficient, $\rho(x)$ is the probability density of the configurational (progression) coordinate x, and $\langle \dot{x} \rangle$ is the average absolute value of the velocity at the transition state $\dot{x}^{\ddagger} = dx/dt|_{x=x^{\ddagger}}$

9.3.2.2 *Progression coordinates for locating transition states*

Structural rearrangements or reactions typically involve more than one coordinate along which the process of interest (P → R) occurs. Only in the simplest of systems (e.g. some proton transfer reactions) can the dominant reaction pathway be "guessed", and only one or a small number of coordinates are involved. In more complex systems, the relevant progression coordinates are less obvious and different choices can lead to different results, such as demonstrated for Na^+ transport through the gramicidin ion channel.[52,53] Another example, where the correct progression coordinate appears to be simple, is the isomerization of a tyrosine residue in the bovine pancreatic trypsin inhibitor (BPTI).[54] If the process is described by using the "obvious" torsion angle of Tyr35, the energy barrier is underestimated by about 5 kcal/mol. To arrive at a more realistic estimate it is important to include additional protein atoms whose non-bonded interactions contribute to the barrier in the definition of the progression coordinate.[54] This situation is similar to simple "coordinate driving" in *ab initio* electronic structure calculations where the system is forced along a predefined progression coordinate. Finally, it is also possible that not only one but a number of pathways contribute to the reaction kinetics. Some paths may be dominant in one temperature or pressure regime, whereas others contribute significantly at different T or p. The aim then is to determine the dominant pathways under the given external conditions, i.e. the respective thermodynamic variables. Methods that include temperature are MaxFlux Reaction Paths,[55,56] Noisy dynamics (such as diffusional paths[57] and Transition Path Sampling[58]), or Reactive Molecular Dynamics.[47,50]

In MaxFlux, the aim is to determine paths $\vec{r}$ between P and R, which minimize the functional $\mathrm{C}(\vec{r}) = \exp[\beta U(\vec{r})]$, where $\beta = 1/k_{\mathrm{B}}T$ is the Boltzmann factor. This method contains temperature explicitly, which may show up by the fact that $\vec{r}$ avoids the saddle point.[55,56]

This method has been applied to the coil-to-helix transition in polyalanine.[59]

In noisy dynamics, the reaction is understood as a diffusion process between reactant and product described by coordinates $\vec{r}_R$ and $\vec{r}_P$.[57] The diffusion is then decomposed into a finite number of intermediate structures $\vec{r}_q$ (a Markov chain of states) and the probability for the reaction $p(\vec{r}_P \mid \vec{r}_R)$ is the integral over all joint probabilities.

$$p(\vec{r}_P \mid \vec{r}_R) \propto \int p(\vec{r}_P \mid \vec{r}_{N-1}) \ldots p(\vec{r}_2 \mid \vec{r}_1) p(\vec{r}_1 \mid \vec{r}_R) d\vec{r}_{N-1} \ldots d\vec{r}_2 d\vec{r}_1 \tag{9.9}$$

Transition pathways are then determined from maximizing $p(\vec{r}_P \mid \vec{r}_R)$. This formulation is similar in spirit to a path integral formulation to solve the Schrödinger equation.[60] Temperature enters through a suitable choice of the transition probabilities

$$p(\vec{r}_i \mid \vec{r}_{i+1}) = (4\pi Ddt)^{-1/2} \exp\left[-\frac{(\vec{r}_{i+1} - \vec{r}_i - D\beta\vec{F}(\vec{r}_i)dt)^2}{4Ddt} \right] \tag{9.10}$$

where D is the diffusion constant. Based on this approach, Chandler and coworkers developed a method, coined transition path sampling (TPS), for calculating rate constants.[58] The approach finds reaction paths by sampling the path ensemble and does not require predefined progression coordinates. TPS is a computationally intensive, iterative simulation method for which at least one pathway of the event of interest is required as a prerequisite. The majority of applications involving TPS has been concerned with finding pathways between different conformational substates.[61] Very recently, the method has also been applied to chemical reactions in conjunction with QM/MM calculations.[62,63]

Reactive Molecular Dynamics (RMD) originally was proposed for and applied to ligand binding in Myoglobin-Ligand systems.[47,49,64] RMD compares the energies of the R and P states V(R) and V(P) along the trajectory. Whenever V(R)- V(P) changes sign, a possible transition state has been found. Recently, RMD has been formulated such that the bond-breaking/bond-forming process is followed in time.[50] Thus, recrossing of the reactive seam is possible and no predefined progression coordinate is required. The essence of RMD is that it uses validated empirical force fields for the R and P states together with one empirical parameter Δ, which describes the asymptotic energetic separation between $V(\mathrm{R})$ and $V(\mathrm{P})$. This procedure has been applied to the rebinding of NO to Myoglobin. The value of Δ can be chosen to reproduce either *ab initio* calculations or experimental data. Given this, conventional MD trajectories are run in time and crossings of the reactive seam are recorded and analysed. Because the process is followed in time, this approach bears the potential to determine and investigate largely unbiased progression coordinates *a posteriori*.[50] The problem of determining progression coordinates has received much recent attention.[58,65–68]

In the following, illustrative examples for the use of conventional and modified force fields as well as mixed QM/MM simulations, which involve barrier-crossing phenomena, are discussed. The topics discussed cover both, work from our laboratory and from other groups.

9.4 Illustrative Examples

9.4.1 *Transition Path Sampling for Protein Folding*

In a series of publications, transition path sampling (TPS) has been applied to the investigation of β-hairpin folding of protein G and the folding of the Trp-cage in explicit solvent.[61,69] As mentioned above, the advantage of TPS is that no pre-defined reaction (or progression) coordinate is required to follow the transition between two states. On the other hand, TPS is computationally very demanding and at least one pathway from educt to product (folded to unfolded) is required. Using TPS for protein folding in explicit solvent, a number of interesting results were found.[61] Most notably, the role of explicit water molecules as a lubricant for the folding process has been suggested. This hypothesis has already been put forward before,[70,71] but was difficult to demonstrate directly from simulations. Furthermore, in protein folding/unfolding simulations, often an extensive MD simulation is carried out from which the free energy profile is then calculated by defining "suitable" progression coordinates. The TPS simulations have convincingly demonstrated that depending upon the choice of the progression coordinate (common choices are the radius of gyration ρ or the number of native hydrogen bonds), the calculated free energy barriers separating the folded, partially folded, and unfolded states can drastically vary.[61] As an example, the free energy barrier between the frayed (F) and the hydrophobic (H) state of the β-hairpin folding of protein G is found to be >7 kcal/mol from transition interface sampling (which is based on TPS). This compares with a barrier of 2 to 3 kcal/mol from replica exchange calculations, depending on the progression coordinates used.[61] Transition path sampling also provides structural information about transition state configurations. For the hydrophobic to unfolded transition one member of the transition state ensemble is shown in Fig. 9.1.

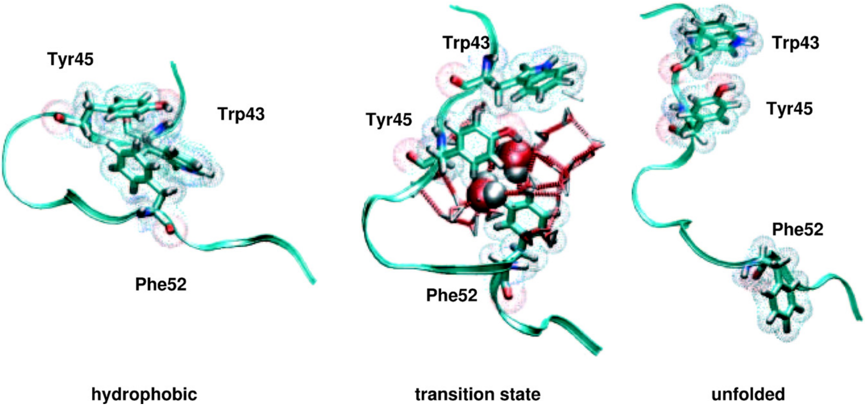


Fig. 9.1 Structures of a metastable (hydrophobic) and the final (unfolded) state together with one representative transition state structure between the two for the 16-residue C-terminal fragment of protein G (Images adapted from Ref. 99). The transition from the H to the U state is accompanied by the disruption of the water network and subsequent water penetration. Residues Trp43, Tyr45, and Phe52 form the hydrophobic core and are represented as a stick model. The backbone is shown in ribbon representation.

9.4.2 *Ligand Rebinding in Myoglobin-CO/NO*

Another area in biomolecular simulations where the transition between products and reactants is of central importance are ligand-binding reactions. The paradigm system for such processes is the recombination of small ligands after photodissociation from the heme group in myoglobin (Mb). Myoglobin is one of the primary model systems for studying protein structure and dynamics in general, and ligand binding and dynamics in particular.[72] The study of the binding of a number of neutral diatomic molecules, such as O_2, CO, or NO has greatly helped in understanding protein function and its relation to structure.[73] The small size and stability of this protein together with the wealth of experimental data has also made it an attractive and meaningful system for computational studies.[16,49,74–78]

9.4.2.1 *Ligand rebinding in MbCO*

Depending on the process under study, different approaches can be used to investigate the rebinding reaction. For MbCO, a model based on the diffusion equation and a simplified representation of the potential energy surface (including two progression coordinates) has been employed by Agmon and Hopfield.[79] Although simple in its formulation, the model captures a number of essential properties of the reaction.

An alternative procedure employed by Wolynes *et al.* uses a Hamiltonian which interpolates between the bound and the unbound system.[48] Structures from a long MD simulation are quenched to the transition state, which are further relaxed, and either end in the product or the reactant well. From several hundred simulations the distribution of barrier heights at different temperatures were determined. Both the width of the distribution at 10 K and the peak position at 300 K qualitatively agree with the experimental data. Because the approach is primarily based on relaxing the transition state structures it is difficult to calculate dynamical properties from it. Furthermore, the effect of the surrounding solvent was not included in this study and is in general difficult to be accounted for with this approach.

With advances in computational methods it was subsequently possible to include all intermolecular interactions in the bound (MbCO, the A state) and the unbound (Mb···CO, including the B state and the ligand in various Xenon pockets) state and to use umbrella sampling of the free energy surface along a meaningful progression coordinate.[17] In this case, the validity of the Fe–CO (center of mass) coordinate q as a progression coordinate was established through extensive MD simulations.[16] This model was also shown to provide quantitative information about the most likely position of the CO molecule in the docking site (B state), the infrared spectrum associated with this state, and the free energy barrier for CO rotation in the B state. Thus, the only unknown quantity is the energetic separation Δ between the bound and the unbound potential energy surface. Based on these observations, the one-dimensional free energy profile for the bound and the unbound state along q can be calculated and used in the Smoluchowski equation to investigate the rebinding kinetics as a function of Δ. For this, the free energy surfaces are diabatized around the crossing point (see Fig. 9.2) to yield a lower and an upper diabat, $G_l(q)$ and $G_u(q)$, respectively. Starting from an initial population (which is produced by a laser pulse in the experiment), the rebinding time is measured by following the relaxation of the initial distribution $p(t = 0)$ to its equilibrium. The experimentally measured rebinding time is $\tau \approx 100$ ns, which corresponds to $\Delta = 4$ kcal/mol. From the kinetic constants, an effective barrier for the B → A transition (see Figs. (B) and (C)) of 4.5 kcal/mol has been calculated,[80] which favorably compares with 4.3 kcal/mol from the simulations.[17] It should be noted that the only free parameter in these simulations is the asymptotic energy separation Δ between the bound and the unbound potential energy surface *in* the protein, which is difficult to determine from experiment, simulations, or calculations.

More recently, this rebinding reaction has been studied in higher dimensions by directly analysing unbiased MD simulations. This was deemed necessary because umbrella sampling can introduce additional stabilization of the ligand through relaxation of the protein. Thus, several nanoseconds of free MD simulation for the migration of

CO between the A state and the Xe4 pocket via the B state were analyzed.[81] The probability density functions $P(u,v)$ — where u and v span a two-dimensional projection of the migration pathway — were used to approximate the free energy surface $G(u,v) \propto - k_B T \log P(u,v)$ Here, k_B is the Boltzmann constant and T is the temperature. This $G(u,v)$ is a rough potential energy surface (see Fig. 9.2), which poses severe problems for solving the Smoluchowski equation. Using a hierarchical approach, following the relaxation of the initial probability distribution in time becomes tractable.[82] Figure 9.3 shows snapshots of the distribution function $p(u, v, t)$ at different times t after photodissociation, from which the rebinding time can be calculated. For different initial populations $p(u, v, t = 0)$, a rebinding time of $\tau \approx 100$ ns yields free energy barriers of 6 kcal/mol between the Xe4 pocket and the B state. For this barrier, no experimental data is available. Other simulations found a barrier of ≈2.5 kcal/mol and 4.5 kcal/mol for this transition.[83,84] The one-dimensional simulations discussed above give a value of 6.8 kcal/mol, which may somewhat overestimate the barrier and reflects the potential bias introduced through umbrella sampling.[17]

9.4.2.2 *Rebinding in MbNO*

For NO, the rebinding processes occur on much faster time scales. Thus, approaches based on the diffusion equation are less likely to be appropriate. Elber and coworkers have considered a model where the rebinding reaction progresses along the Fe-N$_{\mathrm{NO}}$ separation $R_{\mathrm{Fe\text{-}N}}$.[47] To describe the transition between the bound and the unbound state, a switching function depending on $R_{\mathrm{Fe\text{-}N}}$ was employed. This approach was used to investigate the differences in the picosecond recombination rates for mutants at position 29. With this model the correct experimental trends for the picosecond recombination in the different mutants was reproduced. A disadvantage of this approach is that it relies on a geometrical progression coordinate.

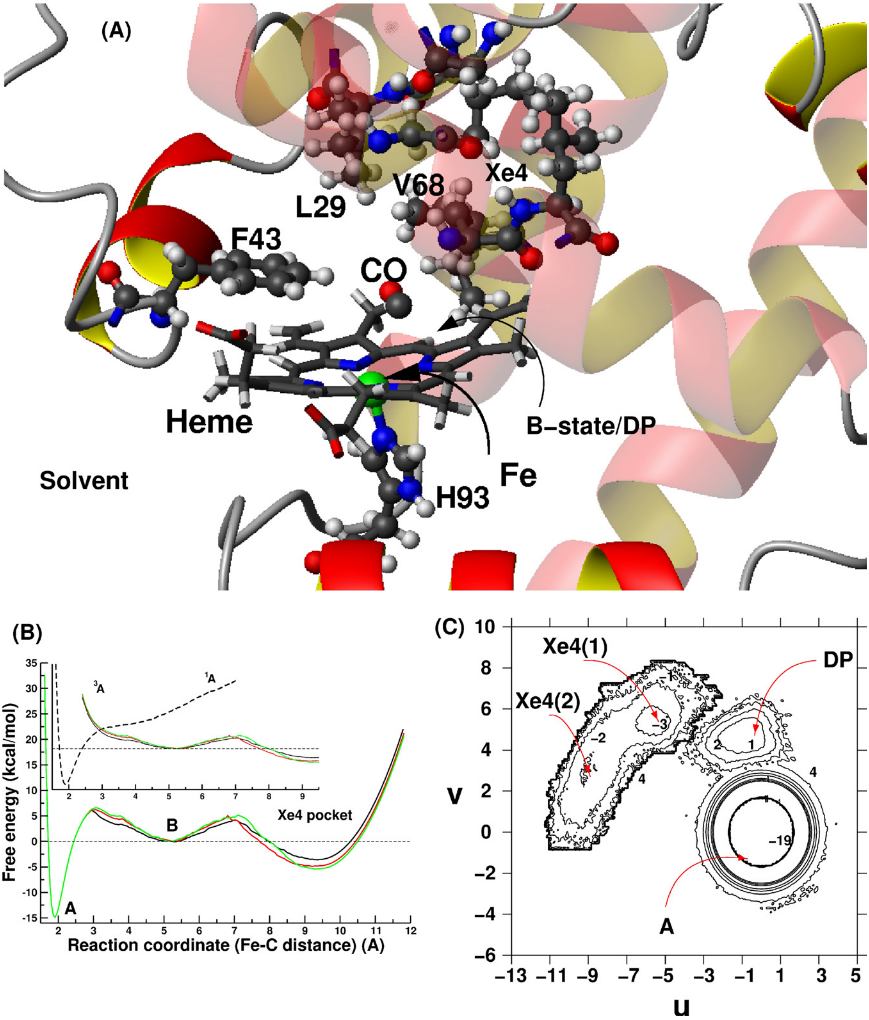


Fig. 9.2 **(A)** The heme pocket and its surrounding in myoglobin (Mb) with CO as a photodissociated ligand. Important residues are drawn in ball-and-stick representation. **(B)** Upper and lower diabats from umbrella sampling along the Fe-CO (center of mass) coordinate. **(C)** The two-dimensional free energy surface from unbiased MD simulations. Coordinates *u* and *v* describe the projection of the center of mass of CO onto the heme plane.

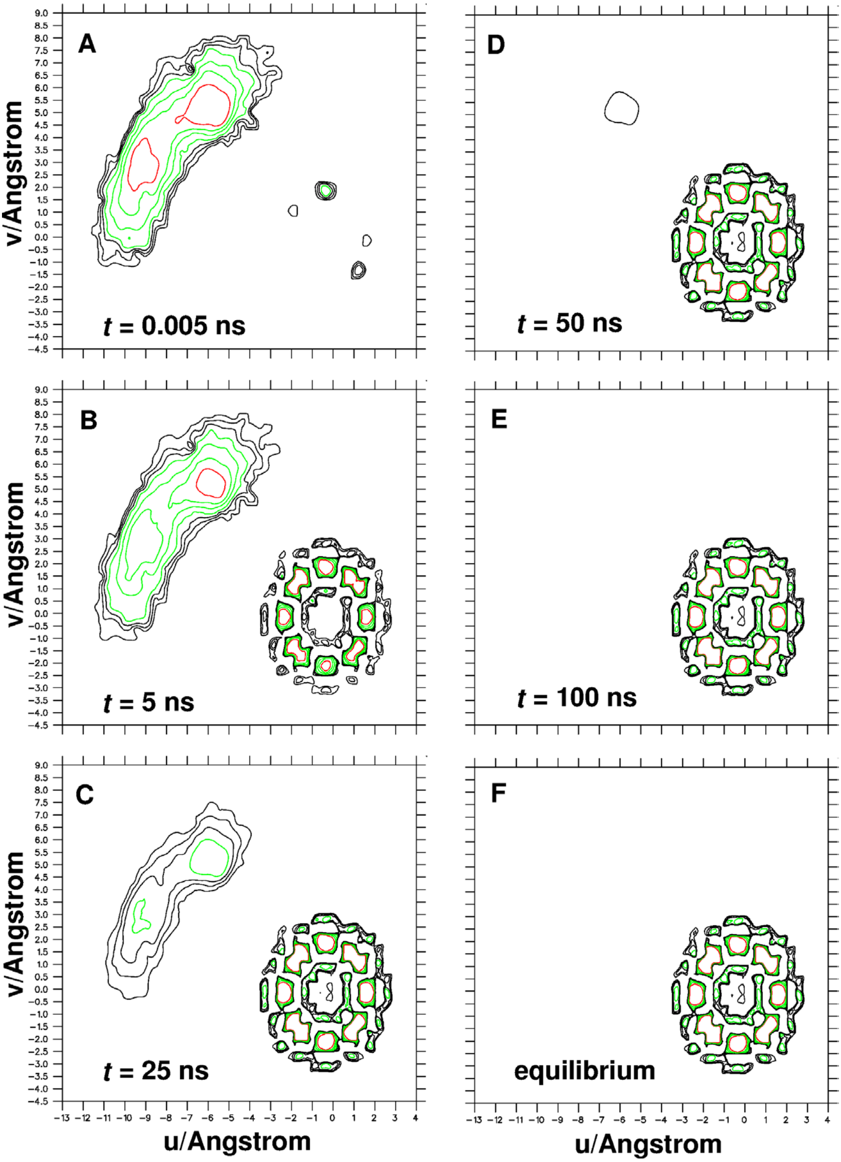


Fig. 9.3 Snapshots of the distribution function (A to E) at various times after photodissociation ($t = 0$) from solving the Smoluchowski equation using the rough free energy surface $G(u, v)$ [see Fig. 9.2 (c)] for T = 300 K. The initial condition is $p(u, v, t = 0) = \delta(u = -9.0, v = 4.0)$ which corresponds to CO in the Xe4 pocket. The last frame (F) shows the equilibrium distribution, which is $P_{\text{eq}} = \exp[-\beta G(u, v)]$. The rebinding time in this case is $\tau \approx 100$ ns.

Recently, a method has been proposed, which follows the reaction in time, allows for crossing and recrossing the reactive seam, and includes the effect of a solvent.[50] It is based on an accurate representation of the asymptotic states in the product and the reactant well (here: Mb···NO and MbNO). Starting on the lower of the two states, the algorithm locates crossings between the two states using an energy criterion. Once a crossing is located, the transition is carried out with probability one over a finite window in time, which depends on the process in question. During the transition, the two potential energy surfaces are mixed according to

$$V(\vec{x}) = f(t) \times V_R(\vec{x}) + (1 - f(t)) \times V_P(\vec{x}) \tag{9.11}$$

where $f(\mathrm{t})$ is a sigmoid function which changes from 1 to 0 over the time window. Figure 9.4 displays an example of a reactive trajectory, which starts on the unbound state, relaxes into the metastable FeON state, recrosses to the unbound state, and finally relaxes into the global FeNO minimum. Starting in the unbound state, several thousand trajectories were followed for the rebinding of NO to Mb. Analysis of the rebinding times revealed two time scales that differ by one order of magnitude. The calculated time constants from a double exponential fit are of the correct order of magnitude but somewhat too small (3.8 and 18.0 ps compared with 28 and 280 ps,[85] 5.3 and 133 ps[86]), in particular for the slower component. However, the ratio between the fast and slower time constants is of similar order as found in experimental data. Given the large differences between the reported experimental data,[85–87] the calculated results can be considered to qualitatively agree with experimental data.

9.4.3 *Enzymatic Reactions*

Transition states are also of central importance in understanding enzymatic reactions. With increasing computer power, the application of mixed quantum mechanical/molecular mechanics methods has become possible. However, it should be mentioned that most studies

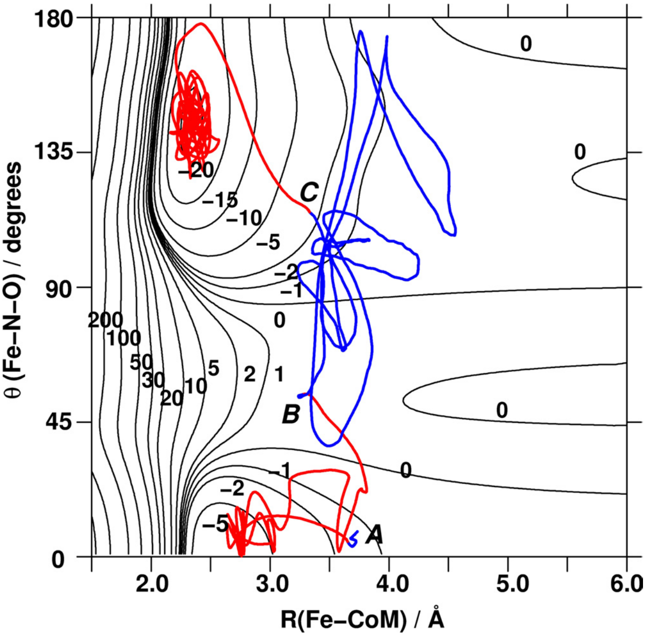


Fig. 9.4 Reactive trajectory from simulations of NO rebinding to Myoglobin. The trajectory starts in the unbound state **(A)** blue trace, encounters a crossing, and switches to the bound state (red trace). After sampling the secondary Fe–ON minimum a further crossing is located **(B)** after which the trajectory samples the unbound state extensively and finally crosses at **C** to relax into the global minimum, which is the Fe–NO configuration. The trajectory is projected onto the bound state potential energy surface $V(R, \theta)$, where R is the Fe–NO (center of mass) distance and θ is the Fe–NO angle. The surface is calculated at the B3LYP//VDZ(Fe,N,O)/3-21G(C,H) level[50] and energy contours are given in kcal/mol.

still employ energy minimization techniques and not true dynamical studies. Also, the majority of applications use semi-empirical methods (such as AM1 or SCC-DFTB) for the quantum part, which may or may not be appropriate.

Two studies have recently used TPS together with semi-empirical QM/MM to investigate the enzymes lactate dehydrogenase and chorismate mutase.[62,63] Studying the catalytic step, and including the dynamics in the active site, the first study showed that both concerted and stepwise hydride/proton transfers are possible. Furthermore, from analyzing the trajectories it was also found that a cooperative mechanism contributes to catalysis.[62] The second study analyzed around 1000 independent transition paths for the Claisen rearrangement reaction. This provided detailed information about the progression coordinates typically used in previous studies.[32] It was found that no single geometrical coordinate is suitable to represent the committor probabilities for the transition path structures. The commonly used reaction coordinate (a difference of two distances) showed some correlation with the committor probability whereas other coordinates (including dihedral angles) were not correlated.[63] However, even by analyzing the transition state structures it was not possible to determine a more suitable progression coordinate. Further analysis showed that the near attack conformations, which were proposed to be involved in the catalytic step,[88] did not appear in the transition state ensemble. As with lactate dehydrogenase, a cooperative compression is involved during the reaction.

9.5 Outlook and New Challenges

The concept of a classical force field together with suitable extensions to incorporate more details in the intermolecular interactions has proven to be a meaningful model for the investigation of complex systems. As briefly discussed here, it is essential that force fields are, however, applied with care due to the assumptions implicit to their parametrization. With this knowledge in mind, it will be possible to devise refined force fields including effects such as polarization or higher multipole moments, which make the model even more powerful for applications in structural biology and biophysics.

For transition states, their localization and significance in different fields of macromolecular simulations, the above examples have illustrated that they are rarely single, well-defined structures or states. Rather, for most applications in structural biology and biophysics, it is more appropriate to consider an ensemble of structures that divide the product from the reactant well. This is due to the large dimensionality of the problems, whereby many paths lead from the product to the reactant.

In the following, a number of potential future directions in the development and use of force fields for simulations relevant to structural biology are briefly described. This collection of thoughts is not intended to be an "Augur's view" of the future developments in atomistic simulations, but rather reflect what the author considers to be interesting topics to work on.

9.5.1 *Combined Coarse-Grained/Atomistic Simulations*

Multi-resolution techniques, where part of the system is represented by an atomistic force field and the rest by a more coarse-grained potential, will provide a significant step forward. Some progress has been made whereby the resolution of a group of atoms can be modified on-the-fly.[89] This means that, depending on the region in phase space, the atoms are treated with the atomistic force field or by coarse-grained potentials. The obvious advantage of such a procedure is the gain in computer time that can be achieved. However, care has to be taken that the resulting force field represents the true energetics and dynamics of the system as if it was simulated at full atomistic detail. This has been carried out for a combined atomistic/coarse-grained study of HIV-1 protease and human β-secretase.[90]

9.5.2 *Separation of Energetics and Nuclear Dynamics*

With free energy methods becoming computationally more and more feasible,[91] the determination of multidimensional free energy surfaces for protein folding, ligand binding, and other processes relevant to structural biology have become possible. This makes it possible to separate the problem of determining a realistic energy landscape on which the system evolves and to solve the problem for the nuclear dynamics, which often can be captured as a diffusion problem. Such procedures are required if the barriers between the different (meta)stable states are considerably larger than $k_B T$ and has been applied to ligand migration in myoglobin[17,81] or to protein folding,[92] which provides rates for barrier crossings from computing mean first passage times. Such studies, ideally carried out in full dimensionality and without bias such as umbrella sampling, provide the most direct contact between computing and actual experiment. Femtosecond time-resolved X-ray spectroscopy was instrumental to bridge the time gap between the computational work and experiment.[7,93] Exploring the ultra-fast time scales provides an ideal point of contact between simulation and experiment, and will allow computational strategies to be further refined.

9.5.3 *Application-Specific Force Fields and Predictive Simulations*

Most force fields for biomolecular simulations have been developed with the purpose to be as universal as possible for particular applications such as the simulation of proteins or DNA. As such, these force fields have proven to be reliable and robust under most circumstances. However, it is rare that quantitatively correct results can be gained from them without modifying particular parameters or even extending the mathematical form of the model. One example is the interaction of ligands with the protein. Relative binding free energies can be calculated from atomistic simulations including the surrounding solvent. However, at best such calculations give satisfactory correlation with experimental data and are not predictive.[94] There are two main reasons for this. First, the intermolecular interactions (e.g. point charges) used in the simulations are most likely not sufficiently accurate. For example, it has been found that the use of accurate charges, such as ones derived from quantum mechanical calculations, and environmental polarization effects can lead to improvements in docking studies.[95] Second, entropic contributions are not included in conventional ligand binding studies, although their effect is assessed in some cases and significant progress has been made recently.[96] It is also possible that mixed QM/MM methods can improve the predictive power of ligand-binding simulations.[97] However, whether the additional computational effort is justified in view of improved results is still a matter of debate.[98]

Other areas for which more detailed force field parametrizations should be considered is the investigation of nuclear magnetic resonance and the role of (electronically) excited states in biomolecular simulations. The latter in particular will be a challenge because the primary source of information — electronic structure calculations — is far less well developed for excited states than for electronic ground states.

One of the ultimate goals for computational structural biology in general and atomistic simulations in particular should be to provide a computational framework that allows *predictive simulations*. An important step towards this is the development of methods that give "the right answer for the right reason". Probably the best way to achieve this is through close collaboration between computational and experimental groups and by constructively challenging the others' findings and interpretations.

Acknowledgments

M. Meuwly acknowledges the contributions of the coworkers for their enthusiasm and their contributions to the work presented here. The studies were supported through generous financial support from the Schweizerischer Nationalfonds and the University of Basel. Computational Resources from the Centro Svizzero di Calcolo Scientifico (CSCS) are appreciated.

References

1. Levitt M, Lifson S. (1969) Refinement of protein conformations using a macromolecular energy minimization procedure. *J Mol Biol* **46**: 269–279.
2. McCammon JA, Gelin BR, Karplus M. (1977) Dynamics of folded proteins. *Nature* **267**: 585–590.
3. Frauenfelder H, Petsko GA, Tsernoglou D. (1979) Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature* **280**: 558–563.
4. Kern D, Zuiderweg ERP. (2003) The role of dynamics in allosteric regulation. *Curr Opin Struct Biol* **13**: 748–757.
5. Vendruscolo M, Dobson CM. (2006) Structural biology: dynamic visions of enzymatic reactions. *Science* **313**: 1586–1587.
6. Hammes-Schiffer S, Benkovic SJ. (2006) Relating protein motion to catalysis. *Ann Rev Biochem* **75**: 519–541.

7. Schotte F, Lim M, Jackson TA, *et al.* (2003) Watching a protein as it functions with 150 ps time-resolved X-ray crystallography. *Science* **300**: 1944–1947.
8. Hamm P, Lim M, Hochstrasser RM. (1998) Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy. *J Phys Chem B* **102**: 6123–6138.
9. Ponder JW, Case DA. (2003) *Adv Protein Chem* **66**: 27–85.
10. Mackerell AD. (2004) Empirical force fields for biological macromolecules: overview and issues. *J Comput Chem* **25**: 1584–1604.
11. Brooks BR, Bruccoleri RE, Olafson BD, *et al.* (1983) CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J Comput Chem* **4**: 187–217.
12. Weiner SJ, Kollman PA, Case DA, *et al.* (1984) A new force-field for molecular mechanical simulation of nucleic-acids and proteins. *J Am Chem Soc* **106**: 765–784.
13. Jorgensen WL, Tirado-Rives J. (1988) The OPLS potential functions for proteins — energy minimizations for crystals of cyclic-peptides and crambin. *J Am Chem Soc* **110**: 1657–1666.
14. Neria E, Fischer S, Karplus M. (1996) Simulation of activation free energies in molecular dynamics system. *J Chem Phys* **105**: 1902–1921.
15. Maple JR, Hwang MJ, Stockfisch TP, *et al.* (1994) Derivation of class-II force-fields. 1. Methodology and quantum force-field for the alkyl functional-group and alkane molecules. *J Comput Chem* **15**: 162–182.
16. Nutt DR, Meuwly M. (2003) Theoretical investigation of infrared spectra and pocket dynamics of photodissociated carbonmonoxy myoglobin. *Biophys J* **85**: 3612–3623.
17. Banushkina P, Meuwly M. (2005) Free-energy barriers in MbCO rebinding. *J Phys Chem B* **109**: 16911–16917.
18. Williams DE. (1988) Representation of the molecular electrostatic potential by atomic multipole and bond dipole models. *J Comput Chem* **9**: 745–763.
19. Stone AJ. (1981) Distributed multipole analysis, or how to describe a molecular charge-distribution. *Chem Phys Lett* **83**: 233–239.
20. Day PN, Jensen JH, Gordon MS, *et al.* (1996) An effective fragment method for modeling solvent effects in quantum mechanical calculations. *J Chem Phys* **105**: 1968–1986.
21. Plattner N, Meuwly M. (2008) The role of higher CO-multipole moments in understanding the dynamics of photodissociated carbonmonoxide in Myoglobin. *Biophys J*, **in print**.
22. Halgren TA, Damm W. (2001) Polarizable force fields. *Curr Opin Struct Biol* **11**: 236–242.
23. Rick SW, Stuart SJ. (2002) Potentials and algorithms for incorporating polarizability in computer simulations. *Rev Comput Chem* **18**: 89–146.

24. van Belle D, Froeyen M, Lippens G, Wodak SJ. (1992) Molecular-dynamics simulation of polarizable water by an extended Lagrangian method. *Mol Phys* **77**: 239–255.
25. Straatsma TP, McCammon JA. (1990) Free energy thermodynamic integrations in molecular dynamics simulations using a non-iterative method to include electronic polarization. *Chem Phys Lett* **167**: 252–254.
26. Rick SW, Berne BJ. (1996) Dynamical fluctuating charge force fields: the aqueous solvation of amides. *J Am Chem Soc* **118**: 672–679.
27. Stuart SJ, Berne BJ. (1996) Effects of polarizability on the hydration of the chloride ion. *J Phys Chem* **100**: 11934–11943.
28. Patel S, Mackerell Jr. AD, Brooks III CL. (2004) Charmm fluctuating charge force field for proteins: II — Protein/solvent properties from molecular dynamics simulations using a non-additive electrostatic model. *J Comput Chem* **25**: 1504–1514.
29. Warshel A, Levitt M. (1976) Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of carbonium-ion in reaction of lysozyme. *J Mol Biol* **103**: 227–249.
30. Alagona G, Ghio C, Kollman PA. (1986) Simple-model for the effect of Glu165->Asp165 mutation on the rate of catalysis in triose phosphate isomerase. *J Mol Biol* **191**: 23–27.
31. Bash PA, Field MJ, Karplus M. (1987) Free-energy perturbation method for chemical-reactions in the condensed phase: a dynamical-approach based on a combined quantum and molecular mechanics potential. *J Am Chem Soc* **109**: 8092–8094.
32. Claeyssens F, Ranaghan KE, Manby FR, Harvey JN, Mulholland AJ. (2005) Multiple high-level QM/MM reaction paths demonstrate transition-state stabilization in chorismate mutase: correlation of barrier height with transition-state stabilization. *Angew Chem Int Ed Engl* **40**: 5068–5070.
33. Zhou HY, Tajkhorshid E, Frauenheim T, Suhai S, Elstner M. (2002) Performance of the AM1, PM3, and SCC-DFTB methods in the study of conjugated schiff base molecules. *Chem Phys* **277**: 91–103.
34. Konig PH, Ghosh N, Hoffmann M, *et al.* (2006) Toward theoretical analyis of long-range proton transfer kinetics in biomolecular pumps. *J Phys Chem A* **110**: 548–563.
35. Altun A, Guallar V, Friesner RA, Shaik S, Thiel W. (2006) The effect of heme environment on the hydrogen abstraction reaction of camphor in P450(cam) catalysis: a QM/MM study. *J Am Chem Soc* **128**: 3924–3925.
36. Mei HS, Tuckerman ME, Sagnella DE, Klein ML. (1998) Quantum nuclear *ab initio* molecular dynamics study of water wires. *J Phys Chem B* **102**: 10446–10458.
37. Meuwly M, Karplus M. (2002) Simulation of proton transfer along ammonia wires: an *ab initio* and semi-empirical density functional comparison of potentials and classical molecular dynamics. *J Chem Phys* **116**: 2572–2585.

38. Cui Q, Elstner T, Karplus M. (2003) A theoretical analysis of the proton and hydride transfer in liver alcohol dehydrogenase (LADH). *J Phys Chem B* **106**: 2721–2740.
39. Meuwly M, Müller A, Leutwyler S. (2003) Energetics, dynamics, and infrared spectra of the DNA base-pair analogue 2-pyridone center dot 2-hydroxypyridine. *PCCP* **5**: 2663–2672.
40. Devi-Kesavan LS, Gao J. (2003) Combined QM/MM study of the mechanism and kinetic isotope effect of the nucleophilic substitution reaction in haloalkane dehalogenase. *J Am Chem Soc* **125**: 1532–1540.
41. Zoete V, Meuwly M. (2004) On the influence of semi-rigid environments on proton transfer along molecular chains. *J Chem Phys* **120**: 7085–7094.
42. Xu D, Wei Y, Wu J, *et al.* (2004) Qm/mm studies of the enzyme-catalyzed dechlorination of 4-chlorobenzoyl-coa provide insight into reaction energetics. *J Am Chem Soc* **126**: 13649–13658.
43. Sauer J, Döbler J. (2005) Gas-phase infrared spectrum of the protonated water dimer: molecular dynamics simulation and accuracy of the potential energy surface. *Comput Phys Commun* **6**: 1706–1710.
44. Warshel A, Weiss RM. (1980) An empirical valence bond approach for comparing reactions in solutions and in enzymes. *J Am Chem Soc* **102**: 6218–6226.
45. Grochowski P, Lesyng B, Bala P, McCammon JA. (1996) Density functional based parametrization of a valence bond method and its applications in quantum classical molecular dynamics simulations of enzymatic reactions. *Int J Quant Chem* **60**: 1143–1164.
46. Kim Y, Corchado JC, Villa J, Xing J, Truhlar DG. (2000) Multi-configuration molecular mechanics algorithm for potential energy surfaces of chemical reactions. *J Chem Phys* **112**: 2718–2735.
47. Li H, Elber R, Straub JE. (1993) Molecular-dynamics simulation of NO recombination to myoglobin mutants. *J Biol Chem* **268**: 17908–17916.
48. Zheng C, Makarov V, Wolynes PG. (1996) Statistical survey of transition states and conformational substates of the sperm whale myoglobin-CO reaction system. *J Am Chem Soc* **118**: 2818–2824.
49. Meuwly M, Becker OM, Stote R, Karplus M. (2002) NO rebinding to myoglobin: a reactive molecular dynamics study. *Biophys Chem* **98**: 183–207.
50. Nutt DR, Meuwly M. (2006) Studying reactive processes with classical dynamics: rebinding dynamics in MbNO. *Biophys J* **90**: 1191–1201.
51. Chandler D. (1978) Statistical-mechanics of isomerization dynamics in liquids and transition-state approximation. *J Chem Phys* **68**: 2959–2970.
52. Roux B, Karplus M. (1994) Molecular-dynamics simulations of the gramicidin channel. *Ann Rev Biophys Biomol Struct* **23**: 731–761.
53. Elber R, Chen DP, Rojewska D, Eisenberg R. (1995) Sodium in gramicidin — an example of a permion. *Biophys J* **68**: 906–924.

54. Northrup SH, Pear MR, Lee C.-Y, McCammon JA, Karplus M. (1982) Dynamical theory of activated processes in globular-proteins. *Proc Natl Acad Sci* **79**: 4035–4039.
55. Berkowitz M, Morgan JD, McCammon JA, Nothrup SH. (1983) Diffusion-controlled reactions — a variational formula for the optimum reaction coordinate. *J Chem Phys* **79**: 5563–5565.
56. Huo SH, Straub JE. (1997) The maxflux algorithm for calculating variationally optimized reaction paths for conformational transitions in many body systems at finite temperature. *J Chem Phys* **107**: 5000–5006.
57. Pratt LR. (1986) A statistical method for identifying transition states in high-dimensional problems. *J Chem Phys* **85**: 5045–5048.
58. Dellago C, Bolhuis PG, Csajka FS, Chandler D. (1998) Transition path sampling and the calculation of rate constants. *J Chem Phys* **108**: 1964–1977.
59. Huo S, Straub JE. (1999) Direct computation of long time processes in peptides and proteins: reaction path study of the coil-to-helix transition in polyalanine. *Proteins* **36**: 249–261.
60. Gillan MJ. (1987) Quantum classical crossover of the transition rate in the damped double well. *J Phys C* **20**: 3621–3641.
61. Bolhuis PG. (2005) Kinetic pathways of β-hairpin (un)folding in explicit solvent. *Biophys J* **88**: 50–61.
62. Basner JE, Schwartz SD. (2005) How enzyme dynamics helps catalyze a reaction in atomic detail: a transition path sampling study. *J Am Chem Soc* **127**: 13822–13831.
63. Crehuet R, Field MJ. (2007) A transition path sampling study of the reaction catalyzed by the enzyme chorismate mutase. *J Phys Chem B* **111**: 5708–5718.
64. Panchenko AR, Wang J, Nienhaus GU, Wolynes PG. (1995) Analysis of ligand-binding to heme-proteins using a fluctuating path description. *J Phys Chem* **99**: 9278–9282.
65. Henkelman G, Jonsson H. (2000) Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J Chem Phys* **113**: 9978–9985.
66. Faradjian AK, Elber R. (2004) Computing time scales from reaction coordinates by milestoning. *J Chem Phys* **120**: 10880–10889.
67. Ma A, Dinner AR. (2005) Automatic method for identifying reaction coordinates in complex systems. *J Phys Chem B* **109**: 6769–6779.
68. Branduardi D, Gervasio FL, Parrinello M. (2007) From A to B in free energy space. *J Chem Phys* **126**: 054103.
69. Juraszek J, Bolhuis PG. (2006) Sampling the multiple folding mechanisms of TRP-cage in explicit solvent. *Proc Natl Acad Sci* **43**: 15859–15864.
70. Sheinerman FB, Brooks III CL. (1998) Calculations on folding of segment b1 of streptococcal protein-G. *J Mol Biol* **278**: 439–455.

71. Garcia AE, Onuchic JN. (2003) Folding a protein in a computer: an atomic description of the folding/unfolding of protein A. *Proc Natl Acad Sci* **100**: 13898–13904.
72. Frauenfelder H, Fenimore PW, McMahon BH. (2002) Hydration, slaving and protein function. *Biophys Chem* **98**: 35–48.
73. Frauenfelder H, McMahon BJ, Austin RH, Chu K, Groves JT. (2001) The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proc Natl Acad Sci* **98**: 2370–2374.
74. Elber R, Karplus M. (1987) Multiple conformational states of proteins — a molecular-dynamics analysis of myoglobin. *Science* **235**: 318–321.
75. Kottalam J, Case DA. (1988) Dynamics of ligand escape from the heme pocket of myoglobin. *J Am Chem Soc* **110**: 7690–7697.
76. Straub JE, Karplus M. (1991) Molecular-dynamics study of the photodissociation of carbon-monoxide from myoglobin — ligand dynamics in the first 10 ps. *Chem Phys* **158**: 221–248.
77. Sagnella DE, Straub JE. (1999) A study of vibrational relaxation of B-state carbon monoxide in the heme pocket of photolyzed carboxymyoglobin. *Biophys J* **77**: 70–84.
78. Nutt DR, Meuwly M. (2004) Migration in native and mutant myoglobin: atomistic simulations for the understanding of protein function. *Proc Natl Acad Sci* **101**: 5998–6002.
79. Agmon N, Hopfield JJ. (1983) CO binding to heme-proteins — a model for barrier height distributions and slow conformational-changes. *J Chem Phys* **79**: 2042–2053.
80. Steinbach PJ, Ansari A, Berendzen J, *et al.* (1991) Ligand-binding to heme-proteins — connection between dynamics and function. *Biochemistry* **30**: 3988–4001.
81. Banushkina P, Meuwly M. (2007) Diffusive dynamics on multidimensional rough free energy surfaces. *J Chem Phys* **127**(13): 135101.
82. Banushkina P, Meuwly M. (2005) Hierarchical numerical solution of the Smoluchowski equation with smooth and rough potentials. *J Chem Theoret Comput* **1**: 208.
83. Bossa C, Anselmi M, Roccatano D, *et al.* (2004) Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin. *Biophys J* **86**: 3855–3862.
84. Cohen J, Arkhipov A, Braun R, Schulten K. (2006) Imaging the migration pathways for O_2, CO, NO, and Xe inside myoglobin *Biophys J* **91**: 1844–1857.
85. Petrich JW, Lambry J.-C, Kuczera K, *et al.* (1991) Ligand binding and protein relaxation in heme proteins: a room temperature analysis of NO geminate recombination. *Biochemistry* **30**: 3975.

86. Kim S, Jin G, Lim M. (2004) Dynamics of geminate recombination of NO with myoglobin in aqueous solution probed by femtosecond mid-IR spectroscopy. *J Phys Chem B* **108**: 20366–20375.
87. Kim S, Lim M. (2005) Protein conformation-induced modulation of ligand binding kinetics: a femtosecond mid-IR study of nitric oxide binding trajectories in myoglobin. *J Am Chem Soc* **127**: 8908–8909.
88. Hur S, Bruice TC. (2003) The near attack conformation approach to the study of the chorismate to prephenate reaction. *Proc Natl Acad Sci* **100**: 12015–12020.
89. Praprotnik M, Delle SL, Kremer K. (2007) A macromolecule in a solvent: adaptive resolution molecular dynamics simulation. *J Chem Phys* **126**: 134902.
90. Neri M, Anselmi C, Cascella M, Maritan A, Carloni P. (2005) Coarse-grained model of proteins incorporating atomistic detail of the active site. *Phys Rev Lett* **95**: 218102.
91. Simonson T, Archontis G, Karplus M. (2002) Free energy simulations come of age: protein-ligand recognition. *Acc Chem Res* **35**: 430–437.
92. Singhal N, Snow CD, Pande VS. (2004) Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J Chem Phys* **121**: 415–425.
93. Srajer V, Ren Z, Teng TY, *et al.* (2001) Protein conformational relaxation and ligand migration in myoglobin: a nanosecond to millisecond molecular movie from time-resolved laue X-ray diffraction. *Biochem* **40**: 13802.
94. Thorsteinsdottir H, Zoete V, Schwede T, Meuwly M. (2006) How inaccuracies in protein structure models affect estimates of protein-ligand interactions: computational analysis of HIV-1 protease inhibitor binding. *Protein Struct Funct Genom* **65**: 407–423.
95. Friesner RA, Guallar V. (2005) *Ab initio* quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Ann Rev Phys Chem* **56**: 389–427.
96. Wang JY, Deng YQ, Roux B. (2006) Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys J* **91**: 2798–2814.
97. Raha K, Merz KM. (2005) Large-scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J Med Chem* **48**: 4558–4575.
98. Peters MB, Raha K, Merz KM. (2006) Quantum mechanics in structure-based drug design. *Curr Opin Drug Discov Devel* **9**: 370–379.
99. Bolhuis PG. (2003) Transition-path sampling of β-hairpin folding. *Proc Natl Acad Sci* **100**: 12129–12134.

Chapter 10

Functional Motions in Biomolecules: Insights from Computational Studies at Multiple Scales

A. W. van Wynsberghe[†], L. Ma[†], X. Chen[‡] and Q. Cui[*,†]

10.1 Introduction

Motions at both the domain and local scales are important to the function of biomolecules. In this chapter, we discuss computational techniques for probing these functional motions. These include atomistic simulations that characterize the energetics of local motions, various normal-mode based methods that capture the directionality of domain scale motions, as well as effective coarse-grained methods necessary for probing motions at very large length and time scales. The values and limitations of these techniques are illustrated by selected applications used to analyze the role of local motions in enzyme catalysis, mechanochemical coupling in signaling proteins and biomolecular motors, and gating of the mechanosensitive channel. A number of outstanding and emerging questions regarding functional motions in biomolecular systems are briefly discussed.

[*] Corresponding author.

[†] Department of Chemistry and Theoretical Chemistry Institute, University of Wisconsin, Madison, 1101 University Ave, Madison, WI 53706. Email: cui@chem.wisc.edu.

[‡] Department of Civil Engineering and Engineering Mechanics, Columbia University, New York City, NY 10027.

Mounting evidence from numerous experimental[1,2] and computational studies[3] has demonstrated that biomolecules have motions that span a wide range of time and spatial scales. Some of those motions reflect the importance of maintaining a "minimal" level of flexibility for function. For example, a recent insightful analysis[3] examined the magnitude of atomic fluctuations in proteins using data from both molecular dynamics simulations and crystallographic Debye-Waller factors. Based on the Lindemann criterion, atomic fluctuations indicate that the surface of proteins is liquid-like while the core is solid-like. This result makes intuitive sense in that the solid core is important for stability while the fluidic surface is essential for the structural changes required by basic functions such as ligand binding. As the temperature approaches the so-called "glass-transition" temperature (~180 K for many proteins), the Lindemann criterion suggests that the entire protein becomes solid-like; at the same temperature, most proteins lose their ability to function.

In addition to such "generic" thermal fluctuations, it is generally agreed that there are also "functional motions", which have specific characters (in direction, magnitude, and time-scale) that make these motions essential to the unique function of a particular biomolecule. These range from structural transitions at the domain scale, which are implicated in the function of many "biomolecular machines"[5] and multi-subunit enzymes,[6] to relatively localized vibrations that have been proposed to facilitate chemical reactions.[7] In this regard, we note that a rather broad notion of "motion" is adopted here, which includes both equilibrium fluctuations in a single state and structural transitions between two (or more) distinct functional states of a system.

Despite their biological importance, functional motions are difficult to identify and characterize at a quantitative level. The multiple length and time scales spanned by these motions pose tremendous challenges to experimental measurements and their interpretation. A significant body of studies has demonstrated that careful computational studies can nicely complement experimental work for better characterizing and understanding the working mechanism of functional motions. In the following, we first briefly review several computational methods that are particularly useful for studying motions in biomolecules at multiple scales; then, we discuss a few examples from our labs to illustrate the value and limitation of these techniques as well as the mechanistic insights derived from computational analyses regarding the nature and functional implication of specific motions in the corresponding systems. Finally, a number of outstanding and emerging questions regarding functional motions in biomolecular systems are briefly discussed.

10.2 Computational Methods

In principle, the most robust computational approach for studying motions in biomolecules is atomistic molecular dynamics (MD).[8] Ever since the first molecular dynamics simulations of proteins 30 years ago,[9] striking progresses have been made in both theoretical/computational algorithms and computational hardware. As a result, sophisticated molecular dynamics simulations have become an indispensable tool in the analysis of structural, energetic, and dynamical properties of biomolecules.[3,8] Nevertheless, for many processes, such as domain motions, atomistic molecular simulations are still too expensive for obtaining statistically meaningful results. Even for relatively local structural transitions, it is challenging to quantify the underlying thermodynamics and kinetics using straightforward molecular dynamics simulations. In those cases, alternative computational approaches have to be used. In the following, we briefly summarize a few computational approaches that are useful for characterizing motions at different scales and evaluating the functional significance of these motions.

10.2.1 *Local Motions: Advanced Atomistic Molecular Dynamics*

Local motions such as side-chain flips, loop displacements and break-formation of salt-bridge interactions play important roles in many systems. For example, isomerization of a histidine from a buried configuration to a solvent exposed orientation is implicated in its proton shuttling function in carbonic anhydrase;[10] the closure of a "lid" composed of an 11-residue loop sets up the active site of triosephosphate isomerase to avoid side reactions;[11] the formation of a critical salt-bridge between two loop motifs in myosin helps to align water molecules properly in the nucleotide binding sites for the subsequent hydrolysis of ATP.[12,13] Due to the presence of free energy barriers higher than k_BT, characterizing the corresponding thermodynamics and kinetics (barriers) for local structural changes is not always straightforward. In principle, these quantities can be estimated from the relevant potential of mean force (PMF, W(ξ))[14] profile [Fig. 10.1(a)] using umbrella samplings[15] to obtain

$$W(\xi) = -k_BT \ln P(\xi) + C \tag{10.1}$$

where ξ is the chosen reaction coordinate, $P(\xi)$ is the probability distribution along ξ, and C is a normalization constant.

In practice, however, even localized structural changes may implicate variations in a handful geometrical parameters and it can be difficult to identify the most important one(s) as the principal "reaction coordinate(s)"; computing PMF along an inappropriate reaction coordinate may lead to significant error in the computed energetics, especially barriers.[16] The situation can be even more complex if there is significant involvement of the solvent degrees of freedom; this might be more prevalent than one may naively assume, and even the isomerization of an alanine dipeptide, for example, has been shown to implicate significant solvent participation.[17] Another example along this line is the dimeric hemoglobin in scallop, where a change in the number of interfacial water molecules is coupled to the rotation of a phenylalanine residue at the dimer interface and is key to the allosteric communication between the two subunits.[18]

In other words, a major challenge for quantifying local motions (including chemical reactions) is the identification of variables whose changes best describe the kinetic bottleneck of the process. In some applications, experience and intuition can be very instructive (see the following discussion on CheY). Nonetheless, a less *ad hoc* approach is highly desirable. In this context, the transition path sampling (TPS) technique proposed by Chandler and co-workers[19] provides a theoretically sound framework for studying reactive processes (either chemical or structural) in complex systems like biomolecules. Unlike the minimum energy path analysis, which is powerful for gas-phase processes but significantly less appropriate for processes in the condensed phase, TPS collects real-time "reactive trajectories," and therefore, samples the true kinetic bottleneck and includes entropic effects [Fig. 10.1(b)]. As described in details in Ref. 20, TPS employs a Monte Carlo procedure to sample the trajectory space with emphasis on reactive trajectories that lead to the structural transitions of interest. This is possible because most local structural transitions are thermally activated, meaning that the rate is low due to significant (free) energy barriers, but the barrier crossing process itself, once activated, is fast (often in the picosecond regime).[21]

Briefly, a TPS simulation starts with a single reactive trajectory, which can be obtained in a number of ways such as by forcing the relevant structural transition to occur via artificial restraints and then gradually reducing the strength of the restraints in a series of "annealing" simulations.[22] Then, new trajectories derived by slightly perturbing the existing trajectory (e.g. by modifying the velocity of certain atoms in a frame) will be generated and accepted based on a Metropolis criterion to ensure detailed balance (i.e. the proper canonical weights of trajectories). This is carried out iteratively until a significant number of uncorrelated reactive trajectories have been collected; the precise number depends on the system and the goal of an application.

Clearly, the TPS approach is computationally intense and typically involves at least collecting thousands of short trajectories on the order of 10–100 ps. Therefore, TPS is ideally suited for studying relatively local structural transitions in biomolecules where the process can be too complex to characterize with a few "obvious" choices of variables, but the intrinsic transition time-scale is still well

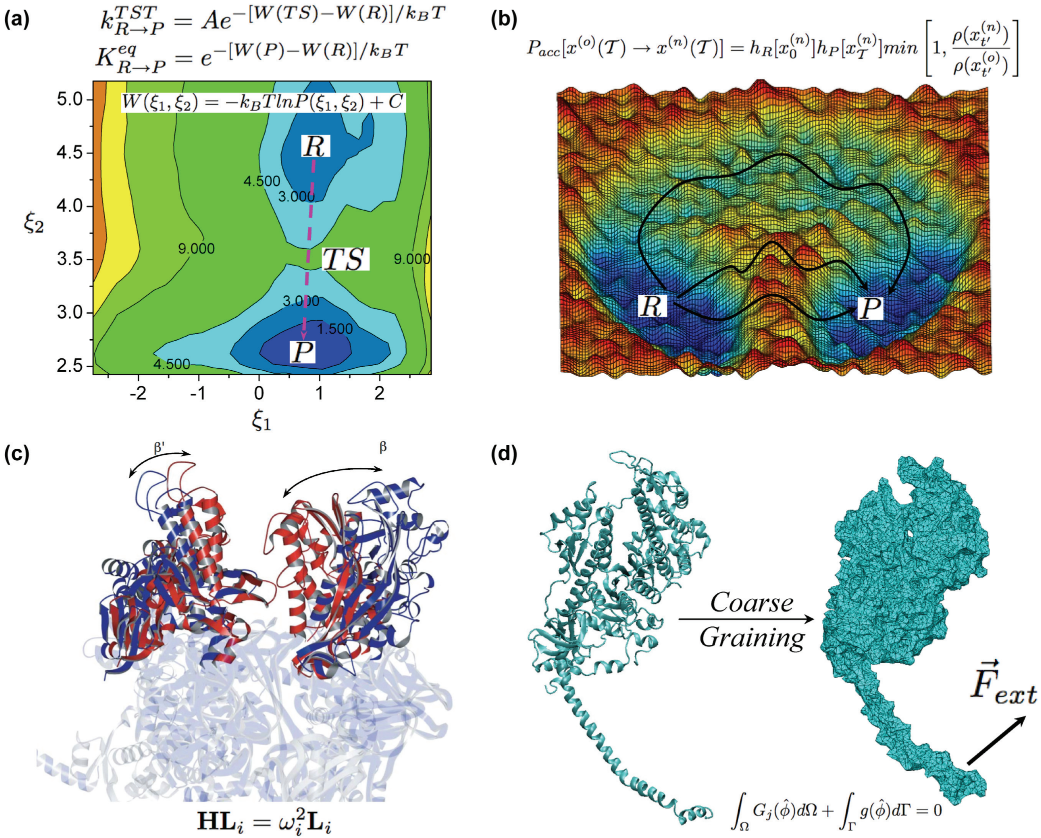

(a)
$k_{R\rightarrow P}^{TST} = Ae^{-[W(TS)-W(R)]/k_BT}$
$K_{R\rightarrow P}^{eq} = e^{-[W(P)-W(R)]/k_BT}$
$W(\xi_1,\xi_2) = -k_BTlnP(\xi_1,\xi_2)+C$
R
TS
P
4.500
3.000
9.000
9.000
3.000
1.500
4.500
5.0
4.5
4.0
3.5
3.0
2.5
ξ_2
-2
-1
0
1
2
ξ_1
(b)
$P_{acc}[x^{(o)}(\mathcal{T}) \rightarrow x^{(n)}(\mathcal{T})] = h_R[x_0^{(n)}]h_P[x_{\mathcal{T}}^{(n)}]min\left[1, \frac{\rho(x_{t'}^{(n)})}{\rho(x_{t'}^{(o)})}\right]$
R
P
(c)
β'
β
$\mathbf{HL}_i = \omega_i^2\mathbf{L}_i$
(d)
Coarse
Graining
$\vec{F}_{ext}$
$\int_\Omega G_j(\hat{\phi})d\Omega + \int_\Gamma g(\hat{\phi})d\Gamma = 0$

Fig. 10.1. (See caption on next page.)

in the sub-nanosecond regime. It is important to recognize, however, that the standard TPS protocol is still a local sampling technique, and therefore, the results are likely dependent on the initial "reactive trajectory" (see example below).[23] Other challenges associated with TPS include analyzing the results,[24] dealing with processes of diffusive nature[25] and/or involving intermediate(s); given the limited space here, we refer readers to recent discussions in the literature.

10.2.2 *Domain Motions: Normal Mode Analysis*

Large-scale structural transitions at the domain scale are involved in many "biomolecular machines" such as molecular motors[5] and allosteric multi-subunit enzymes.[2] They are difficult to study using regular atomistic simulations because they occur at time scales typically at or longer than *ms.* Various "unconventional" molecular dynamics techniques have been proposed accordingly, which either applies specific biasing potentials to artificially speed up the structural transitions[26,27] or aims at identifying the approximate transition path(s) between two functional states.[28]

One interesting alternative that has found great popularity in recent years is the normal mode analysis (NMA).[29] In NMA, one approximates the motion of the system as harmonic vibrations around

Fig. 10.1 Illustration of the computational methods discussed in this chapter. **(a)** Potential of mean force (PMF) simulations along chosen reaction coordinates ($\xi 1$, $\xi 2$), from which the kinetics ($k_{R\rightarrow P}$) and thermodynamics ($k^{eq}_{R\rightarrow P}$) of the relevant transitions can be estimated, are most useful for probing local motions; **(b)** Transition Path Sampling (TPS), which probes reactive (either chemical or conformational) trajectories between different states with proper weights, does not require the pre-selection of reaction coordinates and is suitable for studying complex local structural transitions in systems with rough energy landscape; **(c)** Normal Mode Analysis (NMA) is an approximate method well-suited for describing collective motions at the domain scale; **(d)** Further coarse-grained methods such as those based on continuum mechanics can, if parameterized carefully, probe functional motions at very large length and time scales. All structural figures are made using VMD.[133]

a local minimum on the potential energy surface. Following the diagonalization of the force constant (second-order derivative, or the hessian, $\mathbf{H}$) matrix in mass-weighted coordinates,

$$\mathbf{H}\mathbf{L}_i = \omega_i^2 \mathbf{L}_i, \quad i = 1 \ldots 3N - 6 \tag{10.2}$$

the equations of motion can then be simplified as a set of uncoupled harmonic oscillators of frequencies $\{\omega_i\}$; here N is the total number of atoms. Through the eigenvectors, $\mathbf{L}_i$, the time evolution of the Cartesian coordinates ($\mathbf{L}_i(t)$) can be expressed analytically at all time,

$$q_j^\alpha(t) = m_j^{-1/2} \sum_{i=1}^{3N-6} [A_i \cos(\omega_i t + \phi_i)] L_{ji}^\alpha, \ j = 1 \ldots N; \ \alpha = x, y, z \tag{10.3}$$

where A_i, ϕ_i are the amplitude and phase factor for the ith mode; this allows the calculation of many thermally averaged results such as atomic fluctuations at a given temperature T.[14]

Although clearly approximate, a significant body of research has demonstrated that NMA is uniquely useful for characterizing collective motions in biomolecules [Fig. 10.1(c)].[30–33] In particular, large-scale structural transitions between different functional states have been found to correlate very well with the low-frequency normal modes; in fact, in many cases, a large fraction of the structural transitions can be expressed as the linear combination of a very small number of low-frequency normal modes. This leads to the idea that the flexibility required for the functional transitions is an inherent feature of the system encoded by the structure.

The fact that low-frequency modes are most relevant for characterizing domain-scale motions suggests that further numerical[136] as well as physical approximations can be made to NMA such that the efficiency of the computation can be improved. One idea is to divide the system into a set of "blocks" (e.g. one amino acid per block), and then ignore the internal motion of the blocks when solving the NMA problem;[34] this significantly reduces the size of the eigenvalue problem.[35] Such a "block normal mode" approach has been shown to give very reliable results for low-frequency eigenvalues/eigenvectors, and therefore, can be used to explore structural flexibilities of very large biomolecular complexes (such as protein-DNA complexes and the ribosome) with atomistic interactions.[36,37]

A further approximation is to simplify the interaction potential into that of a set of elastic springs, which leads to the "elastic network model (ENM)".[38]

$$U^{elastic} = \gamma \sum_{i \neq j} \Theta(r_{ij}^{0} - r_{cut})(r_{ij} - r_{ij}^{0})^{2} \tag{10.4}$$

where Θ is a Heaviside step function, r_{cut} a parameter that determines the range of interactions, and γ a scaling factor; r_{ij}^{0} is the distance between atoms i and j in the current structure. In addition to its computational efficiency, a nice feature of ENM is that the potential ensures that the current structure is the energy minimum, and therefore, no energy minimization is needed. A large body of studies have shown that despite its simplicity, ENM produces reliable results for the low-frequency eigenvectors for compact structures[32–33,39] although the results tend to deteriorate as the frequency increases.[40] In a recent analysis,[41] for example, results from several variations of ENM and the block normal mode (BNM) using an atomistic potential were compared to the anisotropic displacement parameters (ADPs) from high-resolution X-ray structures. It was found that most methods produce favorable agreement with the experiment ADPs, although there are notable differences between the eigenvectors from ENM and BNM calculations, except for the first few modes. For very large systems, reliable ENM results can be obtained with a significantly smaller number of interaction sites than the number of atoms. This leads to the impressive application of ENM to systems with low-resolution structural information (such as electron microscopy maps).[42,43]

It is important to emphasize that by "characterizing" domain motions with normal modes, we mean that the *directions* of large-scale flexibility correlate well with a small number of low-frequency normal modes. The time-scale, magnitude, and energetics of motions along these directions, are however, beyond the capability of the NMA approaches discussed above (see Section 4). Moreover, care has to be exercised when interpreting "correlated motions" in biomolecules using only a small number of modes.[44]

10.2.3 *Coarse-Grained Models Beyond the Harmonic Limit*

Another active area of research involves developing effective coarse-grained models such that anharmonic motions (thus beyond normal modes) at long-time scale and large spatial scale can be effectively simulated and analyzed. The most popular approach is to reduce the resolution of the representation by grouping several atoms into a single bead and then parameterize effective interactions between the beads.[45–47] Different approaches have been proposed for parameterizing effective interactions based on atomistic simulations, and it remains a challenging task to develop potentials that are accurate, flexible, and transferable at the coarse-grained level.[48] Nevertheless, impressive results have been obtained for a number of systems, perhaps most notably for lipid systems by Marrink and co-workers.[49]

An alternative direction that we have been exploring recently is to adopt a continuum mechanics framework with the finite element (FEM) representation. Although continuum mechanics have been used in the past to model the mechanical behavior of biomolecules,[50,51] they are usually based on highly idealized geometries and materials properties. The FEM analysis,[52] on the other hand, is widely used in the engineering field for solving mechanical and transport problems and can be applied to systems with complex geometries and boundary conditions.

In our recent study, we have established a proof-of-concept continuum mechanics model for the mechanosensitive channel of the large conductance (MscL).[53] Inspired by its crystal structure,[54] this model treats transmembrane helices of MscL as elastic rods and the lipid membrane as an elastic sheet of finite thickness; in the more recent model,[55] periplasmic loops are also included as elastic springs. In the FEM framework, these continuum components (indicated as domain Ω), are represented by a set of "elements" in the shape of, for example, tetrahedra [Fig. 10.1(d)]. The size of the elements can be determined adaptively, small for regions of interest and large for far-away areas, which makes the simulation framework ideal for very large systems such as a protein complex embedded in a large sheet of membrane. Similar to atoms in particle-based simulations, each element is associated with the materials' properties (e.g. Young's moduli) and parameters that describe inter-element interactions. These important parameters can be derived from calculations using all-atom force fields. The interaction between the continuum components and the surrounding solvent can be treated at the Poisson-Boltzmann level.[56] Currently, we are developing, in a systematic manner, a semi-quantitative framework that treats irregular shapes of continuum components and employs more sophisticated description of materials properties. The solution of an unknown variable (function), ϕ, is then approximated by a series of shape functions, s_i, and a set of unknown parameters, a_i (e.g. nodal displacements), as $\phi \approx \hat{\phi} = \sum_i a_i s_i$. The values of a_i are then determined from equations established from, for example, a variational principle

$$\int_\Omega G(\hat{\phi})d\Omega + \int_\Gamma g(\hat{\phi})d\Gamma = 0 \qquad (10.5)$$

where Γ is the domain boundary and g, G are the relevant energy/work functionals.

Since most interactions in a FEM model are local in nature, the cost of the simulation is modest. Therefore, once parameterized, the continuum mechanics model is ideally suited for studying the structural response of the biomolecule to various external mechanical perturbations of different form and scale. In the simplest application, this involves applying mechanical loads as the boundary condition and evolves the structure of the continuum components (i.e. positions of the FEM nodes) in a quasi-static fashion. Even at this level, interesting insights can be obtained. For example, *qualitatively* different responses of MscL were observed when the membrane was subject to in-plane tension versus out-of-plane bending (see below);[53]

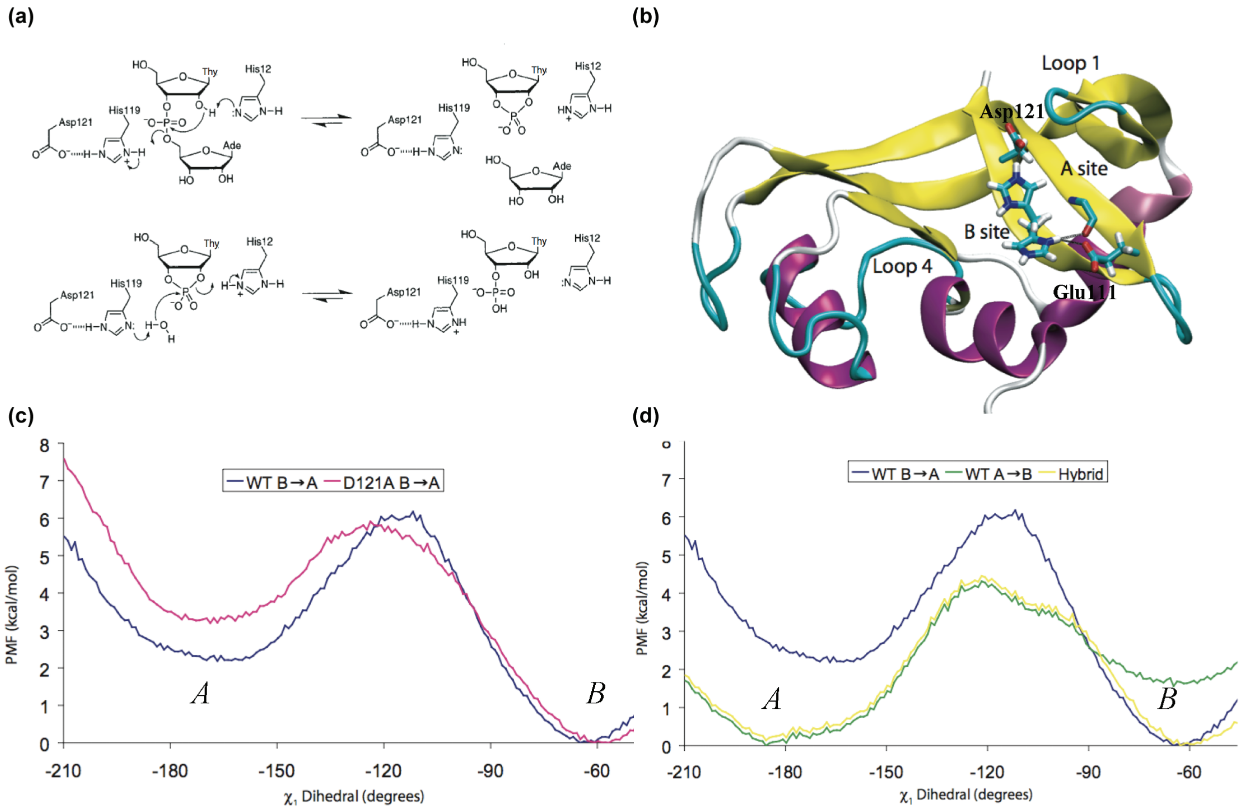


Fig. 10.2. (See caption on next page.)

this is very difficult to achieve with any other widely available simulation techniques. At a more sophisticated level, the real-time dynamics of the continuum system at the finite temperature can be monitored by propagating Langevin dynamics; the potential of such studies on biomolecular systems, however, remains to be fully developed and explored.

10.3 Applications

In this section, we discuss a few applications from our own studies to illustrate how the techniques outlined above are used in providing useful insights into functional motions in biomolecules at multiple scales.

10.3.1 *Functional "Dynamics" of Ribonuclease A*

Whether there are specific motions (or loosely referred to as "dynamics" in the relevant literature) in enzymes that facilitate the catalytic step has been a topic of intense interest. An intriguing recent example is ribonuclease A, for which the motion-catalysis relationship has been analyzed in details by Loria and co-workers using NMR relaxation measurements.[57,58] The most interesting finding concerns the effect of mutating Asp121 near the active site [Fig. 10.2(a)] to an alanine. More than 95% of the activity is lost upon mutation, although neither substrate affinity nor the electrostatic properties of the active site (e.g. as reflected by the catalytic His residues) was significantly perturbed.[58–60] There was, however, interesting changes in the μs-ms motions upon mutation. In the wild type (WT) enzyme, the motions of different motifs (e.g. loop 1, 4, and His119) are very close in time-scale

Fig. 10.2 Study of functional motions in RNase A. **(a)** The basic catalytic cycle; **(b)** The active site of RNase A with both the His119 A and B site rotamers present; the hydrogen bonding in both the A site (Asp121) and the B site (Ala109 main chain and Glu111) is indicated. **(c)–(d)** Potential of mean forces from different calculations along the His119 χ_1 angle that characterizes the transition between the A and B sites.

to each other and to the observed catalytic rate; in the D121A mutant, by contrast, the time-scale for the motions of different structural motifs becomes substantially different, and product release (the rate limiting step in the wild type) actually became faster. These observations led Loria and coworkers to suggest that the synchronicity of global protein motions plays an important role in determining the rates of catalytically important steps, and the loss of catalysis in the D121A mutant is from the disruption of these global dynamics.[58]

10.3.1.1 *Atomic scale hypothesis for D121A effects*

Although the hypothesis that the global millisecond dynamics of RNase A are "coordinated" and "timed" to help catalysis occur is intriguing, it is nonetheless difficult to imagine a detailed atomic level mechanism. We present an alternative hypothesis for the decrease in catalytic rate that involves changes in the free energy landscape of the *apo* enzyme and then attempt to verify this hypothesis using PMF simulations.

The catalytically active His119's side-chain has been observed to exist in two configurations in both crystallographic and NMR experiments.[58,59,62] These two configurations are defined by the *trans* (~180°) or *gauche*⁺ (~–60°) rotamers of the side-chain χ_1 dihedral angle and are known as the "A" and "B" sites respectively [Fig. 10.2(b)]. The enzyme is only active when His119 is in the A site and A/B site conformational exchange is very unlikely if the substrate is bound. If one assumes that the substrate can bind to RNase A when it is in the B conformation, this complex would be unreactive. The kinetic scheme for such a situation, where the enzyme can transition between an active and inactive form with both having the ability to bind the substrate is given by

$$E_I \bullet S \underset{k_{-4}}{\overset{k_4}{\leftrightarrow}} E_I + S \underset{k_{-3}}{\overset{k_3}{\leftrightarrow}} E_A + S \underset{k_{-1}}{\overset{k_1}{\leftrightarrow}} E_A \bullet S \overset{k_2}{\rightarrow} E_A + P \tag{10.6}$$

Analyzing this scheme with the steady state approximation and the assumption that the inactive and active forms bind the substrate with similar affinities results in a modified Michaelis-Menten equation where the apparent catalytic rate, k_2, is multiplied by the fraction of active enzyme, f_A,

$$u = \frac{f_A k_2 [E_T][S]}{K_M + [S]} \tag{10.7}$$

If the assumptions of the above analysis are correct, one can observe a change in the *apparent* catalytic rate, $f_A k_2$, simply by changing the relative populations of the A and B sites. Therefore, our hypothesis for the decrease in catalytic activity upon the D121A mutation is that the A site His119 conformer is destabilized relative to the B site conformer, which leads to a smaller fraction of the *apo* enzyme being in a catalytically active state at any one time.

10.3.1.2 *Simulation studies*

To test our hypothesis, umbrella sampling is used to calculate potentials of mean force (PMFs) along the His119 χ_1 dihedral, which defines the A or B site position, for both the D121A mutant and the wild type enzyme. As a check of convergence, two independent sets of PMF simulations are carried out for each system, starting from either a *trans* (A site) or a *gauche*$^+$ (B site) His119 χ_1 dihedral. Both PMF simulations consist of approximately 2–3 ns of production sampling and the results are referred to as the A → B PMF and the B → A PMF respectively.

As shown by the B → A PMF in Fig. 10.2(c), the A site is destabilized by approximately 1 kcal/mol more with respect to the B site in the mutant than in the wild type. This suggests that the B site is more populated in the mutant enzyme than in the wild type. The A → B simulations show this same trend, although they produce different relative stabilities of the two sites. For the WT enzyme, for example, whereas the B → A PMF predicts that the B site is more stable, the A → B PMF predicts that the A site is more stable. In both cases, however, the $\Delta\Delta G_{B\text{-}A}$ corresponding to $\Delta G_{B\text{–}A}^{D121A} - \Delta G_{B\text{–}A}^{WT}$ is negative, suggesting that the D121A mutation destabilizes the A site with respect to the B site, supporting our hypothesis.

Careful analysis of the simulation trajectories suggests that this hysteresis arises because different His119 χ_2 angles are sampled in the B → A and A → B PMF simulations, especially in regions near site B. To test this, a "hybrid" A/B PMF calculation is set up, which uses the A site and transition barrier windows from the A → B simulations; for windows near the B site, the starting configurations are generated from an A site equilibrium snapshot, which are then constrained during equilibration to give consistent χ_2's. As shown in Fig. 10.2(d), the A site ($-210 \le \chi_1 \le -165$) and transition barrier ($-150 \le \chi_1 \le -105$) regions of this hybrid PMF are exactly (within a small constant) the same as in the A → B PMF, but with the new B site windows ($-90 \le \chi_1 \le -45$), the resulting PMF shows a stabilization of the B site, much like the B → A simulations. In fact, the shape of the hybrid PMF in the A site region matches nearly perfectly with the B → A PMF.

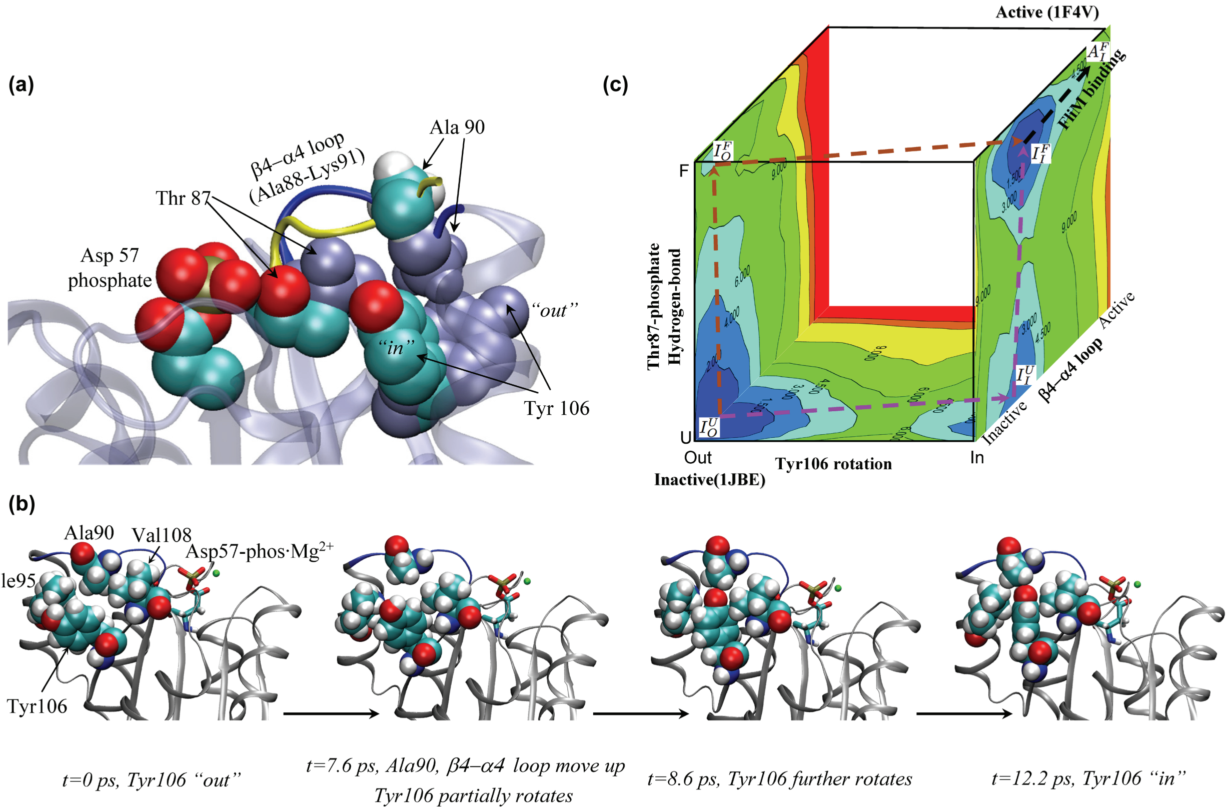
In short, although there remain important hysteresis problems for the computed χ_1 PMFs due to the lack of sufficient sampling in χ_2 (which illustrates the subtlety of such simulations for even a *local* event!), the hypothesis that the D121A mutant's loss of activity came from a shift in population of the His119 from the A site to the B site is feasible. In both the A → B and B → A simulations, a *destabilization* of the A site relative to the B site upon mutation was observed. That is, although $\Delta G_{B\text{–}A}^{D121A}$ and $\Delta G_{B\text{–}A}^{WT}$ change signs between the A → B and B → A simulations, the $\Delta\Delta G_{B\text{-}A}$ stays negative for both cases. The more rigorous way to characterize the relevant energetics associated with the His119 isomerization is to compute a two-dimensional PMF along both the χ_2 and χ_1 coordinates; this is underway.

10.3.2 *Activation of a Signaling Protein: CheY*

Signaling proteins are activated to perform their biological function through a localized event such as phosphorylation or ligand (ion) binding.[63] Understanding how such local modifications lead to striking transitions in the structure, and therefore, activity of signaling proteins is evidently of great value from both fundamental and biomedical perspectives. Recent NMR studies[64] of small signaling proteins in two-component systems suggested that the structural motifs to be activated have a small but non-negligible population in the active conformation prior to phosphorylation; the role of phosphorylation is to shift this population to become the dominant one rather than inducing new conformations. Such a "population shift" framework,[65] which has features of the Monod-Wyman-Changeux (MWC) model[66] for allostery,[66–69] emphasizes the dynamical nature of signaling proteins (and allosteric systems in general) and provides a rather different picture from the "push and pull" type of description as characterized by the stereochemical model for hemoglobin.[70] To fully understand the activation mechanism, however, it is important to characterize the energetics of the relevant motion and reveal how the energetics are modulated by the activation event (i.e. *how* "population shift" is induced).

10.3.2.1 *"Y-T" Coupling versus population shift*

CheY is a 129-residue prototypical response regulator in a two-component signal transduction system.[71] It is activated through phosphorylation, and the most important conformational change in CheY upon activation is the rotation of the Tyr106 side-chain from a solvent exposed orientation to a fully buried state [Fig. 10.3(a)]. The distance between Tyr106 and the phosphorylation site (Asp57) is more than 9.5 Å, which makes CheY a prototypical single-domain protein that exhibits allosteric behavior. The highly conserved Thr87 spatially separates Asp57 and Tyr106, thus the conventional description for CheY activation is the "Y-T coupling" model:[72] phosphorylation of Asp57 displaces Thr87 due to a hydrogen-bonding interaction, which in turn allows the rotation of Tyr106. Since partial activity has been observed for the wild type CheY[73] and the T87A mutant[74] in the absence and presence of phosphorylation, respectively, the "Y-T coupling" model has been questioned. In particular, since the $\beta 4$–$\alpha 4$ loop (Ala88 to Lys91) also undergoes a major displacement upon activation (root-mean-square-deviation for the backbone and all non-hydrogen atoms is 1.9 and 3.6 Å respectively),[75,76] it has been speculated[75,77] that this loop in fact gates the rotation of Tyr106 and the role of phosphorylation, and



t=0 ps, Tyr106 "out"

t=7.6 ps, Ala90, β4–α4 loop move up Tyr106 partially rotates

t=8.6 ps, Tyr106 further rotates

t=12.2 ps, Tyr106 "in"

Fig. 10.3. (See caption on next page.)

Thr87 is to select specific loop configuration, which is reminiscent of the "population shift" model.

10.3.2.2 *Simulation studies*

To gain further mechanistic insights into the activation of CheY as an example of monomeric protein allostery, extensive molecular dynamics[77,78] and PMF simulations[78] were used to explore the coupling between various conformational transitions (e.g. the $\beta4$–$\alpha4$ loop transition, Tyr106 rotation, Thr87 displacement) and phosphorylation in both the wild type CheY and the Thr87Ala mutant. In particular, using transition path sampling (TPS), it was shown through ~−160 natural reactive trajectories [Fig. 10.3(b)] that the isomerization of Tyr106 does not require the displacement of Thr87, and that the hydrogen bond between Thr87 and Asp57 phosphate, an essential element of the "Y-T" scheme, is not formed. Recognizing the local nature of TPS simulations, extensive two-dimensional PMF simulations were also carried out to explore the energetic coupling between key degrees of freedom; each two-dimensional projection in Fig. 10.3(c) is generated using between 100–200 ns of simulations. The results showed that the isomerization of Tyr106 and formation of the Thr87-phosphate hydrogen bond have similar barriers and are

Fig. 10.3 Study of functional motions in CheY.[78] **(a)** Comparison of the inactive[75] and active[76] structures of CheY. Overlay of key residues between the phosphorylation (Asp57) and response sites (Tyr106). Residues in the active structure are colored according to atom types, while those in the inactive structure are colored ice-blue. The inactive and active configurations of the $\beta4$–$\alpha4$ loop are colored dark-blue and yellow respectively. **(b)** Four configurations along an exemplary activation trajectory harvested using TPS; note that the intrinsic time scale of barrier crossing, which is different from the reaction time, is short and on the picosecond scale. Several important residues, including Tyr106, Ala90, Ile95 and Val108, are shown in the van der Waals scheme; the phosphorylated Asp57 is shown in the licorice form; the $\beta4$–$\alpha4$ loop is shown as the blue ribbon. **(c)** A three-dimensional scheme that illustrates the energetics and possible pathways for CheY activation based on the computed two-dimensional PMFs along the key degrees of freedom. The expected fully active state, A_I^F, is not a local free-energy minima in the simulations, presumably due to the absence of the FliM peptide in the model.

thermodynamically coupled; i.e. kinetically, either event can occur first and facilitate the other. The PMF results also showed that the $\beta4$–$\alpha4$ loop transition has substantially higher barriers, and therefore, is unlikely to gate the Tyr106 rotation; rather, the rotation of Tyr106 stabilizes the active configuration of this loop, which is consistent with a statistical analysis of all CheY structures in the PDB.[79,80] Thus, the CheY simulations show that a structural transition at the response site (Tyr106 isomerization) can occur prior to the so-called activation event (Thr87-phosphate hydrogen-bond formation). This suggests that the Tyr orientations are in equilibrium and that the active conformation is stabilized by Thr87-phosphate hydrogen bond formations; kinetically, either event can occur first. In the NMR study of the closely related NtrC,[64] motion associated with the equivalent Tyr was observed to persist in both the unphosphorylated and phosphorylated forms, which led the interpretation that Tyr rotation is "uncoupled" from phosphorylation. Combined TPS/PMF analyses of CheY support that the rotation of Tyr may occur in the absence of phosphorylation, but it *is* coupled thermodynamically with phosphorylation.

10.3.3 *Functional Motions in Molecular Motors at Multiple Scales*

Molecular motors are fascinating systems that convert the chemical free energy in the form of ATP binding/hydrolysis into mechanical work with high efficiency.[81,82] These systems are rich in motions/reactions that span multiple scales ranging from Angstrom-level changes in ATP·HO during hydrolysis, through local structural rearrangements in the nucleotide binding site, to domain-scale conformational transitions associated with displacement of the motor. Understanding the "mechanochemical" coupling in motors clearly requires characterizing these motions individually and revealing how they are coupled.[5,83]

10.3.3.1 *Mechanochemical coupling in myosin*

The specific system we focus on is the conventional myosin (referred to as myosin below), which is involved in muscle contraction.[84] It is one of the few motor systems for which high-resolution X-ray structures (for the motor domain) have been solved for multiple functional states.[85] The two X-ray structures of interest here[86,87] are believed to correspond to the post-rigor and pre-powerstroke states in the kinetic scheme,[84] and the transition from the former to the latter is referred to as the "recovery stroke." In these two states, the motor domain is detached from the actin and ATP hydrolysis is believed to occur only in the pre-powerstroke state. Comparison of the two X-ray structures reveals structural transitions at different scales, and the most notable ones are [Fig. 10.4(a)]: (i) the C-terminal converter sub-domain undergoes a ~60 degrees rotation, which corresponds to a RMSD of more than 20 Å; (ii) the active-site undergoes an open/close transition with a small RMSD of 3 Å; (iii) the relay helix, which connects the active site and the converter, undergoes a significant kink. The fundamental challenge is to understand how these motions are coupled and their relationships to the nucleotide state ($ATP{\cdot}H_2O$ versus $ADP{\cdot}Pi$) in the active site.

10.3.3.2 *Simulation studies*

To meet this challenge, a multitude of computational methods have been combined synergistically in our study.[13,88–91] The general strategy is to characterize the energetics of local events such as ATP hydrolysis[13] and open/close transition of the active site[89] in different X-ray structures; the results provide important information about how these local motions/reactions are coupled to structural changes elsewhere in the motor domain. In addition, approximate transition path calculations, normal mode analyses, and a bioinformatics-based approach are combined to identify residues/interactions that play an important role in the recovery stroke.[90] Due to the limited space, we restrict ourselves to discussions on the structural transitions. Regarding ATP hydrolysis, it suffices to say that QM/MM calculations of the hydrolysis energetics with different active-site structures[13,91] clearly showed that the activity relies on the *complete* closure of the active site, which in turn is coupled to the converter rotation through the relay helix (see below); as a result, the hydrolysis of ATP

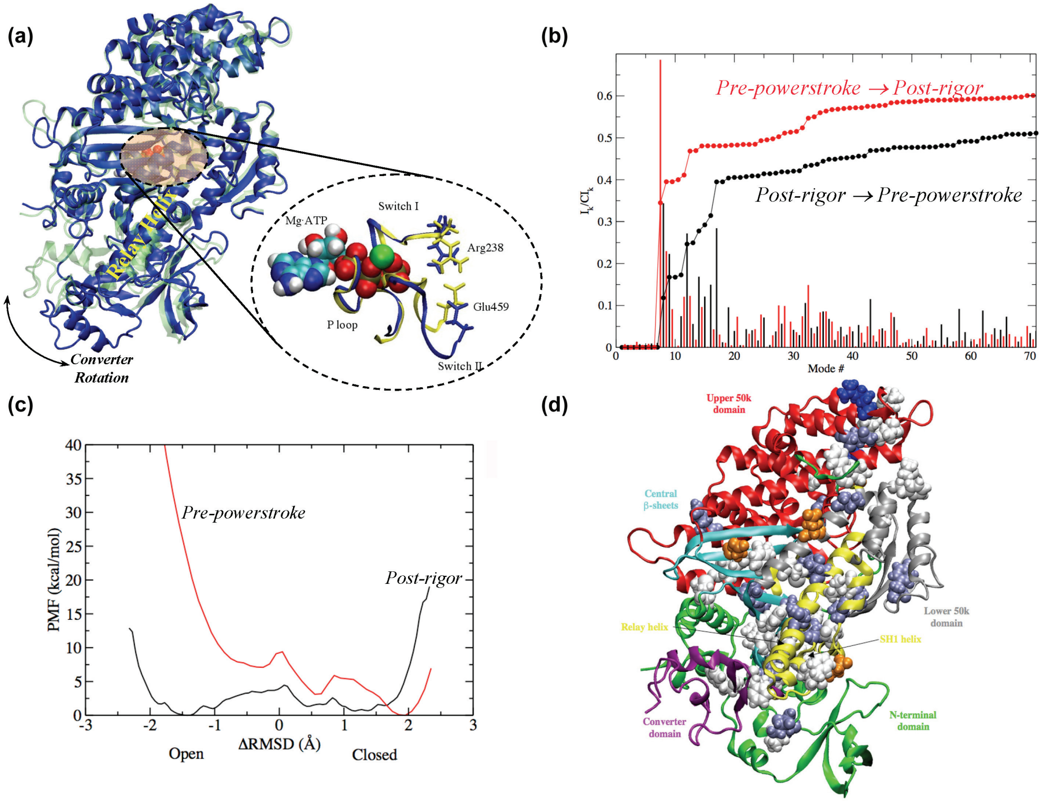


Fig. 10.4. (See caption on next page.)

in the active site is tightly coupled to the converter rotation, despite the separation of more than 40 Å.

First, normal mode calculations found that with either of the two X-ray structures, a small number of low-frequency normal modes sums up to a large fraction of the Cartesian displacements corresponding to the recovery stroke. This is shown more quantitatively with two commonly used descriptors, the involvement coefficients (I_k) and the cumulative involvement coefficients (CI_n),

$$I_k = \frac{\mathbf{X}_1 - \mathbf{X}_2}{|\mathbf{X}_1 - \mathbf{X}_2|} \cdot \mathbf{L}_k \tag{10.8}$$

$$CI_n = \sum_{k=1}^{n} I_k^2 \tag{10.9}$$

where $\mathbf{X}_1$–$\mathbf{X}_2$ is the displacement vector between two conformations ($\mathbf{X}_1, \mathbf{X}_2$) and $\mathbf{L}_k$ is the *k*th eigenvector. As shown in Fig. 10.4(b), using less than 20 lowest-frequency modes, more than 50% of the displacement can be accounted for, indicating that the motor domain has inherent flexibility in the specific direction of the recovery stroke. Comparatively, with the same number of modes, the CI_n for the prepowerstroke state is notably higher than that for the post-rigor state

Fig. 10.4 Study of mechanochemical coupling in myosin.[89,90] **(a)** The difference between the post-rigor (1FMW,[86] in blue) and pre-powerstroke (1VOM,[87] in light green) states; the structures are aligned based on backbone atoms in the first 650 residues. Also shown is the superposition of the active site, where Mg·ATP is in the van der Waals form, and key loops (P-loop, Switch I/II) as ribbons; the active site is "closed" with the salt-bridge between Arg238 and Glu459, and "open" when the salt-bridge is broken; **(b)** Involvement coefficient (I_k) and cumulative involvement coefficient (CI_k) from normal mode calculations (Equations 10.8 and 10.9) for the structural transitions between the two X-ray structures using the modes of either structure; **(c)** PMF for the open/close transition of the active site with different X-ray structures; **(d)** Mapping of the 52 strongly coupled core residues (in van der Waals form) to the structure of *Dictyostelium* myosin motor domain. The coupled residues are colored based on residue type; blue: basic, orange: acidic, ice-blue: polar, white: non-polar.

[Fig. 10.4(b)], which can be interpreted to suggest that the former is more flexible in the direction of the functional transition. This in fact is consistent with the FRET study of Spudich *et al.*[92] who found that the orientation of the converter (lever arm) is relatively rigid in the post-rigor state but spans a broader range of angles in the pre-powerstroke state.

To characterize the open/close transition of the active site, PMF calculations were carried out with both X-ray structures. The differential RMSD with respect to the open and closed configurations of the Switch I and II motifs is used as the reaction coordinate. As shown in Fig. 10.4(c), the results are strikingly different in the two X-ray structures. In the post-rigor state, the PMF profile is very flat, suggesting that the open and closed configurations have similar energetics and the transition between them is a low barrier process. In the pre-powerstroke structure, by contrast, the PMF is strongly titled toward the closed configuration, while the open configuration is at least 8 kcal/mol higher in free energy. Therefore, the PMFs quantitatively showed that rotation of the converter causes structural changes that propagate to the neighborhood of the active site such that the relative stability of the open/close configurations is strongly perturbed. In fact, data from the PMF simulations can also be used to construct the (ϕ, ψ)-free energy profile of residues near the active site. The results (not shown here)[89] clearly indicate that the motion of these residues becomes substantially restricted in the post-rigor state. As mentioned above, since the ATP hydrolysis activity is very sensitive to the active site configuration (including the position of water molecules), this tight coordination between converter orientation and active site stability ensures that the later is also tightly coupled to ATP hydrolysis (i.e. “mechanochemical coupling”).

Finally, to further explore residues/interactions that dictate the coupling between converter rotation and active site closure, the approximate transition path for the recovery stroke was studied using targeted molecular dynamics simulations,[26] as an alternative to minimum energy path analysis.[93] The main goal is to observe the formation of *transient* interactions that are not present in either end-states, and therefore, difficult to identify using the static X-ray structures. Analysis of the results[90] indicates that different types of interactions (polar versus hydrophobic) along the relay helix play an important role during the recovery stroke. Around halfway in the relay helix, the hydrophobic cluster provides stabilization to the kink of the relay helix, while at the joint between the relay helix and the relay loop region, strong polar interactions facilitate co-operative changes in the relay helix, the SH1 helix, and the converter domain. In addition to those *local* interactions, hinge residues in the low-frequency modes with large I_k values were also analyzed; the idea is that disruption of these hinges may perturb the flexibility of the system in important directions, thus the hinge residues should be of functional significance.[88] Among all the hinges identified, a small but significant fraction is highly conserved (>80% across all species), which supports their functional importance. More interestingly, among the 52 residues identified as "strongly coupled (co-evolved)" by the Statistical Coupling Analysis (SCA),[94] most are either a hinge residue or involved in an important interaction in the TMD simulations. This is a significant finding because the SCA algorithm works with sequence information only, thus the identified residues are not guaranteed to be involved in allostery and might instead play a role in, for example, co-operative folding. This observation highlights the value of combining a bioinformatics-based approach with physically motivated analyses for identifying key residues that dictate functional motions.

10.3.4 *Mechanical Response of a Mechanosensitive Channel*

As the last example, we illustrate how continuum mechanics models, even with a simple parameterization at this stage, can offer unique insights regarding functional motions triggered by external mechanical perturbations.

10.3.4.1 *Gating transition of MscL*

The specific system is the mechanosensitive channel of the large conductance (MscL) in *E. coli*, which acts as the "safety valve" for the bacterium by opening up when osmotic pressure is above a certain threshold.[95,96] MscL is one of the first examples that illustrated that mechanical sensing can occur without the involvement of the cytoskeleton.[97] It is now commonly accepted that the sensing process occurs through the mechanical deformation of the lipid membrane and its interaction with the embedded protein, although a complete understanding of the gating mechanism is not yet available.[98] For example, although protein-lipid mismatch has been shown to be important in the gating process,[99] additional force is required to fully open the channel. Moreover, the cytoplasmic S3 helix-bundle was thought to play an important role in the gating process in the first version of structural models.[100] More recently, however, it was argued that the structural changes in the S3 bundle should be substantially smaller.[101] Since the gating process occurs on the millisecond time scale, it is difficult to simulate the transition using atomistic molecular dynamics. For example, even with a steered molecular dynamics approach[27,102] the pore radius reached only 9.4 Å after more than 10 ns, which is significantly smaller than the experimentally estimated radius (~19 Å) for the fully opened state. Approximate open-close transition trajectory can be obtained with targeted molecular dynamics simulations,[103] which, however, requires the detailed knowledge of both the closed and open states. In addition, in these biased all-atom simulations, the "pulling force" on the protein is large in magnitude and artificial since the simulations were short and the lipid bilayer membrane was completely ignored in the TMD simulations.[103]

10.3.4.2 *Simulation studies*

Motivated by the X-ray structures of MscL from *Tb*.[54] and the homology model of the *E. coli* system, we established a simple continuum model for the *E. coli* MscL.[55] As shown in Fig. 10.5, the model contains all the essential structural motifs including the transmembrane (TM1,TM2)/cytoplasmic (S1–S3) helices and periplasmic loops; in an earlier "minimalist" model,[53] only the TM1 and TM2 were included since it was speculated that due to their extensive interactions with the lipid, the transmembrane helices are the most important components in MscL; the performances of the minimalist and full models of protein are compared below. The helices are treated as homogeneous (i.e. no sequence dependence has been included here) and isotropic rods, with the loops as elastic springs. The Young's moduli of the rods are taken from the estimate based on all-atom simulations of Sun *et al.*[104] using the CHARMM force field, and those for the springs are established by fitting the lowest three normal modes of the isolated springs from the continuum model and from an all-atom CHARMM calculation.[105] The material properties of the elastic membrane are approximated by values in the literature for DPPC.[106,107] Non-bonded interactions between different continuum components were estimated based on CHARMM force field energy calculations and fitted into simple functional forms similar to the Lennard-Jones interactions. The membrane sheet-helix rod interaction was estimated by rotating the corresponding helix in an implicit membrane using the Generalized Born model.[108] The fitted parameters can be applied to several structural configurations or different relative orientations of the channel in the close, open, and intermediate states.[100]

Once the model is parameterized, different mechanical stress can be applied to the membrane and quasi-static structural response of the channel can be solved using the finite element framework. In the published studies so far,[53,55] further simplifications were made in which the deformation of the lipid hole that contains the channel and structural response of the channel were calculated separately. This simplification is based on the assumption that the deformation of lipid dominates that of the protein, which reduces computational cost and can be easily removed by considering the full coupling between lipid and protein; in fact, such a comparison may yield important insights regarding the dominance of lipid mechanics during gating (Tang *et al.*, work in progress). Due to limited space, we do not discuss the quantitative aspects (e.g. estimate of proper tension, pore evolution profile) of the simulations, which can be found in Refs. 53 and 55. We restrict ourselves to two examples that illustrate the unique value of the FEM framework.

With the "minimalist" model, the structural response of MscL to in-plane biaxial stretching and out-of-plane bending was studied.[53]

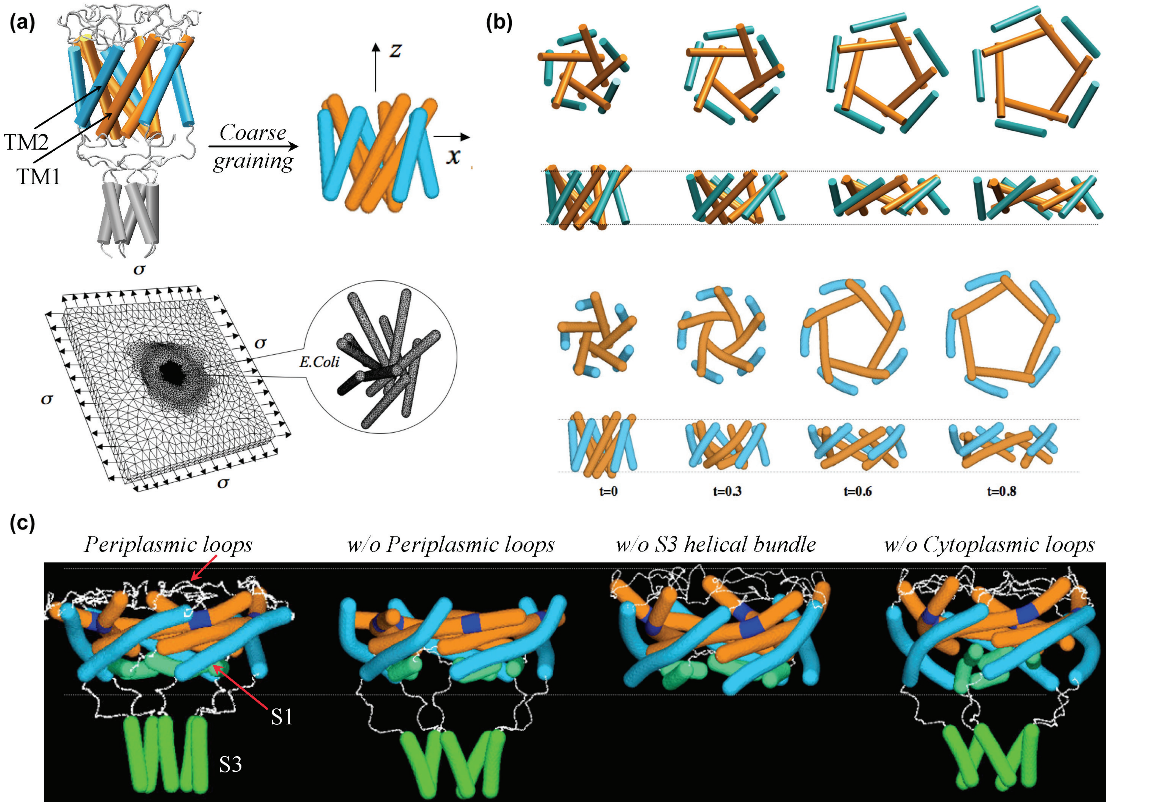


Fig. 10.5. (See caption on next page.)

With the proper magnitude (~35 MPa) of in-plane stretching consistent with the experimental value, the channel was indeed observed to make a transition from the closed to the open configuration. The structural evolution from the FEM simulations compare very favorably to the structural model of Guy and co-workers;[100] this is significant because the continuum mechanics model was established largely based on the closed state only! With the out-of-plane bending, although the tilting angle of the transmembrane helices changed significantly, due to the lack of in-plane membrane deformation, the structural response of the protein model is minimal. This drastic difference clearly illustrates that the structural response of MscL depends sensitively on the form of the mechanical perturbation and the gating transition relies critically on the in-plane tension in the membrane rather than the curvature of the membrane.

As the second example, we turn to recent results with the more complete structural model[55] at the continuum level and focus on the effect of in-plane stretching of the membrane. Similar to the results for the "minimalist" model, the channel fully opens under the proper magnitude of in-plane tension. At the same membrane strain level, the pore radius of the final state is, however, about 20% smaller than that in the "minimalist" model, which indicates that the presence of additional structural motifs other than the transmembrane helices (most likely the periplasmic loops, see below) tend to reduce pore opening. Nevertheless, the relatively small difference between results

Fig. 10.5 Study of structural response of MscL from *E. coli* based on a continuum mechanics model. **(a)** The basic philosophy of the coarse-graining procedure (replacing transmembrane helices by homogeneous and isotropic elastic rods, bilayer by elastic sheet) and set-up of the finite element simulation;[53] **(b)** Comparison of the transmembrane helices in terms of their packing and tilting angles during the gating transition from the structural model of Guy *et al.*[100] and from the finite element simulations;[53] **(c)** The side view of the structure at the end of the simulated gating transition using a more complete continuum model at the continuum level,[55] along with the corresponding structure when the periplasmic loops, the S3 helical bundle, and the cytoplasmic loops, respectively, are removed from the model. The dotted lines in **(b)**–**(c)** approximately indicate the location of the membrane-water interface.

from the complete and the "minimalist" model indicates that the gating process is dominated by the iris-like expansion of transmembrane helix bundles. Interestingly, the structural variation of the S3 helical bundle during the gating transition is very small, which supports the recent modification of the structural model.[101] The cytoplasmic S1 helical bundle, on the other hand, moves into the transmembrane region and opens up; the periplasmic loops also closely follow the trajectory of the transmembrane helices.

Another type of interesting study is to remove a specific structural component and observe the effect on the gating transition; this is a unique aspect of computational analysis since the corresponding exercise with experiments will be complicated by factors such as major structural distortions prior to channel activation. Here, we have tested the role of three structural motifs: the S3 helical bundle, the periplasmic loops, and the cytoplasmic loops that connect S1 and TM1 helices. As expected, based on the above discussion, removing S3 helices did not cause much change in the gating behavior, once again confirming the insignificant role of S3 for the opening of MscL. Removing the periplasmic loops causes major variations in the configurations (e.g. tilting angle) of the transmembrane helices, and therefore, the final pore size; in addition, removing the cytoplasmic loops makes the S1 bundle distorted, thus also affecting the pore radius. Both observations regarding the importance of these loops are consistent with the recent experimental observation of Sukharev and co-workers.[101,98]

10.4 Concluding Discussions and Future Outlook

There is little doubt that biomolecules are flexible objects and rich in motions of different temporal and spatial scales. Characterizing the nature of these motions and how they are perturbed by changes in the environment (e.g. osmotic stress) or ligation state is a fundamental challenge in structural biology and biophysics. It is even more challenging, however, to identify *functional motions* that in fact play a major role in facilitating the function of biomolecules, which can be striking domain-scale rearrangements that "propel" a molecular motor forward or subtle local changes that set up the proper active site or interface for the subsequent catalysis or binding.

Through the examples in this chapter, we hope to illustrate that modern computational approaches are making rapid advances so that processes at multiple scales can be investigated. As a result, computational analysis can play a major role in the study of functional motions, in terms of both helping better interpret experimental data and stimulating new hypotheses regarding the nature of such motions and mechanisms by which they are regulated. For example, our study of RNase A helped establish a concrete hypothesis regarding how the perturbation of a catalytic residue's motion may lead to a significant decrease in the observed catalytic rate for the D121A mutant. This hypothesis, which is more molecular in nature than the original proposal[58] that underlines the importance of "coordinated" and "time" global dynamics (however, see discussions below regarding current challenges), should be tested by further mutation studies.

In most cases, the functional motion implicates multiple structural changes, and therefore, an important issue is to understand how these structural changes are coupled and whether there is distinct "causality" (or sequence of event) between them. In CheY, for example, a key question is whether the Tyr rotation is dependent on the hydrogen-bond formation between Thr87-phosphate. In the molecular motor myosin, a fundamental question regards whether ATP hydrolysis in the active site triggers structural transitions that eventually propagate to the converter or the converter rotation occurs first,[109] which then leads to changes in the nucleotide binding site that activate the ATP hydrolysis. Since it remains difficult to directly observe those events in real-time, either computationally or experimentally, the best approach is to characterize the *energetic coupling* between different processes, which can be achieved with careful potential of mean force (PMF) computations.

In CheY, extensive multi-dimensional PMF simulations[78] revealed the energetics of different motions, and therefore, how they are coupled. The results support the idea that the isomerization of a key Tyr residue can occur prior to the activation event (hydrogen-bonding formation between Thr87-phosphate) with a modest barrier; unlike the original experimental interpretation of the similar NtrC,[64] however, the calculations clearly indicate that the Tyr rotation and the activation event are coupled. In myosin, PMF calculations[89] for the active-site open/close transition with different X-ray structures also convincingly showed how converter rotation propagates to structural changes near the active site, such that the open/close energetics get affected significantly. As a result, the ATP hydrolysis activity is tightly coupled to the converter rotation, despite the large separation of more than 40 Å.

Due to the complex nature of functional motions, it is productive to combine multiple computational techniques, similar to the use of multiple approaches in an experimental investigation. In the study of CheY activation,[78] the collection of hundreds of reactive trajectories from TPS simulations are instructive but not conclusive due to the local nature of the employed TPS algorithm. However, these *natural* reactive trajectories played a major role in identifying the proper coordinates for the subsequent multi-dimensional PMF simulations. In the study of myosin,[90] the recovery stroke is a highly complex process that involves both domain-scale motions and extensive rearrangements at the loop or side-chain levels. To identify residues that play a key role in the process, combining normal mode-based hinge analysis, targeted molecular dynamics, and the statistical coupling analysis was productive, because these techniques are based on different fundamental assumptions, and therefore, complement each other well.

Another form of combining different methods is propagating information from all-atom simulations to an effective coarse-grained model, which is illustrated here with the continuum mechanics model of MscL.[53,55] Even with rather simple proof-of-concept type of models, new insights have been obtained regarding the impact of different forms of mechanical perturbation on the gating transition of MscL and the role of various structural motifs in the process. Although still in its infancy, if done carefully and properly, this type of strategy can be very powerful for analyzing the functional motions of very large biomolecular complexes at very large length (even cellular) scales.

10.4.1 *Outstanding and Emerging Challenges*

Although it is always presumptuous to speculate too much into the future, we briefly ponder several subjects for which we would like to see further studies. Instead of discussing these from the perspective of technical developments, which clearly will continue on multiple fronts and at multiple scales, we point out a number of questions regarding "functional motions" that the authors believe are particularly interesting to explore.

10.4.1.1 *What are the roles of "slow (μs-ms) motions" in enzyme catalysis?*

A significant body of computational studies has been focused on analyzing the impact of motion on enzyme catalysis. However, essentially *all* calculations focused on relatively fast motions on the order of pico- to nanoseconds, due to either limits in the computational resources or the fact that the goal was to study the impact of enzyme motions on the barrier crossing process, which *does* occur at the picosecond time scale for most chemical reactions. Therefore, an important issue that has not been extensively analyzed at the molecular level concerns the possibility that slow (*μ*s -ms) motions may significantly modulate the enzyme (active-site) structure so that a significant number of chemical turnovers, in fact, occur in "excited state" conformation(s). In this regard, we note that we do not consider "simple" cases where a conformational change (e.g. closure of the active-site upon substrate binding) is a kinetically distinct step prior to catalysis;[110] rather, we focus on systems where the most catalytically active conformation is rarely populated and distinct from the most stable Michaels-complex as observed by, for example, crystallography. Another way of stating the issue is that the most populated conformation observed under a specific experimental condition may not be the most functionally active one. Single molecule experiments demonstrated that the "rate-constant" (or apparent barrier) of an enzyme (e.g. cholesterol oxidase) catalyzed reaction is in fact time-dependent,[111] presumably due to structural transitions between different conformational sub-states; the molecular nature of such transition has not been illustrated and the magnitude of the apparent barrier fluctuation is usually small. The existence of important "global" μs-ms motion has been suggested in several enzyme systems based on NMR relaxation measurements,[112,1] including the RNase system discussed earlier.[57,58] Although alternative explanations *may* exist, a conclusive analysis of the nature of "functional motions" is not complete unless motions on the μs-ms time-scale are accessed and their impact on the catalysis analyzed at the atomic level. This is clearly a challenging task for computations because a meaningful description of such long-time motions requires both extensive sampling and reliable force fields. Before *tour de force* analyses can be done, it is likely that combining QM/MM analysis and enhanced sampling techniques, such as "conformational flooding",[113] may produce instructive insights.

10.4.1.2 *What are the bottlenecks for large-scale functional motions?*

To completely *understand* functional motion, it is important to identify the kinetic bottleneck of the process among all the implicated structural transitions. The above discussion of myosin made it clear that functional motions likely involve both domain-scale changes *and* important local structural transitions. Domain motions are more striking in scale while the local transitions more subtle, but the spatial magnitude of changes does *not* necessarily correlate with kinetic significance. As noted above, many studies found that large-scale structural transitions are correlated with low-frequency modes, which implies that biomolecules tend to have intrinsic structural flexibilities that ensure domain-scale motions are largely diffusive in nature; therefore, the kinetic bottleneck of a functional transition may, in fact, consist of key local structural changes that are thermally activated. Such considerations highlight the importance of revealing the free energy landscape of functional motions, for which detailed computational analysis beyond a simple harmonic picture is indispensable.

The realization that local structural changes may constitute the kinetic bottleneck of complex structural transitions has important implications regarding strategies for constructing meaningful coarse-grained models in the context of studying functional motions. For instance, although it seems sensible to coarse-grain biomolecules into rigid domains, the predictive power of such models might be significantly compromised if important local features (e.g. repacking of hydrophobic side-chains in the rely loop/helix in myosin) are ignored. In this regard, an important emerging challenge is to make quantitative connections, at multiple resolutions and scales, between computational models and experiments that report on the time and spatial scales of biomolecular motions; most notable examples include small-angle X-ray scattering,[114,115] diffuse X-ray scattering,[116] fluorescence resonance transfer (FRET),[117] electron spin resonance,[118] and two-dimensional infrared spectroscopy,[119] which span a broad range of time resolutions and spatial scales. Φ analysis, which is commonly used in protein folding analysis[120] and recently applied to study motions in the acetylcholine receptor,[121] also provides extremely valuable data regarding whether specific residues are involved in the transition state ensemble of the transition. Making explicit comparison to experiments provides not only important validations for the computational model but also the opportunity of gleaning additional information from experimental data. Recent applications of elastic network models in the refinement of X-ray structures, EM structures, electron tomography, and FRET data are good examples.[32,33,122,134,135]

10.4.1.3 *Can functional motions be modulated in a predictive manner?*

Although significant motions are implicated in the functional cycle of biomolecules, protein engineering studies have largely been guided by static structural considerations, which reflects our lack of thorough understanding of factors that dictate the features of functional motions. Although the situation will improve steadily, the most productive avenue for incorporating molecular motions into protein design in the near future likely involves combining clever genetic approaches, molecular simulations, and informatics motivated models.

In an impressive recent study,[123] for example, a novel gene synthesis approach was used to construct chimeras between the mesophilic and thermophilic adenylate kinases, in which different domains from the two enzymes are combined randomly (eight were considered). Measurement of thermostability and enzyme activity (which is limited by a structural transition that implicate the active site closure) revealed that it *is* possible to enhance the flexibility of key domains without affecting the thermostability. Further studies of this sort, supplemented by simulation and bioinformatics analysis,[124] may lead to new avenues of manipulating protein functions through rationally modulating essential motions. In addition to enzymes, interesting targets are molecular motors and other allosteric systems, in which specific mutations are known to disrupt functional motions such that communication between different sites is abolished.[125–127] Since these mutations often lead to serious diseases, devising effective methods for restoring key functional motions has great biomedical implications.

10.4.1.4 *Are there major differences between "functional motions" in vitro and those in vivo?*

Finally, a recent trend in biophysical studies is to contrast molecular behaviors under *in vitro* and *in vivo* conditions. It has been recognized for some time that the cellular environment is far from the dilute solution condition in *in vitro* experiments; molecular crowding, non-specific binding and other features associated with non-ideality may significantly affect the structure, stability, and association of biomolecules in cells.[128–130] Functional motions of biomolecules, especially those large-scale motions that strongly implicate nearby water/solute molecules, might be substantially different *in vivo* compared to *in vitro*. To what extent this is true and what are the corresponding functional implications are clearly very interesting questions that deserve careful analysis. Along this line, recent studies of both the protein[131] and solvent[132] dynamics in reverse micelles provided interesting clues. Clearly, more studies are needed from both the theoretical/computational and experimental perspectives to fully understand biomolecular motions that are most important under the physiological condition.

Acknowledgments

Q. Cui and X. Chen acknowledge other group members and collaborators who made significant contributions to the work discussed here. The studies reviewed have been generously supported from a Research Innovation Award from the Research Corporation and the National Institutes of Health (R01-GM071428-01) to Q. Cui and a CAREER award to X. Chen (CMMI-0643726) from the National Science Foundation. Q. Cui also acknowledges an Alfred P. Sloan Research Fellowship. Computational resources from the National Center for Supercomputing Applications at the University of Illinois are greatly appreciated.

References

1. Boehr DD, Dyson HJ, Wright PE. (2006) An NMR perspective on enzyme dynamics. *Chem Rev* **106**: 3055–3079.
2. Jardetzky O. (1996) Protein dynamics and conformational transitions in allosteric proteins. *Prog Biophys Mol Biol* **65**: 171–219.
3. Karplus M, Kuriyan J. (2005) Molecular dynamics and protein function. *Proc Natl Acad Sci USA* **102**: 6679–6685.
4. Zhou YQ, Vitkup D, Karplus M. (1999) Native proteins are surface-molten solids: application of the Lindemann criterion for the solid versus liquid state. *J Mol Biol* **285**: 1371–1375.
5. Vale RD, Milligan RA. (2000) The way things move: looking under the hood of molecular motor proteins. *Science* **288**: 88–95.
6. Changeux J-P, Edelstein SJ. (2005) Allosteric mechanisms of signal transduction. *Science* **308**: 1424–1428.
7. Benkovic SJ, Hammes-Schiffer S. (2003) A perspective on enzyme catalysis. *Science* **301**: 1196–1202.
8. Karplus M, McCammon JA. (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **9**: 646–652.
9. McCammon JA, Gelin BR, Karplus M. (1977) Dynamics of folded proteins. *Nature* **267**: 585–590.
10. Silverman DN. (1995) Marcus rate theory applied to enzymatic proton transfer. *Meth Enzymol* **249**: 479–503.
11. Joseph-McCarthy D, Petsko GA, Karplus M. (1990) Anatomy of a protein conformational change: hinged "lid" motion of the triosephosphate isomerase loop. *Science* **249**: 1425–1428.

12. Onishi H, Ohki T, Mochizuki N, Morales MF. (2002) Early stages of energy transduction by myosin: roles of Arg in switch I, of Glu in switch II, and of the salt-bridge between them. *Proc Natl Acad Sci USA* **99**: 15339–15344.
13. Li GH, Cui Q. (2004) Mechanochemical coupling in myosin: a theoretical analysis with molecular dynamics and combined QM/MM reaction path calculations. *J Phys Chem B* **108**: 3342–3357.
14. McQuarrie DA. (1973) *Statistical Mechanics*. Harper and Row, New York.
15. Frenkel D, Smit B. (2002) *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, San Diego.
16. Geissler PL, Dellago C, Chandler D. (1999) Kinetic pathways of ion pair dissociation in water. *J Phys Chem B* **103**: 3706–3710.
17. Bolhuis PG, Dellago C, Chandler D. (2000) Reaction coordinates of biomolecular isomerization. *Proc Acad Natl Sci USA* **97**: 5877–5882.
18. Zhou Y, Zhou H, Karplus M. (2003) Co-operativity in *Scapharca* dimeric hemoglobin: simulation of binding intermediates and elucidation of the role of interfacial water. *J Mol Biol* **326**: 593–606.
19. Bolhuis PG, Chandler D, Dellago C, Geissler PL. (2002) Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Ann Rev Phys Chem* **53**: 291–318.
20. Dellago C, Bolhuis PG, Geissler PL. (2002) Transition path sampling. *Adv Chem Phys* **123**: 1–78.
21. McCammon JA, Karplus M. (1979) Dynamics of activated processes in globular proteins. *Proc Natl Acad Sci USA* **76**: 3585–3589.
22. Hagan MF, Dinner AR, Chandler D, Chakraborty AK. (2003) Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA. *Proc Natl Acad Sci USA* **100**: 13922–13927.
23. Juraszek J, Bolhuis PG. (2006) Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. *Proc Acad Natl Sci USA* **103**: 15859–15864.
24. Ma A, Dinner AR. (2005) Automatic method for identifying reaction coordinates in complex systems. *J Phys Chem B* **109**: 6769–6779.
25. Bolhuis PG. (2003) Transition-path sampling of beta-hairpin folding. *Proc Natl Acad Sci USA* **100**: 12129–12134.
26. Schlitter J, Engels M, Krüger P, Jacoby E, Wollmer A. (1993) Targeted molecular dynamics simuation of conformational change: application to the T→R transition in insulin. *Mol Simul* **10**: 291–308.
27. Sotomayor M, Schulten K. (2007) Single-molecule experiments *in vitro* and *in silico*. *Science* **316**: 1144–1148.
28. Elber R, Ghosh A, Cardenas A. (2002) Long time dynamics of complex systems. *Acc Chem Res* **35**: 396–403.
29. Cui Q, Bahar I. (eds). (2006) *Normalmode Analysis: Theory and Applications to Biological and Chemical Systems*. Chapman and Hall/CRC, New York.

30. Cui Q, Li GH, Ma JP, Karplus M. (2004) A normal mode analysis of structural plasticity in the biomolecular motor F(1)-ATPase. *J Mol Biol* **340**: 345–372.
31. Xu CY, Tobi D, Bahar I. (2003) Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T, R transition. *J Mol Biol* **333**: 153–168.
32. Ma JP. (2005) Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* **13**: 373–380.
33. Tama F, Brooks CL. (2006) Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Ann Rev Biophys Biomol Struct* **35**: 115–134.
34. Tama F, Gadea FX, Marques O, Sanejouand YH. (2000) Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* **41**: 1–7.
35. Li GH, Cui Q. (2002) A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca^{2+}-ATPase. *Biophys J* **83**: 2457–2474.
36. Van Wynsberghe AW, G Li, Cui Q. (2004) Normal mode analysis suggests protein flexibility modulation throughout RNA polymerase's function cycle. *Biochemistry* **43**: 13083–13096.
37. Li GH, Van Wynsberghe A, Demerdash ONA Cui, Q. (2006) *Normal Mode Analysis: Theory and Application to Biological and Chemical Systems*, pp. 65–89. Chapman and Hall/CRC, New York.
38. Atilgan AR, Durell SR, Jernigan RL, *et al.* (2002) Anisotropy of fluctuation dynamics of proteins with an elastic network model *Biophys J* **80**: 505–515.
39. Bahar I, Rader AJ. (2005) Coarse-grained normal mode analysis in structural biology. *Curr Opin Struct Biol* **15**: 586–592.
40. Van Wynsberghe AW, Cui Q. (2005) Comparisons of mode analyses at different resolutions applied to nucleic acid systems. *Biophys J* **89**: 2939–2949.
41. Kondrashov DA, van Wynsberghe AW, Bannen RM, Cui Q, Phillips GN. (2007) Protein structural variation in computational models and crystallographic data. *Structure* **15**: 169–177.
42. Ming D, Kong YF, Lambert MA, Huang Z, Ma JP. (2002) How to describe protein motion without amino acid sequence and atomic coordinates. *Proc Acad Natl Sci USA* **99**: 8620–8625.
43. Tama F, Valle M, Frank J, Brooks CL. (2003) Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc Natl Acad Sci* **100**: 9319–9323.
44. Van Wynsberghe AW, Cui Q. (2006) Interpreting correlated motions using normal mode analysis. *Structure* **14**: 1647–1653.
45. Izvekov S, Voth GA. (2005) A multiscale coarse-graining method for biomolecular systems. *J Phys Chem B* **109**: 2469–2473.

46. Shelley JC, Shelley MY, Reeder RC, Bandyopadhyay S, Klein ML. (2001) A coarse-grain model for phospholipid simulations. *J Phys Chem B* **105**: 4464–4470.
47. V Tozzini. (2005) Coarse-grained models for proteins. *Curr Opin Struct Biol* **15**: 144–150.
48. Noid WG, Chu JW, Ayton GS, Voth GA. (2007) Multiscale coarse-graining and structural correlations: connections to liquid-state theory. *J Phys Chem B* **111**: 4116–4127.
49. Marrink SJ, de Vries AH, Mark AE. (2004) Coarse-grained model for semiquantitative lipid simulations. *J Phys Chem B* **108**: 750–760.
50. Wang HY, Oster G. (1998) Energy transduction in the F_1 motor of ATP synthase. *Nature* **396**: 279–282.
51. Zumdieck A, Lagomarsino MC, Tanase C, *et al.* (2005) Continuum description of the cytoskeleton: ring formation in the cell cortex. *Phys Rev Lett* **95**: 258103.
52. Huebner KH, Thornton EA, Byrom TG. (1995) *Finite Element Method for Engineers.* Wiley, New York.
53. Tang Y, Cao G, Chen X, *et al.* (2006) A finite element framework for studying mechanical response of macromolecules: application to the gating of the mechanosensitive channel. *Biophys J* **91**: 1248–1263.
54. Chang G, Spencer RH, Lee AT, Barclay MT, Rees D. (1998) Structure of the MscL homolog from Mycobacterium tuberculosis: a gated mechanosensitive ion channel. *Science* **282**: 2220–2226.
55. Tang Y, Chen X, Yoo J, Cui Q. (2007) Numerical simulations of patch clamp and nanoindentation experiments on mechanosensitive channels of large conductance. *J Exp Mech*, **in press**.
56. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Acad Natl Sci USA* **98**: 10037–10041.
57. Beach H, Cole R, Gill ML, Loria JP. (2005) Conservation of μs-ms enzyme motions in the apo- and substrate-mimicked state. *J Am Chem Soc* **127**: 9167–9176.
58. Kovrigin EL, Loria JP. (2006) Enzyme dynamics along the reaction coordinate: critical role of a conserved residue. *Biochem* **45**: 2636–2647.
59. Schultz LW, Quirk DJ, Raines RT. (1998) His-Asp catalytic dyad of ribonuclease A: structure and function of the wild-type, D121N, and D121A enzymes. *Biochemistry* **37**: 8886–8898.
60. Quirk DJ, Raines RT. (1999) His-Asp catalytic dyad of ribonuclease A: histidine pKa values in the wild-type, D121N, and D121A enzymes. *Biophys J* **76**: 1571–1579.
61. Borkakoti N, Moss DS, Palmer RA. (1982) Ribonuclease A least-squares refinement of the structure at 1.45 Å resolution. *Acta Crystallogr B* **38**: 2210–2217.

62. Rico M, Santoro J, Gonzalez C, *et al.* (1991) 3D structure of bovine pancreatic ribonuclease A in aqueous solution: an approach to tertiary structure determination from a small basis of 1H NMR NOE correlations. *J Biomol NMR* **1**: 283–298.
63. Alberts B, Bray D, Lewis J, *et al.* (1994) *Molecular Biology of the Cell.* Garland Publishing, Inc., New York.
64. Volkman BF, Lipson D, Wemmer DE, Kern D. (2001) Two-state allosteric behavior in a single-domain signaling protein. *Science* **291**: 2429–2433.
65. Kern D, Zuiderweg ERP. (2003) The role of dynamics in allosteric regulation. *Curr Opin Struct Biol* **13**: 748–757.
66. Monod J, Wyman J, Changeux J-P. (1965) On the nature of allosteric transitions: a plausible model. *J Mol Biol* **12**: 88–118.
67. Koshland DEJ, Nemethy G, Filmer D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochem* **5**: 365–385.
68. Eaton WA, Henry ER, Hofrichter J, Mozzarelli A. (1999) Is cooperative oxygen binding by hemoglobin really understood? *Nat Struct Biol* **6**: 351–358.
69. Szabo A, Karplus M. (1972) A mathematical model for structure-function relations in hemoglobin. *J Mol Biol* **72**: 163–197.
70. Perutz MF. (1970) Stereochemistry of cooperative effects in hemoglobin. *Nature* **228**: 726–739.
71. Stock AM, Robinson VL, Goudreau PN. (2000) Two-component signal transduction. *Ann Rev Biochem* **69**: 183–215.
72. Cho HS, Lee SY, Yan DL, *et al.* (2000) NMR structure of activated CheY. *J Mol Biol* **297**: 543–551.
73. Barak R, Eisenbach M. (1992) Correlation between phosphorylation of the chemotaxis protein CheY and its activity at the flagellar motor. *Biochemistry* **31**: 1821–1826.
74. Appleby JL, Bourret RB. (1998) Proposed signal transduction role for conserved CheY residue Thr87, a member of the response regulator active-site quintet. *J Bacter* **180**: 3563–3569.
75. Simonovic M, Volz K. (2001) A distinct meta-active conformation in the 1.1 Å resolution structure of wild-type apo-CheY. *J Biol Chem* **276**: 28637–28640.
76. Lee S-Y, Cho HS, Pelton JG, *et al.* (2001) Crystal structure of an activated response regulator bound to its target. *Nat Struct Biol* **8**: 52–56.
77. Formaneck MS, Ma L, Cui Q. (2006) Reconciling the “old” and “new” views of protein allostery. A molecular simulation study of chemotaxis Y protein (CheY). *Proteins* **63**: 846–867.
78. Ma L, Cui Q. (2007) Activation mechanism of a signaling protein at atomic resolution from advanced computations. *J Am Chem Soc* **129**: 10261–10268.

79. Dyer CM, Dahlquist FW. (2006) Switched or not? The structure of unphosphorylated CheY bound to the N terminus of FliM. *J Bacter* **188**: 7354–7363.
80. Stock AM, Guhaniyogi J. (2006) A new perspective on response regulator activation. *J Bacter* **188**: 7328–7330.
81. Howard J. (2001) *Mechanics of Motor Proteins and the Cytoskeleton*. Sinauer Associates, Inc., Sunderland, Massachusetts.
82. Schliwa M. (ed). (2002) *Molecular Motors*. Wiley-VCH, New York.
83. Gao Y, Karplus M. (2004) Biomolecular motors: the F_1-ATPase paradigm. *Curr Opin Struct Biol* **14**: 250–259.
84. Geeves MA, Holmes KC. (2005) The molecular mechanism of muscle contraction. *Adv Protein Chem* **71**: 161–193.
85. Houdusse A, Sweeney HL. (2001) Myosin motors: missing structures and hidden springs. *Curr Opin Struct Biol* **11**: 182–194.
86. Bauer CB, Holden HM, Thoden JB, Smith R, Rayment I. (2000) X-ray structures of the Apo and MgATP-bound states of *Dictyostelium discoideum* myosin motor domain. *J Biol Chem* **275**: 38494–38499.
87. Smith CA, Rayment I. (1996) X-ray structure of the Magnesium(II)·ADP· Vanadate complex of the *Dictyostelium discoideum* myosin motor domain to 1.9 Å resolution. *Biochemistry* **35**: 5404–5417.
88. Li GH, Cui Q. (2004) Analysis of functional motions in Brownian molecular machines with an efficient block normal mode approach: myosin-II and Ca^{2+} — ATPase. *Biophys J* **86**: 743–763.
89. Yu H, Ma L, Yang Y, Cui Q. (2007) Mechanochemical coupling in myosin motor domain, I. Equilibrium active site simulations. *PLoS Comput Biol* **3**: 0199–0213.
90. Yu H, Ma L, Yang Y, Cui Q. (2007) Mechanochemical coupling in myosin motor domain, II. Analysis of critical residues. *PLoS Comput Biol* **3**: 0214–0230.
91. Yang Y, Yu H, Cui Q. (2007) **manuscript in preparation**.
92. Shih WM, Gryczynski Z, Lakowicz JR, Spudich JA. (2000) A FRET-based sensor reveals large ATP hydrolysis-induced conformational changes and three distinct states of the molecular motor myosin. *Cell* **102**: 683–694.
93. Fischer S, Windshugel B, Horak D, Holmes KC, Smith JC. (2005) Structural mechanism of the recovery stroke in the myosin molecular motor. *Proc Natl Acad Sci USA* **102**: 6873–6878.
94. Lockless SW, Ranganathan R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**: 295–299.
95. Hamill OP, Martinac B. (2001) Molecular basis of mechanotransduction in living cells. *Physiol Rev* **81**: 685–740.
96. Anishkin A, Kung C. (2005) Microbial mechanosensation. *Curr Opin Neuro* **15**: 397–405.
97. Kung C. (2005) A possible unifying principle for mechanosensation. *Nature* **436**: 647–654.

98. Perozo E. (2006) Gating prokaryotic mechanosensitive channels. *Nat Rev Mol Cell Biol* **7**: 109–119.
99. Perozo E, Kloda A, Cortes DM, Martinac B. (2002) Open channel structure of MscL and the gating mechanism of mechanosensitive channels. *Nat Struct Biol* **9**: 696–703.
100. Sukharev SI, Durell SR, Guy HR. (2001) Structural models of the MscL gating mechanism. *Biophys J* **61**: 917–936.
101. Sukharev S, Anishkin A. (2004) Mechanosensitive channels: what can we learn from "simple" model systems? *Trends Neurosci* **27**: 345–351.
102. Gullingsrud J, Schulten K. (2003) Gating of MscL studied by steered molecular dynamics. *Biophys J* **85**: 2087–2099.
103. Kong Y, Shen Y, Warth TE, Ma JP. (2002) Conformational pathways in the gating of *Escherichia coli* mechanosensitive channel. *Proc Natl Acad Sci USA* **99**: 5999–6004.
104. Choe S, Sun SX. (2005) The elasticity of α-helices. *J Chem Phys* **122**: 244912.
105. MacKerell ADJ, Bashford D, Bellott M, *et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102**: 3586–3616.
106. Evans E, Rawicz W. (1990) Entropy-driven tension and bending elasticity in condensed-fluid membranes. *Phys Rev Lett* **64**: 2094–2097.
107. Lindahl E, Edholm O. (2000) Spatial and energetic-entropic decomposition of surface tension in lipid bilayers from molecular dynamics simulations. *J Chem Phys* **113**: 3882–3893.
108. Im W, Feig M, Brooks CL. (2003) An implicit membrane generalized born theory for the study of structure, stability, and interactions of membrane proteins. *Biophys J* **85**: 2900–2918.
109. Miyashita O, Onuchic JN, Wolynes PG. (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci USA* **100**: 12570–12575.
110. Wolft-Watz M, Thai V, Henzler-Wildma KN, *et al.* (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat Struct Mol Biol* **11**: 945–949.
111. Lu HP, Xun LY, Xie XS. (1998) Single-molecule enzymatic dynamics. *Science* **282**: 1877–1882.
112. Eisenmesser EZ, Bosco DA, Akke M, Kern D. (2002) Enzyme dynamics during catalysis. *Science* **295**: 1520–1523.
113. Grubmüller H. (1995) Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys Rev E* **52**: 2893–2906.
114. Svergun DI, Koch MHJ. (2003) Small-angle scattering studies of biological macromolecules in solution. *Rep Prog Phys* **66**: 1735–1782.
115. Lipfert J, Doniach S. (2007) Small-angle X-ray scattering from RNA, proteins and protein complexes. *Ann Rev Biophys Biomol Struct* **36**: 307–327.

116. Benoit JP, Doucet J. (1995) Diffuse scattering in protein crystallography. *Q Rev Biophys* **28**: 131–169.
117. Michalet X, Kapanidis AN, Laurence T, *et al.* (2003) The power and prospects of fluorescence microscopies and spectroscopies. *Ann Rev Biophys Biomol Struct* **32**: 161–182.
118. Fanucci GE, Cafiso DS. (2006) Recent advances and applications of site-directed spin labeling. *Curr Opin Struct Biol* **16**: 644–653.
119. Zanni MT, Hochstrasser RM. (2001) Two-dimensional infrared spectroscopy: a promising new method for the time resolution of structures. *Curr Opin Struct Biol* **11**: 516–522.
120. Fersht AR, Matouschek A, Serrano L. (1992) The folding of an enzyme 1. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol* **224**: 771–782.
121. Purohit P, Mitra A, Auerbach A. (2007) A stepwise mechanism for acetylcholine receptor channel gating. *Nature* **446**: 930–933.
122. Zheng WJ, Brooks BR. (2006) Modeling protein conformational changes by iterative fitting of distance constraints using reoriented normal modes. *Biophy J* **90**: 4327–4336.
123. Bae E, Phillips GN. (2006) Roles of static and dynamic domains in stability and catalysis of adenylate kinase. *Proc Natl Acad Sci USA* **103**: 2132–2137.
124. Huang SW, Hwang JK. (2005) Relationship between local structural entropy and protein thermostability. *Proteins: Struct Funct Bioinform* **59**: 802–809.
125. Suel GM, Lockless SW, Wall MA, Ranganathan R. (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* **10**: 59–69.
126. Sasaki N, Ohkura R, Sutoh K. (2003) Dictyostelium myosin II mutations that uncouple the converter swing and ATP hydrolysis cycle. *Biochemistry* **42**: 90–95.
127. Murphy CT, Rock RS, Spudich JA. (2001) A myosin II mutation uncouples ATPase activity from motility and shortens step size *Nat Cell Biol* **3**: 311–315.
128. Minton AP. (1998) Molecular crowding: analysis of effects of high concentration of inert cosolutes on biochemical equilibria and rates in terms of volume exclusion. *Meth Enzymol* **295**: 127–149.
129. Ellis RJ. (2001) Macromolecular crowding: obvious but underappreciated. *Trends Biochem Sci* **26**: 597–604.
130. Record MT, Courtenay ES, Cayley S, Guttman HJ. (1998) Biophysical compensation mechanisms buffering *E. coli* protein-nucleic acid interactions against changing environments. *Trends Biochem Sci* **23**: 190–194.
131. Wand AJ, Ehrhardt MR, Flynn PF. (1998) High-resolution NMR of encapsulated proteins dissolved in low-viscosity fluids. *Proc Acad Natl Sci USA* **95**: 15299–15302.

132. Piletic IR, Tan HS, Fayer MD. (2005) Dynamics of nanoscopic water: vibrational echo and infrared pump-probe studies of reverse micelles. *J Phys Chem B* **109**: 21273–21284.
133. Humphrey W, Dalke A, Schulten K. (1996) VMD — visual Molecular Dynamics. *J Mol Graph* **14**: 33–38.
134. Delarue M, Dumas P. (2004) On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proc Natl Acad Sci USA* **101**: 6957–6962.
135. Poon BK, Chen XR, Lu MY, *et al.* (2007) Normal mode refinement of anisotropic thermal parameters for a supramolecular complex at 3.42-Å crystallographic resolution. *Proc Natl Acad Sci USA* **104**: 7869–7874.
136. Mouawad L, Perahia D. (1993) Diagonalization in a mixed basis: a method to compute low-frequency normal modes for large macromolecules. *Biopolymers* **33**: 599–611.

This page intentionally left blank

Chapter 11

Protein-Protein Interactions and Aggregation Processes

R. I. Dima*

11.1 Introduction

It is generally believed that any protein, under certain conditions (low pH, high temperature, high concentration), can form ordered aggregates composed of amyloid-like fibrils.[1,2] Still, only a limited set of about 25 proteins[3] can form, under near-native conditions, highly toxic oligomeric species and amyloid fibrils associated with a number of protein deposition diseases such as Alzheimer's disease,[4] transmissible spongiform encephalopathies (TSE),[5] and Huntington's disease.[6] The existence of this special class of proteins is surely puzzling because these soluble proteins often play essential biological functions under normal cellular conditions. However, in the disease-related state, provoked by a change in environment or genetic predisposition, parts of the native structure are lost, thus exposing a template for the growth of aggregates. Amyloid fibrils are typically composed of a varying number of protofilaments, each with a central spine of β-strands running perpendicular to the fibril axis (termed a cross-β spine conformational architecture).[3]

The ability of virtually any protein to assemble into ordered aggregates would suggest that the propensity to aggregate is a general

*Department of Chemistry, University of Cincinnati, Cincinnati, OH 45221. Email: dimari@ucmail.uc.edu.

property of the polypeptide backbone. Nevertheless, a number of studies[7,8] revealed that the tendency towards amyloid formation goes beyond the backbone being dependent on the amino acid composition (patterns) and on interactions between side-chains. For example, the propensity towards amyloidogenicity is correlated with high levels of hydrophobicity and reduced concentration of charged residues. To fully probe aggregation, one would have to have access to high-resolution structures of amyloid-fibrils formed by both proteins with a tendency to aggregate and by normal, non-amyloidogenic, proteins. Unfortunately, despite their ordered arrangement, amyloid fibrils are at best para-crystalline, so X-ray crystallography cannot be used to determine their structures. In addition, liquid state NMR cannot be applied either, because the fibrils are non-soluble. As a result, only one amyloid cross-β spine structure of a small seven-residue peptide from Sup35 has been determined to atomic detail.[9] Thus, approaches such as solid state NMR, site-directed spin labeling, cryo-electron microscopy, and proline-scanning mutagenesis have been used to determine the structures of amyloids formed by a set of small peptides. Moreover, determination of the amyloid-ready conformation(s) of a protein or a peptide is also highly non-trivial. In general, such structural species are at best metastable in monomeric form so techniques developed to measure the equilibrium (stable) structures of polypeptide chains are not useful here. Techniques such as high pressure NMR,[10,11] single-molecule approaches, such as force AFM, and H/D exchange are used instead to provide glimpses into the characteristics of the monomeric non-native conformations. Because of the difficulties in probing amyloid fibrils using experimental techniques, carefully planned computational approaches,[7,12–17] sometimes combined with experimental methods, have become the norm in this field as such approaches have increased our understanding of the intricate picture of the protein aggregation process. The scarcity of detailed structural information about (i) the amyloid-ready monomeric states, (ii) the soluble oligomeric ensembles that are believed to be precursors of the full amyloid fibrils, and (iii) the fibrils themselves, is a major drawback in simulation approaches too. In the absence of reliable starting conformations, two main avenues are opened to computational studies. The first is to perform long all-atom and explicit solvent MD simulations starting from pre-formed structures. These are usually produced with docking techniques that rely on presumed high-quality structures of the monomer and low-resolution (e.g. cryoEM) information of the oligomeric ensemble structure. This class of studies then probes the stability of each selected docking conformation and the contribution of environmental factors (such as water, temperature, and pH) to oligomer stability and propagation. The second avenue is to perform all-atom and implicit-solvent MD simulations or coarse-grained simulations of the oligomerization process starting from collections of fluctuating monomers. In addition, bioinformatics approaches that employ searches through large databases of sequences and structures can be applied by themselves or in conjunction with the simulation methods described above to probe the propensity towards aggregation of a proposed sequence.

11.2 Pathways to the Formation of Aggregation Prone Conformations and Mechanisms of Oligomerization

11.2.1 *Formation of Aggregation Prone Conformations*

Gaining an understanding of the molecular details of the series of events that lead to the formation of species that can nucleate and grow into full amyloid fibrils is of the utmost importance for the development of efficient methodologies to prevent disease-related aggregation. Such efforts are somewhat impeded by the large degree of fluctuations and the metastability of the species found along the aggregation pathways. Still, a number of conformational conversion mechanisms have been identified and described at various levels of detail by both experimental and computational investigations. The best-studied mechanism of formation of "amyloid-ready" states was done by the partial unfolding of the native protein structure or by partial folding of the unfolded state of the protein[18] (see Scenario I in Fig. 11.1). This pathway, typical, for example, in transthyretin (TTR), therefore refers to the formation of either folding or unfolding

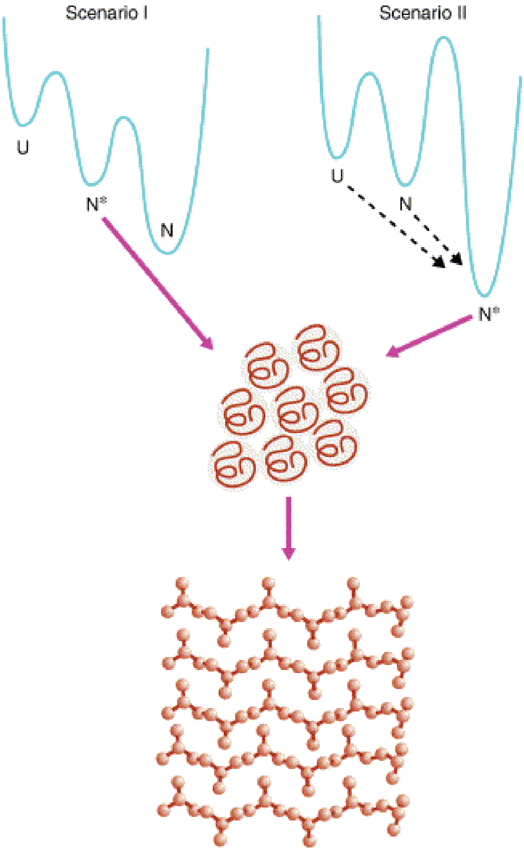


Fig. 11.1 Schematic diagram of the two plausible scenarios of fibril formation based on free energy landscape perspective. According to scenario I, the assembly competent state N* is metastable with respect to the monomeric native state N and is formed due to partial unfolding. In scenario II, N* is formed upon structural conversion either of the native state N (as in prions) or directly from the unfolded state U (as in Aβ-amyloid peptides). In both cases, proteins (or peptides) in N* states must coalescence into larger oligomers capable of growth into fibrils. This figure is reproduced with permission from Fig. 4 in Ref. 73.

intermediates that are assembly-competent structures (N*), and depends on the propensity of the given protein to populate, under normal folding conditions, beside the native basin of attraction of other low-lying basins of attraction. In the case of TTR, extensive experiments[19] have shown that the N* state, which has a higher free energy than the native state N, is formed upon unraveling of the β-strands C and D at the edge of the structure. Using the energy landscape approach to protein folding, a second mechanism for the formation of the amyloid state can also be rationalized. This Scenario II depicted in Fig. 11.1 corresponds to instances where N^* has a lower free energy than N, thus making the folded (functional state) state metastable. This pathway is likely to be found in proteins that require almost complete conformational changes upon the formation of the amyloid-ready state such as the prion protein associated with various TSEs and the Aβ peptides associated with Alzheimer's disease. The Aβ peptides populate mostly loop/coil conformations under infinite dilution conditions. These peptides acquire definite secondary and tertiary structures only upon complex formation with other copies of themselves. If the newly formed oligomeric species are assembly competent, they continue to grow, and the typical cross-β structure of amyloid fibrils is thus formed. The causative agent in TSE diseases is believed to be the aggregated form (PrP^{SC} = scrapie) of the prion protein.[20] The transition to the scrapie form, which is believed to be mostly β-sheet, involves a large conformational change from the native, functional structure, PrP^{C}, which is predominantly α-helical. According to the "protein-only" hypothesis,[5] PrP^{Sc} serves as a template to induce conformational transitions in PrP^{C} that can subsequently be added to PrP^{Sc}. It is believed that the PrP^{C} state is only a kinetic trap, with the disease-related scrapie form, PrP^{Sc}, being the true free-energy minimum state.[21] The transition between the two forms is likely to occur by populating an intermediate state, PrP^{C^*}, which is a species able to undergo transition to the assembly-competent structure and having a lower energy than PrP^{C}. Because of the large degree of structural variation between the PrP^{C} and the PrP^{Sc} conformations, the $PrP^{C} \rightarrow PrP^{C^*}$ transition is also likely to involve the crossing of a high energy barrier. Therefore, in both scenarios the growth kinetics must be initially determined by the "unfolding" barriers separating N^* from either N or U (unfolded state). Based on the energy-landscape perspective for aggregation (Fig. 11.1), this suggests that the free energy of stability may not be a good indicator of fibril growth kinetics. Rather, growth kinetics should correlate with unfolding barriers.

11.2.2 *Mechanisms of Oligomerization*

One of the proposed mechanisms of fibril formation, the nucleated conformational conversion (NCC) model,[22] reproduces many experimental findings. This model, proposed from the study of the assembly kinetics of Sup35 into [PSI$^+$] in *Saccharomyces cerevisiae* combines parts of the templated assembly and nucleation-growth mechanisms. The hallmark of the NCC model[22] is the formation of a critically sized mobile oligomer, in which Sup35 adopts a conformation that may be distinct from its monomeric random coil or the one it adopts in the aggregated state. The formation of a critical nucleus to which other Sup35 can assemble involves a conformational change to states that it adopts in the self-propagating [PSI$^+$].

Fibrillogenesis was proposed to depend mainly on the relative stability of "amyloid-competent states" of the monomer.[23] In other words, peptides that populate predominantly such states form fibrils readily and without passing through any intermediates. A corollary of this proposal is that for peptides found in the amyloid-protected state, the kinetics of protein aggregation must be slow. The main energetic contributions to aggregation are believed to be the van der Waals interactions between side-chains and backbone hydrogen bonds.[24] In addition, shape complementarity between neighboring molecules plays a key role as well.[25] In accord with the NCC mechanism of fibril formation, which requires the existence of a nucleus, a combination of experimental[9] and computational[24] studies indicate that the minimal nucleus seed for fibril formation consists of only a few peptides because larger oligomers do not disassociate quickly due to slow diffusion coefficients. The presence of the aggregation nucleus both facilitates the transition into the cross-β conformation and substantially lowers the free energy barrier of the transition.[26] This suggests an autocatalyzed, nucleation-like mechanism for the formation of β-amyloid. Also, in accord with the NCC model,[22] a number of computational studies studying the formation of fibril by the GNNQQNY peptide.[26,27] revealed that, in the formation of amyloid fibrils, the nucleation step is rate-limiting, while the growth step is fast. In particular, electrostatic interactions of peptide backbone dipoles are found to contribute significantly to the stability of the *β*-amyloid state and water exclusion and interactions of polar side-chains are driving forces of amyloid formation: the cross-*β* conformation is stabilized by burial of polar side-chains and inter-residue hydrogen bonds in the presence of an amyloid-like seed.

The NCC model has received a large degree of support from the time it was first proposed. Still, this is not the only conceivable aggregation mechanism. It is plausible that fibril formation is nucleation-dependent when it occurs only after a lag time that decreases with increasing peptide concentration and increases with temperature. At the same time, fibril formation can appear to be a conformational conversion process consisting of the steps: small amorphous aggregates → *β*-sheets → ordered nucleus → subsequent rapid growth of a small stable fibril or protofilament. Unlike the NCC model, in which fibril growth occurs through the addition of globular multimers to fibril ends, the formation of fibrils can then involve both *β*-sheet elongation, in which the fibril grows by adding individual peptides to the end of each *β*-sheet, and lateral addition, in which the fibril grows by adding already formed *β*-sheets to its side. The initial rate of fibril formation can thus increase with increasing concentration and decrease with increasing temperature. Such a mechanism shares elements with all three proposed standard mechanisms of fibril formation, i.e. templated assembly, nucleated polymerization, and nucleated conformational conversion. However, none of them gives a completely satisfactory description. It has been found, for example, during the investigation of the kinetics of fibril formation of systems containing 48–96 model polyalanine (Ac-KA$_{14}$K-NH2) peptides.[28]

11.2.3 *Applications to the Kinetics of Fibril Formation of Aβ Peptides*

The kinetic model, by which the A*β* peptides associated with Alzheimer's disease aggregate into amyloid fibrils, is believed to follow the NCC model, with a lag-phase of several days. In general, to understand the kinetics of fibril formation, it is necessary to characterize the early events and pathways that lead to the formation of the critical nucleus. In terms of the energy landscape, the structures of N^*, the ensemble of transition state structures, and the conformations of the critical nuclei must be known to fully describe the assembly kinetics. Teplow and coworkers have followed the growth of fibrils for 18 peptides, including $A\beta_{1-40}$ and $A\beta_{1-42}$.[29] They showed that the formation of amyloids is preceded by the transient population of the intermediate oligomeric state with high α-helical content. This is remarkable given that both the monomers and fibrils have little or no α-helical content. An obligatory α-helical intermediate for the formation of fibrillar conformations was found also in simulations of the oligomerization of a collection of three $A\beta_{16-22}$ peptides by Klimov and Thirumalai.[13] Therefore, the formation of oligomers rich in α-helical structure may be a universal mechanism for $A\beta$ peptides, and this α-helical intermediate may well correspond to the mobile oligomer from the NCC aggregation mechanism that has the "wrong" conformation to induce further assembly. Klimov and Thirumalai rationalized the formation of this on-pathway α-helical intermediate using arguments based on confinement and the "minimization of frustration" principle: the initial steps in oligomerization were driven by hydrophobic interactions leading to a reduction of the effective available volume to each $A\beta$ peptide. In the confined space, peptides adopt a α-helical structure similar to the behavior of newly synthesized chains in the ribosomal channel. Because further structural evolution is determined by the requirement to maximize the number of favorable hydrophobic and electrostatic interactions, i.e. that the oligomeric structure must obey the "minimum frustration" principle, this can be achieved only when the $A\beta$ peptides adopt extended β-strand-like conformations.[13]

For the Alzheimer's $A\beta_{10-35}$ peptide, computational studies[30,31] showed that it populates a number of collapsed globular states that are in rapid dynamic equilibrium with each other. This conformational ensemble is dominated by random coil and bend structures

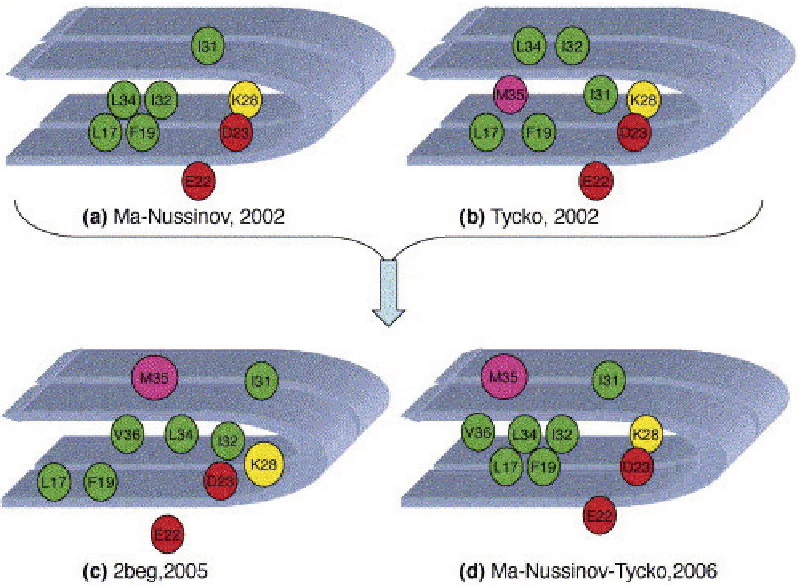


Fig. 11.2 Fibrillar structures of the Aβ_{10-35} peptide. **(a)** The Ma-Nussinov-2002 model is based on extensive MD simulations.[30] The key feature is the salt bridge between D23 and K28. **(b)** The Tycko-2002 model has been proposed on the basis of solid state NMR-derived backbone torsion angles of Aβ_{1-40}.[32] **(c)** The structure of Aβ_{1-42} derived from hydrogen-bonding constraints from quenched hydrogen/deuterium-exchange NMR, and side-chain packing constraints from pair-wise mutagenesis studies (PDB entry 2BEG). **(d)** Experimental support[49] for this model is from solid-state NMR measurements on Aβ_{1-40} fibrils providing critical residue contacts, confirming the Ma-Nussinov-2002 key features. This figure was reproduced with permission from Fig. 2 in Ref. 25.

with insignificant presence of an α-helical or β-sheet structure. Still, both the structure of the peptide as well as that of the resulting fibril are characterized by a salt bridge formed between the side-chains of K28 and D23, as illustrated in Fig. 11.2. This salt bridge was subsequently observed experimentally.[32] The identification of the structure of this fibril using MD simulations, which was later confirmed experimentally, is one of the important contributions of computational approaches to the field of protein aggregation.

11.2.4 *Application to the Early Steps of Prion Proteins Fibril Formation*

In prion proteins, the mechanism of conversion and oligomerization into the amyloid-fibril form has been proposed to follow a templated assembly route.[5] This proposal, which stems from the "protein-only" hypothesis, states that the conformational transition from PrP^{C} to PrP^{Sc} is facilitated by the presence of a pre-formed template consisting of PrP^{Sc} structures. Since the protein-only hypothesis has been proven only in non-mammalian prions,[33] the details of the conformational change in prion proteins are still much investigated by researchers. A study to probe such details for a prion-like peptide was recently performed by Dokholyan and coworkers.[35] They found that at low temperatures these peptides favor α-helical structures, while an increase in temperature drives them to convert into a mainly β-sheet conformation. Most strikingly, during the course of simulations with hexamers, several peptides form a β-sheet that acts as a template to convert an α-helix into a β-strand. This occurred in the absence of any artificial constraints, in perfect agreement with the template-based aggregation scheme for prion proteins.[5] Because the stability of PrP^{C} is due to the C-terminal end (which forms its structural core), the transition to PrP^{C*} requires global unfolding of PrP^{C}.[36] This explains the origin of the high free energy barrier of 20 kcal/mole[21,37] separating PrP^{C} and PrP^{C*}.[7] More importantly, conformational fluctuations that originate in the C-terminal part of H2 are essential in the formation of PrP^{C*}.[10,11] The requirement for the conformational fluctuations of PrP^{C}, needed to populate PrP^{C*}, suggests that the earliest event involves extensive unfolding of the monomeric PrP^{C}. MD simulations designed to study the degree of conformational fluctuations in the various helical segments of mPrP^{C}[15] revealed that H1 remains helical for the duration of the simulation (≈0.09 μs), in agreement with experimental observations.[38–40] In contrast, simulations of peptides encompassing H2 and H3 (together with their connecting loop), including the intact disulfide bond (Cys179-Cys214), showed that residues in the second half of H2, clustered around positions 187–188, have a large conformational flexibility and a non-zero preference for β-strand or coil-like structures.[15] Based on these results, we mapped the plausible structures of the aggregation prone PrP^{C*} (depicted in Fig. 11.3).

11.2.5 *Applications to the Study of Fibril Formation in Polyglutamine Disease-Related Peptides*

Nine human neurodegenerative diseases, including Huntington's disease, are collectively known as "polyglutamine diseases." The time of disease onset and the severity of the symptoms are linked to the increasing number of glutamine repeats when that number exceeds the threshold value of 35–40. Therefore, abnormally long glutamine repeats render their host protein toxic to nerve cells, and all polyglutamine diseases are believed to progress via common molecular mechanisms. A possible mechanism of cell death is that the long sequence of glutamines acquires a shape that prevents the host protein from folding into its functional conformation. A 37-mer polyQ (Q_{37}) peptide populates random coil conformations both at low and at high temperatures, while at intermediate temperatures it adopts a β-strand conformation.[41] The existence of the β-strand conformation is directly correlated with the presence of specific side-chain to backbone hydrogen bonding interactions. In the absence of such interactions, the peptide only populates α-helices in the ground state, and at higher temperatures, the helices melt to form a random coil but no β-strands appear. The authors[41] propose that side-chain to backbone interactions lead to the formation of β-strands by a single polyQ peptide, which is the nucleating structural transition observed in polyQ-peptide aggregation. This is similar to the β-helix nanotube model proposed experimentally.[42] The β-helix acts as the aggregation nucleus because its long-time stability is expected to be sufficient for further propagation of the aggregate. Moreover, this particular secondary structure is acquired only in polyQ segments longer than the critical value found in disease, and as expected, the longer the glutamine tract, the higher is the propensity to form β-helices. The formation of a β-helix from a random coil is accompanied by entropy

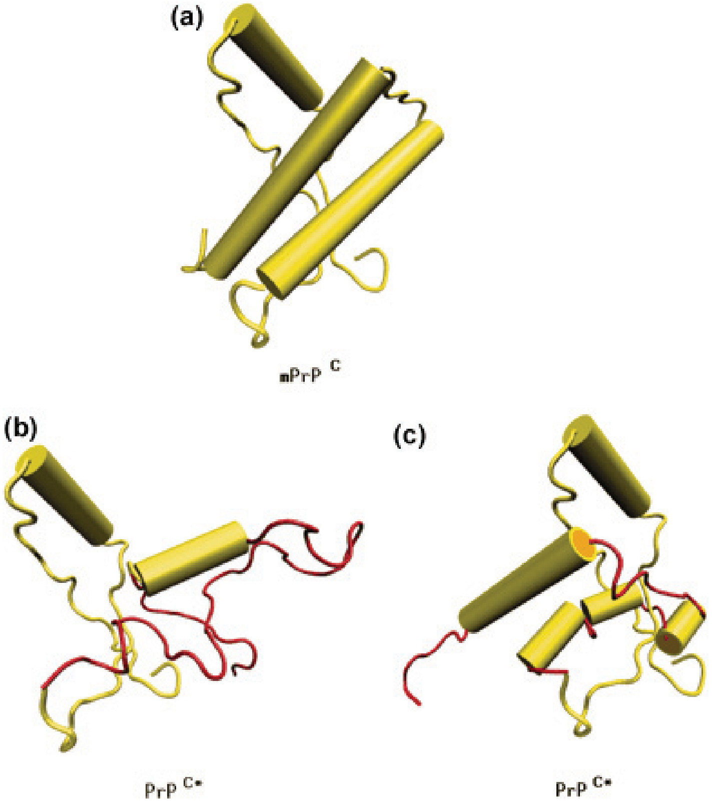


Fig. 11.3 Schematic representation of PrP^{C} ◊ PrP^{C*} transition, where the conformation for PrP^{C} is taken from the PDB file 1ag2. The conformations for PrP^{C*} contain H1 from 1ag2, while the residues encompassing H2+H3 are shown in a conformation (red) reached towards the end of our MD simulations using the NAMD package [3(b)] or the simulations using the MOIL package [3(c)]. The schematic PrP^{C*} structures are representatives from ensembles of fluctuating conformations. In the representative PrP^{C*} structure obtained using NAMD simulations, the H1 region, together with the adjacent loops and the β-strands, and residues (205–212) from H3 retain their original conformations and are therefore depicted with the same color as for PrP^{C}. In the MOIL representative PrP^{C*} structure, the H1 region, together with the adjacent loops and the β-strands, and residues (175–179), (184–188), (193, 194) from H2, and residues (203–218) from H3, retain their original conformations and are therefore depicted with same color as for PrP^{C}. The figures are rotated such that the orientation of H1 is the same in all of them. The figure was reproduced with permission from Fig. 4 in Ref. 15.

loss, leading to a free energy barrier that can account for the lag times observed in experiments of polyQ peptide aggregation. The energy barrier is increased for longer peptides because the enthalpy gain upon *β*-helix formation compensates for the entropy loss in the transition.[41,43,44]

11.3 Self-Association Processes Under Various Cellular Conditions

Self-association routes depend not only on the protein, but also on the specific cellular conditions. For example, the transition of the recombinant prion protein to the scrapie-form is strongly promoted under low concentrations of urea, i.e. under conditions that are conducive to the formation of intermediates. On the other hand, there is no conformational conversion under conditions that favor either the native structure (absence of denaturing agents) or the completely unfolded state (high concentration of denaturing agents). Also PrP^{Sc} accumulates in endosomes of scrapie-infected cells, which have mostly acidic pH values (between 4.0 and 6.0). This finding led to the proposal that the formation of the scrapie form from the normal cellular form of prion proteins, PrP^{C}, is favored under low pH conditions.[19,45] The fact that the conformational conversion in prion proteins is pH-dependent indicates that electrostatic interactions are likely to be involved in the transition. To further probe the effect of strongly acidic conditions (Glu, Asp, Lys, Arg, and protonated His) on prion conformation, molecular dynamics simulations have been employed. In the study by Daggett and coworkers[46] of a number of mammalian prion proteins with known native structure at low pH, it was revealed that the protein exhibits a higher conformational mobility and the sheet-like structure increases both by lengthening of the native *β*-sheet and by addition of a portion of the N-terminus to widen the sheet by two additional strands. The role of more moderate acidic pH conditions, which are more akin to physiological conditions, on the dynamics of the human prion conversion to the scrapie form have been probed using MD simulations.[47] By focusing on the effect of histidine protonation on the conformational behavior of human PrP^{C} globular domain, they found a significant loss of α-helix content under mildly acidic conditions (pH = 4.5, all His positively charged, and Glu and Asp remaining unprotonated), due to the loss of ordered secondary structure in the C-terminal part of the second α-helix and a transient lengthening of the native β-sheet. This study supports the central role played by the C-terminal end of H2 in the conformational transition between the cellular and scrapie forms of the prion protein highlighted in Refs. 7 and 15.

11.3.1 *The Effect of pH on Aggregation Processes*

To probe the influence of the solvent and pH on the formation and stability of different Alzheimer $A\beta_{16-22}$ fragment oligomers, Parinello and coworkers used atomic-detail molecular dynamics simulations with explicit solvent.[48] They found that only large oligomers form a stable β-sheet aggregate, with the minimum nucleus being of the order of eight to 16 peptides. This is due to better hydrophobic contacts and a better shielding of backbone-backbone hydrogen bonds from the solvent in bigger assemblies. This argues in favor of the crucial role played by the solvent (water in this case) in amyloid-fibril formation. Additionally, depending on the stacking interface between the sheets, the simulations reveal straight or twisted structures. Under neutral pH, APLFA, which displays a twisted structure, is more likely to form than PARKVFE, which displays a flat structure, because of the better solvation of the charged Lys and Glu side-chains.[48] By contrast, under acidic or basic pH, where one of the side-chains would be neutralized, this effect must be smaller so that the formation of PARKVFE becomes more likely. The authors propose that the stacking of different interfaces could be related to different fibril morphologies found *in vitro* at different pH, which is in agreement with the suggestion made for the $A\beta_{1-40}$ peptide by Tycko and collaborators.[49]

11.3.2 *The Role of Water in the Oligomerization of Proteins*

Fernandez and coworkers have recently found[50] that some proteins that readily form amyloids have a significant number of backbone hydrogen bonds that are exposed to the solvent, suggesting that these regions have a propensity toward protein interaction and aggregation. This is, for example, the case in the sheep and human C-terminal prion proteins PrP(90–231). The experimentally elucidated 3D structures revealed a large number of under-dehydrated hydrogen bonds (UDHBs) and partially buried water molecules. De Simone and collaborators[51] discovered that regions with a high concentration of UDHBs are structurally more labile. This is due to the fact that UDHBs are backbone hydrogen bonds, which are not protected against water interaction by flanking hydrophobic residue, leading to a less stable packing of hydrogen bonds. Therefore, the loci of these defects on the protein surface are correlated with local destabilization and the favoring of partially unfolded structures with a consequent potential for aggregation.

The fundamental role played by water molecules in the early steps of oligomerization in Aβ-proteins has been probed using all-atom MD simulations in explicit water. Thirumalai and collaborators[52] studied the formation of the intramolecular salt bridge between D23 and K28 in the isolated Aβ_{10-35} monomer. It is known that this loop is formed in the structure of the monomer from the amyloid fibrils formed from long fragments of the amyloid β-protein and ensures that unpaired charges are not buried in the low-dielectric interior. The computed free energy disconnectvity graph shows that the ensemble of compact random coil conformations can be clustered into four basins that are separated by free energy barriers ranging from 0.3 to 2.7 kcal/mol. The extent of solvation of the peptides in the four basins varies greatly, which underscores the dynamical fluctuations in the monomer. These results suggest that the early event in the oligomerization process must be the expulsion of discrete water molecules that facilitates the formation of stable structures driven by interpeptide interactions with an intramolecular D23-K28 salt bridge and an intact VGSN turn. A major conclusion of this work is that discrete water molecules, which solvate charges and facilitate hydrogen bond formation, play a key role in preventing the formation of the D23-K28 salt bridge in the monomer. Therefore, the authors propose that a molecular description of the early events in the oligomerization of Aβ-proteins requires explicit inclusion of water molecules.[52]

11.4 Formation of Soluble Oligomers in the Early Steps of Fibril Formation

Due to the long time scales (minutes to days) involved in protein aggregation, a computational study of complexation processes using a full atomistic description of the chains as well as of the solvent molecules is currently beyond the timescale of nanoseconds for classical all-atom MD simulations. Still, experimental studies revealed that soluble oligomers formed in the early stages of fibril formation are even more pathogenic than the full fibrillar associations.[53] The cytotoxicity of prefibrillar aggregates is, among other factors, dependent on the size of these misfolded oligomers, and occurs according to a universal mechanism. The mechanism of toxicity of protein aggregates remains unclear, but accumulating evidence suggests that it is related to the interaction of protein aggregates and the cell membrane through formation of channels in the membrane.[8] Therefore, the study of the pathways of formation and energy landscapes of these early steps in aggregation is of the utmost importance to the efforts to block it. The advantage is that the formation of structures in such early stages can occur on far shorter timescales (hundreds of nanoseconds up to milliseconds), which are closer to the computationally accessible intervals.

In general, during the initial steps in oligomerization, an increase in the concentration of the chains reduces the melting temperature. This decrease can be attributed to the fact that oligomeric structures are less structured/stable than the native structure of the monomer. Indeed, in multichain systems, free-energy landscapes for folding show an increased preference for misfolded states. As expected, misfolding is accompanied by an increase in inter-protein interactions even if, near the folding temperature, the transition from folded to misfolded chains is entropically driven. The majority of the most probable inter-protein contacts are also native contacts, suggesting that native topology plays a role in early stages of aggregation. Such behavior has been found in a number of lattice model investigations into the origin of the driving force behind the initial steps in oligomerization.[12,54] It has also been suggested[23] that an increase in the relative stability of the β-prone state of the polypeptide changes the aggregation type from disordered into fibril formation with the presence of oligomeric on-pathway intermediates. Further increase in the stability of the β-prone state of the polypeptide leads to fibril formation without intermediates, i.e. according to a downhill pathway. In conclusion, the main difference between functional and pathological fibril formation is in the degree of stability of the β-prone state of the monomer, with very stable monomeric states favoring the formation of functional amyloids according to a downhill pathway scenario.[55] Another characteristic of the early steps of aggregation is that they are dominated by side-chain to side-chain interactions, which allow partial assembly of distinct disordered β-sheets in nonnative registries. As stated above, folding proceeds by concomitant optimization of hydrogen bonding and hydrophobic interactions through amorphous aggregates leading to the formation of multiple early aggregates. These early aggregates with amorphous structure act as building blocks for the nucleus from which rapid growth of fibril can occur and convert either directly (higher probability) or indirectly (lower probability through the second type of topology) to cross-β-sheet-like structures (the first type of topology).[56]

A number of pathways to the formation of fibrils from an ensemble of fluctuating peptide conformations could exist. One such pathway can consist, for example, in direct aggregation to an amyloid fibrillar-like structure, with no evidence of intermediate amorphous states. Along another pathway, the peptides can form amorphous metastable aggregates, which would evolve to amyloid fibril-like structures in agreement with the NCC model.[22] Such a variety of pathways is found indeed in simulations of the process of aggregation in NFGAIL (residues 22–27 of the human islet amyloid polypeptide).[57,58] The authors provide an explanation for the occurrence of fast and slow routes starting from the finding that a slow variation in the density of the peptides is necessary, but not sufficient to avoid amorphous aggregates. A more critical parameter is found in the fact that the initial collapse must not go above a threshold of interchain contacts, i.e. each chain must not be involved in too many connections with the other chains, so that the energy costs for rearrangement and elongation of the chains are low enough. This implies that the acquisition of a β-sheet oligomeric structure must, as stated elsewhere (see, for example, Ref. 52), obey the principle of minimal energetic and topological frustration. In addition, for longer chains, which populate mostly random coil conformations in solution, the free energy barrier between amorphous aggregates and amyloid fibrillar-like structures should increase.[58]

Formation of non-fibrillar soluble oligomers in Aβ has been recently investigated by Thirumalai and collaborators.[59] They monitored the early events that direct the assembly of the amyloidogenic peptide Aβ_{16-22}. Using multiple all-atom MD simulations in water totaling 6.9 μs to probe the dynamics of formation of (Aβ_{16-22})$_n$ by adding a monomer to a preformed (Aβ_{16-22})$_{n-1}$ (n = 4–6) oligomer in which the peptides are arranged in an antiparallel β-sheet conformation, they discovered that the oligomer grows by a two-stage dock-lock mechanism. The first dock stage is rapid (50 ns) and involves a substantial increase in the β-strand content of the monomer from a low starting value. The second phase, the lock stage, is slow and corresponds to rearrangements in the monomeric structure to form in register antiparallel structures. Surprisingly, the simulations reveal also that the mobile structured oligomers undergo large conformational changes in order to accommodate the added monomer. This finding, together with possible arrangements of the hexameric ensemble, are illustrated in Fig. 11.4. Based on the speed of oligomer growth, the authors suggest that the critical nucleus size must exceed six. In addition, stable antiparallel structure formation is found to exceed hundreds of nanoseconds even though frequent inter-peptide collisions occur at the elevated monomer concentrations used in these simulations. In conclusion, the authors propose that the dock-lock mechanism should be a generic mechanism for growth of oligomers of amyloidogenic peptides.[59]

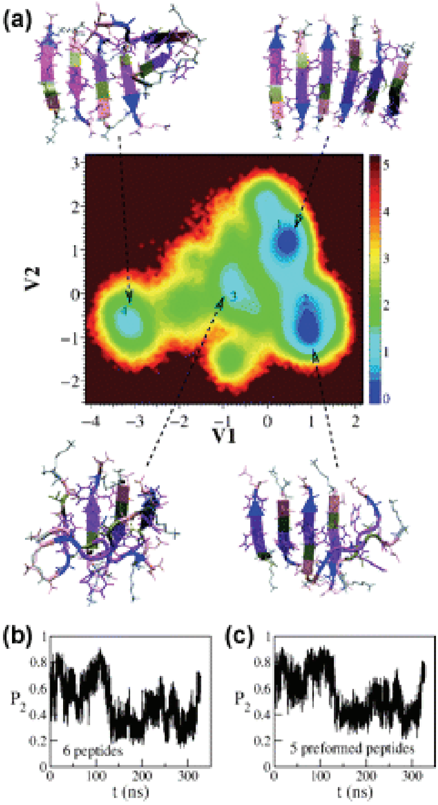


Fig. 11.4 Snapshots of hexameric conformations for $A\beta_{16-22}$ and evidence for the structural re-arrangements of the template to accommodate newly attached peptides.[59] **(a)** Free-energy diagram projected onto the first two principal components, V1 and V2 of the Principal Component Analysis (PCA) for the hexamer. The free-energy scale is given on the right. The structure in the basin labeled 1 shows that the monomers are arranged in antiparallel fashion. The energy-minimized structures from the second basin would also correspond to an ordered hexamer. The presence of other minima could potentially act as kinetic traps that delay oligomerization. **(b)** Time dependence of the changes in the order parameter for the template upon monomer addition. The initial value of P2 for the pentamer exceeds 0.8, which implies shortly (1 ns) the added monomer induces fluctuations in the structured oligomer. **(c)** Dynamics of P2 for the structured pentamer that roughly mirrors (a). The large fluctuations show that the initially ordered pentamer orientationally melts (disorders) to accommodate the added monomer. This figure was reproduced with permission from Fig. 3 in Ref. 59.

11.5 Future Outlook

From a biophysical perspective, there are a number of open problems in the study of amyloid-fibrils formation. Are there common pathways involved in the self-assembly of fibrils? Because of the paucity of the structural description of the intermediates involved in an aggregation process, a definitive answer cannot be currently provided. One avenue that is just starting to be explored for structural studies of amyloid fibrils is to perform single-molecule experiments. These include approaches such as AFM or Laser Optical Tweezers (LOT) pulling experiments or FRET, which have been successful in probing various aspects of the structure and free-energy landscape in a variety of proteins and nucleic acids.[60–63]

Recently, Karsai and coworkers[64] used mechanical manipulation through AFM pulling experiments to probe the structure of $A\beta_{1-42}$ fibrils. They showed that $A\beta_{1-42}$ sheets can be mechanically unzipped from the fibril surface with constant forces in a reversible transition. The measured unzipping force was 23 pN, which is significantly lower than the critical force of unfolding units in a protein tandem of <(100–200) pN. This finding suggests that the inter-sheet contacts in a fibril are significantly weaker than the nonbonded contacts responsible for the folding of a protein chain. In addition, this force value is found to be significantly lower even than that observed earlier for fibrils formed from the $A\beta_{1-40}$ peptide (33 pN).[65] Based on this result, the authors propose that the presence of the two extra residues at the C-terminus end of the $A\beta_{1-42}$ peptide leads to a mechanical destabilization of the fibril.

Despite their importance, single-molecule experiments often cannot easily assign structural information to the observed force peaks due to the intrinsic set-up, which only allows for the measurement of elongation distances versus force or versus time. As a result, it is easy to conceive that, at least in proteins with complicated architectures as is the case in amyloid fibrils, the same elongation of the chain can correspond to the unraveling of a variety of structural elements. To assist with this problem, as well as to provide an in-depth picture of the response of the protein to force, computational approaches are crucial.[66,67] Therefore, further progress in probing details of the amyloid fibrils structure and energy landscape using mechanical manipulation techniques, requires development and application of computational approaches in conjunction with experiments.

The energy landscape perspective, summarized briefly in Fig. 11.1, suggests that multiple scenarios for fibril assembly must exist. Although the generic nucleation and growth governs fibril formation, the details can vary considerably. The microscopic basis for the formation of distinct strains in mammalian prions and in yeast prions remains a mystery. Are these merely associated with the heterogeneous seeds or are there unidentified mechanisms that lead to their growth? What factors may determine the variations in the kinetics of fibril formation for the wild type and the mutants? A tentative proposal is that the kinetics of polymerization is determined by the rate of production of N^* (Fig. 11.1),[68] which in turn is controlled by barriers that separate N and N^*.[7,69] In this scenario, the stability of N plays a secondary role. The generality of this observation has not yet been established. As noted above by the findings in proteins that form functional amyloids, one mechanism used by nature to block toxic aggregation is for proteins to undergo very fast amyloid formation, i.e. without the presence of intermediates. This route is plausible especially in short peptides that have reduced energetic and topological frustration compared to full proteins. But the majority of proteins that undergo toxic aggregation *in vivo* are long polypeptide sequences that are likely to present well-populated intermediates on the pathway to aggregation. The work of Nussinov and collaborators[70] would then suggest that making such intermediates either unstable or highly stable would reduce the amyloid formation propensity of the protein. Their results are interesting, but have been checked only in a peptide, the $A\beta_{25-35}$ fragment.

With the exception of the above-mentioned simulations for prion proteins and a study of the initial stages in the aggregation of TTR,[71] simulations on full proteins are still rare. Therefore, to shed light on the routes to block aggregation in proteins in general, computational approaches performed on longer sequences are critical. Finally, how can one design better therapeutic agents based on enhanced knowledge of the assembly mechanism? This is highly non-trivial, as, for example, in the case of sickle cell disease, viable therapies began to emerge only long after the biophysical aspects of gelation were understood.[72]

References

1. Chiti F, Webster P, Taddei N, *et al.* (1999). Designing conditions for *in vitro* formation of amyloid protofilaments and fibrils. *Proc Natl Acad Sci USA* **96**: 3590–3594.
2. Fandrich M, Forge V, Buder K, Kittler M, *et al.* (2003). Myoglobin forms fibrils by association of unfolded polypeptide segments. *Proc Natl Acad Sci USA* **100**: 15463–15468.
3. Bennett MJ, Sawaya MR, Eisenberg D. (2006). Deposition diseases and domain-swapping. *Structure* **14**: 811–824.
4. Selkoe DJ. (2001). Alzheimer's disease: genes, proteins, and therapy. *Physiol Rev* **81**: 741–766.
5. Prusiner SB. (1998). Prions. *Proc Natl Acad Sci USA* **95**: 13363–13383.
6. Chiti F, Dobson CM. (2006). Protein misfolding, functional amyloid, and human disease. *Ann Rev Biochem* **75**: 333–366.
7. Dima RI, Thirumalai D. (2002). Exploring the propensities of helices in PrP^C to form β-sheet using NMR structures and sequence alignments. *Biophys J* **83**: 1268–1280.
8. Rousseau F, Schymkowitz J, Serrano L. (2006). Protein aggregation and amyloidosis: Confusion of the kinds? *Curr Opin Struct Biol* **16**: 118–126.
9. Nelson R, Sawaya M, Balbirnie M, *et al.* (2005). Structure of the amyloid spine. *Nature* **435**: 773–778.
10. Kuwata K, Li H, Yamada H, *et al.* (2002). Locally disordered conformer of the hamster prion protein: a crucial intermediate to PrP^{Sc}? *Biochem* **41**: 12277–12283.
11. Kuwata K, O.Kamatari K, Akasaka K, James TL. (2004). Slow conformational dynamics in the hamster prion protein. *Biochemistry* **43**: 4439–4446.
12. Dima RI, Thirumalai D. (2002). Exploring protein aggregation and self-propagation using lattice models: phase diagram and kinetics. *Protein Sci* **11**: 1036–1049.
13. Klimov DK, Thirumalai D. (2003). Dissecting the assembly of A beta (16–22) amyloid peptides into antiparallel beta sheets. *Structure* **11**: 295–307.
14. Gsponer J, Haberthur U, Caflisch A. (2003). The role of side-chain interactions in the early steps of aggregation: molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35. *Proc Natl Acad Sci USA* **100**: 5154–5159.
15. Dima RI, Thirumalai D. (2004). Probing the instabilities in the dynamics of helical fragments from mouse PrP^C. *Proc Natl Acad Sci USA* **101**: 15335–15340.

16. Tarus B, Straub JE, Thirumalai D. (2005). Probing the initial stage of aggregation of the Aβ_{10-35}-protein: assessing the propensity for peptide dimerization. *J Mol Biol* **345**: 1141–1156.
17. Buchete NV, Tycko R, Hummer G. (2005). Molecular dynamics simulations of Alzheimer's beta-amyloid protofilaments. *J Mol Biol* **353**: 804–821.
18. Fink AL. (1998). Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des* **3**: R9–R23.
19. Kelly JW. (1998). The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr Opin Struct Biol* **8**: 101–106.
20. Riek R, Hornemann S, Wider G, Billeter M, Glockshuber R, Wuthrich K. (1996). NMR structure of the mouse prion protein domain PrP(121–231). *Nature* **382**: 180–182.
21. Baskakov IV, Legname G, Prusiner SB, Cohen FE. (2001). Folding of prion protein to its native α-helical conformation is under kinetic control. *J Biol Chem* **276**: 19687–19690.
22. Serio TR, Cashikar AG, Kowal AS, *et al.* (2000). Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science* **289**: 1317–1321.
23. Pellarin R, Caflisch A. (2006). Interpreting the aggregation kinetics of amyloid peptides. *J Mol Biol* **360**: 882–892.
24. Zheng J, Ma B, Tsai CJ, Nussinov R. (2006). Structural stability and dynamics of an amyloid-forming peptide GNNQQNY from the yeast prion Sup-35. *Biophys J* **91**: 824–833.
25. Ma B, Nussinov R. (2006). Simulations as analytical tools to understand protein aggregation and predict amyloid conformation. *Curr Opin Chem Biol* **10**: 445–452.
26. Lipfert J, Franklin J, Wu F, Doniach S. (2005). Protein misfolding and amyloid formation for the peptide GNNQQNY from yeast prion protein Sup35: Simulation by reaction path annealing. *J Mol Biol* **349**: 648–658.
27. Tsemekhman K, Goldschmidt L, Eisenberg D, Baker D. (2007). Cooperative hydrogen bonding in amyloid formation. *Protein Sci* **16**: 761–764.
28. Nguyen HD, Hall CK. (2005). Kinetics of fibril formation by polyalanine peptides. *J Biol Chem* **280**: 9074–9082.
29. Kirkitadze MD, Condron MM, Teplow DB. (2001). Identification and characterization of key kinetic intermediates in amyloid β-protein fibrillogenesis. *J Mol Biol* **312**: 1103–1119.
30. Ma B, Nussinov R. (2007). Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers Aβ_{16-22}, Aβ_{16-35}, and Aβ_{10-35}: sequence effects. *Proc Natl Acad Sci USA* **99**: 14126–14131.
31. Baumketner A, Shea J. (2007). The structure of the Alzheimer amyloid Aβ_{10-35} peptide probed through replica-exchange molecular dynamics simulations in explicit solvent. *J Mol Biol* **366**: 275–285.

32. Petkova AT, Ishii Y, Balbach JJ, *et al.* (2002). A structural model for Alzheimer's β-amyloid fibrils based on experimental constraints from solid-state NMR. *Proc Natl Acad Sci USA* **99**: 16742–16747.
33. Sparrer HE, Santoso A, Szoka FCJ, Weissman JS. (2000). Evidence for the prion hypothesis: Induction of the yeast [Psi$^+$] factor by *in vitro*-converted Sup35 protein. *Science* **289**: 595–599.
34. Ding F. LaRocque JJ, Dokholyan NV. (2005). Direct observation of protein folding, aggregation, and a prion-like conformational conversion. *J Biol Chem* **280**: 40235–40240.
36. Hosszu LP, Baxter NJ, Jackson GS, *et al.* (1999). Structural mobility of the human prion protein probed by backbone hydrogen exchange. *Nat Struct Biol* **6**: 740–743.
37. Baskakov IV, Legname G, Baldwin MA, Prusiner SB, Cohen FE. (2002). Pathway complexity of prion protein assembly into amyloid. *J Biol Chem* **277**: 21140–21148.
38. Liu A, Riek R, Zahn R, Hornemann S, Glockshuber R, Wuthrich K. (1999). Peptides and proteins in neurodegenrative diseases: helix propensity of a polypeptide containing helix 1 of the mouse prion protein studied by NMR and CD spectroscopy. *Biopolymers* **51**: 145–152.
39. Speare JO, Rush TS III, Bloom ME, Caughey B. (2003). The role of Helix 1 aspartates and salt bridges in the stability and conversion of prion protein. *J Biol Chem* **278**: 12522–12529.
40. Ziegler J, Sticht H, Marx UC, *et al.* (2003). CD and NMR studies of prion protein (PrP) Helix 1. *J Biol Chem* **278**: 50175–50181.
41. Khare SD, Ding F, Gwanmesia KN, Dokholyan NV. (2005). Molecular origin of polyglutamine aggregation in neurodegenerative diseases. *PLoS Comput Biol* **1**: 231–235.
42. Perutz MF, Finch JT, Berriman J, Lesk A. (2002). Amyloid fibers are water-filled nanotubes. *Proc Natl Acad Sci USA* **99**: 5591–5595.
43. Stork M, Giese A, Kretzschmar HA, Tavan P. (2005). Molecular dynamics simulations indicate a possible role of parallel beta-helices in seeded aggregation of poly-Gln. *Biophys J* **88**: 2442–2451.
44. Merlino A, Esposito L, Vitagliano L. (2006). Polyglutamine repeats and α-helix structure: molecular dynamics study. *Proteins* **63**: 918–927.
45. Hornemann S, Glockshuber R. (1998). A scrapie unfolding intermediate of the prion protein domain PrP(121-231) induced by acidic pH. *Proc Natl Acad Sci USA* **95**: 6010–6014.
46. DeMarco M, Daggett V. (2004). From conversion to aggregation: protofibril formation of the prion protein. *Proc Natl Acad Sci USA* **101**: 2293–2298.
47. Improta ELR, Barone V. (2004). Checking the pH-induced conformational transition of prion protein by molecular dynamics simulations: effect of protonation of histidine residues. *Biophys J* **87**: 3623–3632.

48. Rohrig UF, Laio A, Tantalo N, Parrinello M, Petronzio R. (2006). Stability and structure of oligomers of the Alzheimer peptide A$\beta_{16\text{–}22}$: from the dimer to the 32-mer. *Biophys J* **91**: 3217–3229.
49. Petkova AT, Yau WM, Tycko R. (2006). Experimental constraints on quaternary structure in Alzheimer's β-amyloid fibrils. *Biochemistry* **45**: 498–512.
50. Fernandez A, Kardos J, Scott LR, Goto Y, Berry RS. (2003). Structural defects and the diagnosis of amyloidogenic propensity. *Proc Natl Acad Sci USA* **100**: 6446–6451.
51. De Simone A, Dodson GG, Verma CS, Zagari A, Fraternali F. (2005). Prion and water: tight and dynamical hydration sites have a key role in structural stability. *Proc Natl Acad Sci USA* **102**: 7535–7540.
52. Tarus B, Straub JE, Thirumalai D. (2006). Dynamics of Asp23-Lys28 salt-bridge formation in A$\beta_{10\text{–}35}$ monomers. *J Am Chem Soc* **128**: 16159–16168.
53. Kayed R, Head E, Thompson JL, *et al.* (2003). Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science* **330**: 486–489.
54. Cellmer T, Bratko D, Prausnitz JM, Blanch H. (2005). Protein-folding landscapes in multichain systems. *Proc Natl Acad Sci USA* **102**: 11692–11697.
55. Fowler DM, Koulov AV, Balch WE, Kelly JW. (2007). Functional amyloid — from bacteria to humans. *Trends Biochem Sci* **32**: 217–233.
56. Melquiond A, Mousseau N, Derreumaux P. (2006). Structures of soluble amyloid oligomers from computer simulations. *Proteins* **65**: 180–191.
57. Wu C, Lei H, Duan Y. (2005). Elongation of ordered peptide aggregate of an amyloidogenic hexapeptide NFGAIL observed in molecular dynamics simulations with explicit solvent. *J Am Chem Soc* **127**: 13530–13537.
58. Melquiond A, Gelly J-C, Mousseau N, Derreumaux P. (2007). Probing amyloid fibril formation of the NFGAIL peptide by computer simulations. *J Chem Phys* **126**: 065101–065107.
59. Nguyen PH, Li MS, Stock G, Straub JE, Thirumalai D. (2007). Monomer adds to preformed structured oligomers of Aβ-peptides by a two-stage dock-lock mechanism. *Proc Natl Acad Sci USA* **104**: 111–116.
60. Rief M, Gautel M, Oesterhelt F, Fernandez J, Gaub H. (1997). Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science* **276**: 1109–1112.
61. Rief M, Pascual J, Saraste M, Gaub H. (1999). Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. *J Mol Biol* **286**: 553–561.
62. Onoa B, Dumont S, Liphardt J, *et al.* (2003). Identifying kinetic barriers to mechanical unfolding of the *t*-thermophila ribozyme. *Science* **299**: 1892–1895.
63. Dietz H, Rief M. (2004). Exploring the energy landscape of GFP by single-molecule mechanical experiments. *Proc Natl Acad Sci USA* **101**: 16192–16197.
64. Karsai A, Martonfalvi Z, Nagy A, Grama L, Penke B, Kellermayer M. (2006). Mechanical manipulation of Alzheimer's amyloid A$\beta_{1\text{–}42}$ fibrils. *J Struct Biol* **155**: 316–326.

65. Kellermayer MS, Grama L, Karsai A, *et al.* (2005). Reversible mechanical unzipping of amyloid beta-fibrils. *J Biol Chem* **280**: 8464–8470.
66. Marszalek P, Lu H, Li H, *et al.* (1999). Mechanical unfolding intermediates in titin modules. *Nature* **402**: 100–103.
67. Hyeon C, Dima RI, Thirumalai D. (2006). Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure* **14**: 1633–1645.
68. Ramirez-Alvarado M, Merkel JS, Regan L. (2000). A systematic exploration of the influence of the protein stability on amyloid fibril formation *in vitro*. *Proc Natl Acad Sci USA* **97**: 8979–8984.
69. Hammarstrom P, Jiang X, Hurshman AR, Powers ET, Kelly JW. (2002). Sequence-dependent denaturation energetics: a major determinant in amyloid disease diversity. *Proc Natl Acad Sci USA* **99**: 16427–16432.
70. Ma B, Nussinov R. (2006). The stability of monomeric intermediates controls amyloid formation: $A\beta_{25-35}$ and its N27A mutant. *Biophys J* **90**: 3365–3374.
71. Sorensen J, Hamelberg D, Schiott B, McCammon AJ. (2007). Comparative MD analysis of the stability of transthyretin providing insight into the fibrillation mechanism. *Biopolymers* **86**: 73–82.
72. Eaton WA, Hofrichter J. (1995). The biophysics of sickle cell hydroxyurea therapy. *Science* **268**: 1142–1143.
73. Thirumalai D, Klimov DK, Dima RI. (2003). Emerging ideas on the molecular basis of protein and peptide aggregation. *Curr Opin Struct Biol* **13**: 146–159.
74. Humphrey W, Dalke A, Schulten K. (1996). VMD — Visual Molecular Dynamics. *J Mol Graph* **14**: 33–38.

Chapter 12

Modeling and Simulation of Ion Channels

S. Bernèche*,† and B. Roux‡

12.1 Introduction

Ion channels are intrinsic membrane proteins that have the ability to enable and control the passage of ions across the cell membrane. Of particular interest are the molecular features that are responsible for the three principal functional aspects of ion channels, which are permeation, selectivity, and gating. The availability of high-resolution crystallographic structures, together with the development of detailed atomic models and molecular dynamics (MD) simulations methodologies, provide a unique opportunity to refine our understanding of these systems. Although the complexity of these channels does present a formidable challenge to theoretical studies, even with modern computational resources, it is particularly encouraging to note that many of the recent results from simulations have been consistent with the information emerging from higher resolution structural data (for a recent review, see Ref. 1). This relative success relies for a large part on computational strategies involving free energy simulations.

*Corresponding author.

†Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland. E-Mail: simon.berneche@isb-sib.ch.

‡Center for Integrative Science, University of Chicago, Chicago, IL 60637.

The aim of the present chapter is to present an overview of the approaches used to study the microscopic mechanisms underlying the function of ion channels (permeation, selectivity, and gating).

12.2 Structures of Ion Channels

For more than 50 years, ion channels have been studied using a wide range of biophysical approaches, providing an invaluable amount of functional data.[2] But it is only in 1998, less than 10 years ago, that the first structure of a physiological selective ion channel — the KcsA K^+ channel — was determined at atomic resolution.[3] Until then, the development and application of all-atom simulation techniques to examine fundamental principles governing ion transport relied largely on model systems, such as the small channel-forming peptides gramicidin and alamethicin.[4,5] Proteins such as OmpF porin have also served as useful models to study ion permeation through wide aqueous pores.[1] A few other structures of ion channels or selective transporters have since been solved. However, because the number of available experimental structures remains limited, it is often necessary to construct models using the structural information from homologous proteins or using various constraints deduced indirectly from experiments.

12.2.1 *Available High Resolution Structures*

Determining the three-dimensional structure of membrane proteins to atomic resolution has been, and remains, a great challenge. Though our ability to control the crystallization of membrane proteins remains incomplete, obvious progress has been made during the last decade as testified by the structures listed in Table 12.1. Including ion channels and transporters, about 20 structures are available from 12 different families. The K^+ channel family is particularly well characterized with eight crystallized proteins. As X-ray crystallography becomes a standard technique, it is increasingly used to go beyond the initial elucidation of a protein structure. Thus, many variant structures crystallized under different conditions (e.g. mutants,

Table 12.1 Available High-Resolution Structures of Ion Channels and Transporters

Description	PDB Codes	References
Channels:		
KcsA potassium channel: H^+ gated, *Streptomyces lividans*, 3.2 Å.	1BL8	Doyle *et al.* (1998)[3]
KcsA potassium channel: H^+ gated, *Streptomyces lividans*, Fab complex, 2.0 Å.[a]	1K4C 1K4D	Zhou *et al.* (2001)[6]
NaK channel: Bacillus cereus, Na^+ complex, 2.4 Å; K^+ complex, 2.8 Å.	2AHY 2AHZ	Shi *et al.* (2006)[7]
MthK potassium channel: Ca^{2+} gated, *Methanobacterium thermoautotrophicum*, 3.3 Å.	1LNQ	Jiang *et al.* (2002)[8]
KvAP voltage-gated potassium channel: *Aeropyrum pernix*, full-length channel, 3.2 Å; voltage sensor domain, 1.9 Å.	1ORQ 1ORS	Jiang *et al.* (2003)[9]
KirBac1.1 inward-rectifier potassium channel: *Burkholderia pseudomallei*, closed state, 3.65 Å.	1P7B	Kuo *et al.* (2003)[10]
KirBac3.1 inward-rectifier potassium channel: *Magnetospirillum magnetotacticum*, intermediate state 1, 2.60 Å; intermediate state 2, 2.85 Å.	1XL4 1XL6	Gulbis *et al.* (2004) *To be published.*
Kir3.1 inward-rectifier potassium channel: prokaryotic chimera expressed in *Escherichia coli*, 2.2 Å	2QKS	Nishida *et al.* (2007)[11]
Kv1.2 voltage-gated potassium channel: *Rattus norvegicus* (expressed in Pichia pastoris), 2.9 Å.	2A79	Long *et al.* (2005)[12]
Kv1.2/Kv2.1 voltage-gated potassium channel chimera: *Rattus norvegicus* (expressed in *Pichia pastoris*), 2.4 Å (with resolved lipids)	2R9R	Long *et al.* (2007)[12a]
ASIC1 acid-sensing ion channel: Gallus gallus, 1.9 Å	2QTS	Jasti *et al.* (2007)[12b]

(*Continued*)

Table 12.1 (*Continued*)

Description	PDB Codes	References
Nicotinic Acetylcholine Receptor Pore: *Torpedo marmorata*, electron diffraction, 4.0 Å.	1OED 2BG9	Miyazawa *et al.* (2003)[13] Unwin (2005)[14]
MscL mechanosensitive channel: *Mycobacterium tuberculosis*, 3.5 Å.	2OAR	Chang *et al.* (1998)[15]
MscS voltage-modulated mechanosensitive channel: *Escherichia coli*, 3.7 Å.	2OAU	Bass *et al.* (2003)[16]
Transporters:		
CorA Mg^{2+} Transporter: *Thermotoga maritima*, 3.9 Å	2BBJ 2IUB	Lunin *et al.* (2006)[16a] Eshaghi *et al.* (2006)[16b]
MgtE Mg^{2+} Transporter: *Thermus thermophilus*, 3.5 Å	2YVX	Hattori *et al.* (2007)[16c]
ClC Cl^-/H^+ exchanger (formerly ClC chloride channel): *Salmonella typhimurium*, 3.0 Å; *Eschericia coli*, 3.5 Å, 2.51 Å.[a]	1KPL 1KPK 1OTS	Dutzler *et al.* (2002)[17] Dutzler *et al.* (2003)[18]
CLC-ec1 Cl^-/H^+ exchanger: *Escherichia coli*, 3.2 Å.[a]	2FEE	Accardi *et al.* (2005)[19]
NhaA Na^+/H^+ exchanger: *Escherichia coli*, 3.45 Å.	1ZCD	Hunte *et al.* (2005)[20]
Calcium ATPase: Rabbit sarcoplasmic reticulum E1 state with bound calcium, 2.4 Å[a]	1SU4	Toyoshima *et al.* (2000)[20a]
Na,K-ATPase: Pig Kidney, 3.5 Å	3B8E	Morth *et al.* (2007)[20b]
Plasma Membrane H^+-ATPase: *Arabidopsis thaliana*, 3.6 Å	3B8C	Pedersen *et al.* (2007)[20c]

All proteins were solved by X-ray crystallography, unless otherwise specified.

See the following websites for a complete list of available structures of membrane proteins: http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html, http://www.mpdb.ul.ie/, or http://pdbtm.enzim.hu.

[a] Different variants are available for these proteins.

ionic concentration, blockers …) are now available for a given channel. Combined with the tremendous amount of available electrophysiological data, this offers a fertile ground for theoretical studies aimed at bridging structural and functional experiments.

12.2.2 *Homology and Knowledge-based Modeling*

Considering the limited number of ion channels of known structure and the challenge that it represents to determine the structure of new ones, computer modeling based on sequence similarity ("homology modeling") is often an obligatory route. Of course, the general limitations of homology modeling apply to ion channels as well (see Chapter 1 and Ref. 21). However, the abundance of functional and biophysical experimental data offers the opportunity to improve the accuracy of the models (see below). Modeling of ion channels can also benefit from the spatial and topological constraints imposed by the membrane environment. For example, recent progress in the development of hydrophobicity scales helps to better define the secondary structure of ion channels in reference to the cellular membrane's hydrophobic core[22] (see the following web page from Prof. Stephen White's laboratory for useful tools: http://blanco.biomol.uci.edu).

Whether it is possible to transfer information between channels from different families or between bacterial and eukaryotic counterparts remains an unresolved question. For example, it is likely that ion channels from the large and important voltage-gated family, which includes K^+, Na^+, as well as Ca^{2+} channels, share common structural elements. On the other hand, while the pore region of K^+ channels is highly conserved, it is most probable that sequence variations affect their function more specifically. Assessing the accuracy of an ion channel model is difficult.

Whenever possible, it remains preferable to perform all-atom simulations based on high-resolution experimental structures. Homology models are nonetheless well-suited for some type of simplified calculations (e.g. based on continuum electrostatic approximations) and useful for designing structural or functional experiments.

12.2.3 *Modeling the Voltage-sensor of a Kv Channel*

In view of the paucity of structural information about ion channels, it is highly desirable to make full use of all available experimental information from homologous channels. A critical examination of models of the Shaker K^+ channel constructed prior to the X-ray structure of

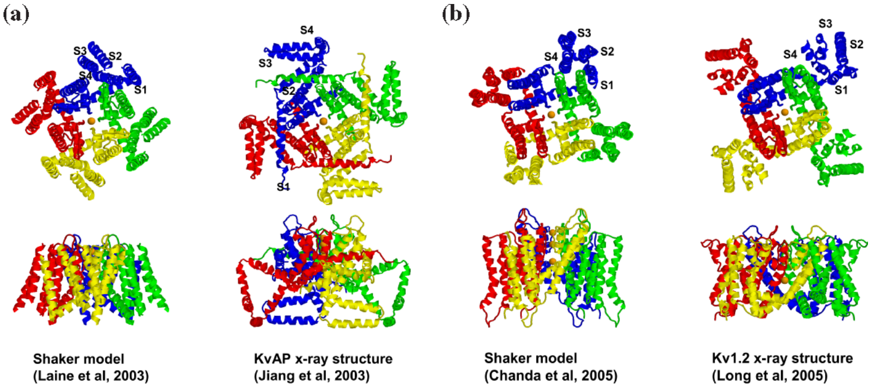


Fig. 12.1 Comparison of models of the Shaker channel with the X-ray structures of the KvAP and Kv1.2 channels.

Kv1.2 shows that this can be done with some confidence. The various models are compared with X-ray structures in Fig. 12.1. Such models are generated by combining a broad range of experimental data from biochemistry, electrophysiology, and X-ray crystallography. For example, the pore domain of an early model of *Shaker* by Laine *et al.*,[23] generated prior to the X-ray structure of the KvAP channel,[9] was based on the X-ray structure of the MthK channel.[8] Proximity of residues in helices S2 and S3 were imposed on the basis of a Mg^{2+} binding site engineered in the *eag* Kv channel.[24] Proximity of residues in helices S4 and S5 in the pore domain were imposed on the basis of a histidine-Zn^{2+}-histidine bridge engineered in the *Shaker* Kv channel.[23] In addition, the amino acid positions along helices S1 and S2 identified as functionally tolerant to tryptophan substitutions were constrained to be lipid-exposed.[25,26]

A subsequent model of *Shaker* by Chanda *et al.*[27] incorporated additional structural elements from the X-ray structure of the KvAP channel (i.e. the pore domain and the voltage sensor module).[9] Though there are clearly some structural differences, the main structural features of the model from Chanda *et al.*[27] are in excellent agreement with the X-ray structure of the Kv1.2 channel. As predicted, the voltage sensor is formed by a bundle of four anti-parallel transmembrane helices, S1-S4, each with their N- and C-terminal ends exposed alternatively to the intra and extracellular solution. The transmembrane helices forming the voltage sensor are packed counterclockwise, and the voltage sensor makes contact with the adjacent subunit in the clockwise direction.[23] The approximate models captured the main topological features of the Kv1.2 structure. This structure shows that about 75 to 80% of the total molecular surface area of S4 is covered by S1-S3 and by the contact with the S5 helix in the pore domain. Only the remaining 20 to 25% is exposed to lipids. In comparison, the fraction of the surface of S4 exposed to lipids was ~12% in the *Shaker* models by Laine *et al.*[23] and ~22% in the model by Chanda *et al.*[27] The *Z*-position of the Cα of the first arginine in S4 (R294 in Kv1.2 and R362 in *Shaker*) is at 12.7 Å from the center of the bilayer in Kv1.2,[28] and is at 11.1 Å and 12.5 Å in the *Shaker* models by Laine *et al.*[23] and Chanda *et al.*[27] respectively. The excellent agreement between the X-ray structure and results from numerous functional and biophysical experiments considerably strengthens the growing consensus about voltage-gated K^+ channels. It also increases confidence in the modeling of membrane proteins based on a wide range of biophysical data.

12.3 Explicit Membrane System

To take advantage of all the information that the high-resolution structure of an ion channel can provide, it is essential to model its environment with all atomic details. With the currently available computational power, one can routinely work with systems containing up to about 100 000 atoms, usually enough to contain the alpha-subunit of a channel and surrounding lipids and bulk solvent. A system of that size would allow for reasonably long simulations, as well as for demanding free energy calculations, and eventually for some computer experiments with different setups or mutants. While expensive in terms of computer time, such systems can be simulated without being at the limit of usual computational resources. Larger systems would most probably be used for straight MD simulation purpose only. As the accuracy of free energy calculations rely on the sampling of relevant conformations of the molecular system, truncated systems requiring restraints to maintain the integrity of the system should be avoided whenever possible.

12.3.1 *Assembling an Ion Channel and Membrane System*

The setup of an integral membrane protein atomic system for simulations can follow one of two commonly used approaches. One is to start with a pre-equilibrated pure lipid bilayer from which a number of lipids are removed to make space for the protein to be inserted.[29] Alternatively, one can build the lipid bilayer around the membrane protein.[30] In this approach, Lennard-Jones (LJ) spheres are first equilibrated on two planes to simulate the effective packing of the lipid head-groups around the channel protein (see Fig. 12.2). These planes are positioned in reference to the protein's aromatic residues, which are usually found at the lipid-water interface. The acyl chain length of

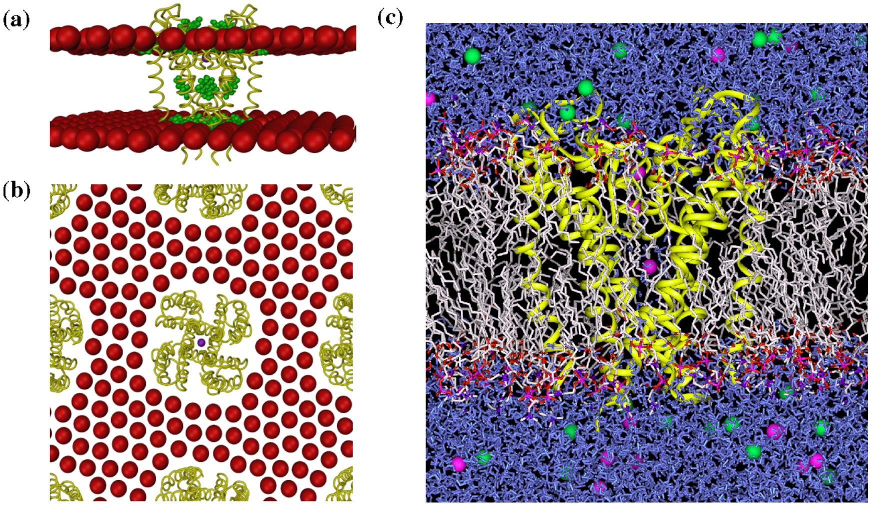


Fig. 12.2 Explicit membrane system of the KcsA K^+ channel: Side **(a)** and top view **(b)** of the initial setup phase with Lennard-Jones spheres mimicking the polar head of lipids. In (a), aromatic residues used as reference to setup the membrane bilayer are represented in green. **(c)** Complete system with the KcsA channel, 112 DPPC lipid molecules, and water bulk containing 12 K^+ and 23 Cl^- (these ions neutralize the system and represent a salt concentration of ~150 mM).

the lipids is chosen so that it matches the distance between the two planes, effectively corresponding to the hydrophobic portion of the membrane. After equilibration, the LJ spheres are replaced by explicit lipid-molecules randomly taken from a library of pre-equilibrated and pre-hydrated lipids. The initial configurations are then refined by rigid body rotation and translation, followed by energy minimization.

For both approaches, the number of lipids on each leaflet has to be adequately calculated to complement the surface area of the protein, which might be different on each side of the bilayer. The area per lipid ratio is unfortunately not well-defined. It varies significantly with lipid type, and seems to be highly dependent on experimental conditions. For example, values from 56 to 72 Å^2 per molecule have been reported for fluid phase DPPC bilayers at 50°C.[31] For the setup of microscopic molecular systems, an area between 59 and 64 Å^2 per molecule is usually assumed for phosphatidylcholine lipids.[32,33] In some simulation work, octane molecules replace lipids, effectively mimicking the hydrophobic core of the membrane but not its densely packed polar-head region. While it does not represent an experimentally stable system, the approach has yielded coherent results for studies focusing on the channel's pore.[34] Membrane proteins have also been successfully simulated in detergent micelles, allowing for comparison with experimental measurements performed under these conditions (for example, see Ref. 35).

In all cases, the system should be equilibrated with respect to the channel protein, i.e. the structure of the protein being the given experimental data, it is important to preserve its integrity. In that order, the membrane system can be equilibrated with slowly decreasing restraints maintaining the different components in their relative positions, using slightly stronger restraints on the backbone of the ion channel than other parts of the membrane system. At the end of this equilibration period, the system is freed of all restraints to perform the actual simulation. (Detailed computational protocols for membrane system setup, including scripts based on the CHARMM biomolecular simulation program,[36] can be found on the following web page: http://thallium.bsd.uchicago.edu/RouxLab, http://www.charmm-gui.org, and in the CHARMM distribution documentation.)

12.3.2 *Periodic Boundary Condition: Advantages and Caveats*

For simulations of membrane proteins, it is common to use periodic boundary conditions (PBC). It has the advantages of mimicking the extent of a cellular lipid membrane with no edges. It also allows for the use of Particle Mesh Ewald (PME) summation to take in account all electrostatic interactions,[37] which obviously play a fundamental role in ion permeation and have effects over long distances.[38] However, PBC/PME may induce artificial ordering that enhances the systems stability.[39] Modeling a finite concentration of ions in the bulk solution, and adjusting the number of added charges so that the system is neutral, can help reduce such undesirable artifacts arising from long-range dipole-dipole interactions. By the same occasion, the system is made closer to the physiological conditions by incorporating some salt concentration.

An important theoretical issue concerns the lateral pressure that should be maintained on a membrane system.[40] Because membrane properties are usually better reproduced in these conditions (see below), it is common to keep the area of the membrane constant, and allow the perpendicular dimension to fluctuate in response to an isobaric-isothermal thermodynamic ensemble. However, considering that many ion channels are sensitive to membrane pressure, it is highly desirable to have better control over the applied lateral tension. Further studies will be necessary to have a better understanding of the impact of lipid bilayer surface tension on membrane protein simulations. Constant development of the lipid force field should also help improve the situation.[41,42]

12.3.3 *Force Field Limitations*

Significant progress is being made in the development of atomic force fields for describing membrane bilayer systems empirically. Parameters for lipid molecules are optimized to reproduce structural data of hydrated phospholipid bilayers as obtained from neutron and X-ray diffraction experiments.[41] Lipid force field parameters remain perfectible, and simulations of a pure membrane system usually show better agreement with experimentally measured properties when performed in the NPTA (constant number of atoms, pressure, temperature, area/lipid) than in the more standard NPT ensemble.[42] As always, when defining a set of empirical parameters, compromises need to be found. For example, the accuracy of the CHARMM27 force field varies for ethane, propane, butane, and hexadecane, illustrating that parameters for extended-chain *n*-alkanes cannot be directly transferred from the short-chain *n*-alkanes. It could then be appropriate to optimize alkane parameters for lipid simulations based on long-chain *n*-alkanes. Such an approach, however, makes it difficult to maintain compatibility with hydrocarbons on other molecules when performing simulations of heterogeneous systems (e.g. protein-lipid systems). While more studies are needed to rigorously address this compatibility issue, the CHARMM27 force field sacrifices the long-chain *n*-alkane pure-solvent properties in order to maintain the overall consistency of the force field.[41] As only a few simulations have been performed with mixtures of different kinds of lipids,[43–45] further studies will be necessary to assess the force field compatibility in this case as well.

To study the function of an ion channel, it is important to pay special attention to the parameters describing the interactions involving ions. The selectivity and conductance of an ion channel results from a delicate balance of very strong microsopic interactions, the large energetic loss of dehydration at the entry of the pore being roughly compensated by coordination with pore lining residues. Gas phase experiments on model systems provide the most direct information concerning these individual microscopic interactions.[46] High-level quantum-mechanical *ab initio* calculations can also be used to supplement the (often scarce) information available from experiments.[47] Since selectivity and conductance are primarily governed by relative free energies, it is essential to also consider thermodynamics properties in the parametrization of the potential function (i.e. the force field).

For simulations of K^+ channels, the most relevant microscopic interactions are between ions, water molecules, and the carbonyl moieties of the protein backbone. Solvation free energy of ions in liquid water and liquid N-methylacetamide (NMA), a model of the backbone of proteins, are thus particularly important for calibrating a proper potential function. In the case of ions solvated in water, it is possible to reproduce both the microsopic interactions and the solvation free energy of ions with the current biomolecular potential functions.[48,49] In contrast, MD free energy calculations indicate that it is very difficult to reproduce both the cation-NMA microscopic energy and the solvation free energy in liquid NMA (see Table 12.2).[50] In MD simulations of K^+ channels, the solvation free energy of the ions in liquid NMA (ΔG_{FF}) impacts directly on the stability of ions in the selectivity filter of the channel: if ΔG_{FF} is higher than the experimental value, ions are artificially trapped in the selectivity filter; if it is smaller, ions are expelled. In general, it ought to be possible to calibrate potential function to reproduce solvation free energies by adjusting the parameters. For example, modifying the Lennard-Jones interaction distance of the K^+-carbonyl oxygen pairs, such as to reduce or increase the microscopic K^+-NMA interaction energy can be used to modulate the resulting solvation free energy of K^+ in liquid NMA.

Clearly, if the potential function were an exact representation of the Born-Oppenheimer energy surface, success in reproducing the microscopic interactions would automatically lead to accurate

Table 12.2 Interaction Energy and Solvation Free Energy (absolute values) for K^+ in Water and N-methylacetamide (NMA) as Represented by Different Force Fields.[a]

	Interaction Energy (kcal/mol)		**Solvation Free Energy (kcal/mol)**	
Force field	**Water**	**NMA**	**Water**	**NMA (Charging Only)[b]**
Experimental	17.9	28.3-32.3	79.3	~80-82[c]
AMBER94/TIP3	18.2	23.7	80.9	81.8
CHARMM27/TIP3	18.9	24.2	81.5	89.2
CHARMM27/TIP3 (modified)	18.9	21.6	81.5	82.0
GROMOS87/SPC	17.6	16.6	77.2	71.6

[a] All values taken from Ref. 51.

[b] A non-polar contribution of about 2 kcal/mol should be subtracted to obtain the total solvation free energy.

[c] Approximation based on data from other liquid amides.[52]

thermodynamic properties. But current biomolecular potential functions try to account for many-body polarization effects in an average way using an effective parametrization of the atomic partial charges. Because of this approximation, the optimal parametrization is the result of a compromise between an accurate representation of the microscopic energies and bulk solvation properties. (This is true for both the CHARMM and AMBER force fields reported in Table 12.2.) Nevertheless it may be hoped that such potential functions can yield meaningful results of semi-quantitative accuracy (see also Section 12.8).

12.4 Permeation and Conductance

The subtleties of an ion channel function are found in how its conductance varies in response to the transmembrane electro-chemical potential. The conductance of the channel is governed by intrinsic properties of the permeation pore and by gating events, which will be discussed in a following section. Ultimately, for a complete description of ion permeation, it is of fundamental importance to establish a direct link between electrophysiological data and the crystallographic protein structure by reproducing the I(V) relation of a given channel. At the present time, this cannot be done using simple "brute force" MD simulations, as the time-scale to observe the translocation of a single ion is on the order of a typical MD trajectory. Despite the steady increase in computer power, the direct simulation of ionic fluxes across a selective biological channel with all-atom MD remains computationally prohibitive; it might, however, be feasible in a not too distant future.[53] Designing a theoretical framework able to rigorously extend the simulation time-scale to calculate ion fluxes is one of the most important goals in computational studies of ion channels.[54]

12.4.1 *Transmembrane Voltage as an Analytical Function*

At the microscopic level, the transmembrane electrostatic potential arises from a small charge imbalance distributed in the neighborhood of the membrane/solution interface. The net charge per area for a transmembrane potential of 100 mV corresponds roughly to only one atomic unit charge per surface of 130•130 Å^2. Controlling the transmembrane potential through explicit ionic concentrations would thus require molecular systems much larger then those usually seen. A potential created from an explicit ion concentration asymmetry was simulated by Woolf and co-workers[33] using a double bilayer membrane system. While this approach provides the most realistic realization of a transmembrane potential, from a practical point of view, it is more advantageous to rely on an analytical function describing the electrostatic potential. Simulations have been performed with a linear function extending across the whole simulation cell along the axis perpendicular to the membrane plane.[53,55]

The actual transmembrane potential along the pore of a channel can be modeled quite effectively by using continuum electrostatic approximations. Specifically, it can be calculated by using a modified Poisson-Boltzmann (PB) theory, in which the intra- and extracellular bulk regions are kept in equilibrium with electrodes at a potential difference of V_{mp}, yielding the PB-Voltage equation:[56]

$$\begin{aligned}
&\nabla\cdot\left[\varepsilon(\mathbf{r})\nabla\phi_{\mathrm{mp}}(\mathbf{r})\right]=0 && \text{Pore region}\\
&\nabla\cdot\left[\varepsilon(\mathbf{r})\nabla\phi_{\mathrm{mp}}(\mathbf{r})\right]-\kappa^2(\mathbf{r})\left[\phi_{\mathrm{mp}}(\mathbf{r})\right]=0 && \text{Bulk region, side I}\\
&\nabla\cdot\left[\varepsilon(\mathbf{r})\nabla\phi_{\mathrm{mp}}(\mathbf{r})\right]-\kappa^2(\mathbf{r})\left[\phi_{\mathrm{mp}}(\mathbf{r})-V_{\mathrm{mp}}\right]=0 && \text{Bulk region, side II}
\end{aligned}\tag{12.1}$$

where $\varepsilon(\mathbf{r})$ and $k(\mathbf{r})$ are the space-dependent dielectric constant and Debye-Hückel screening factor. It should be noted that the protein and ion charges must be formally turned off in this calculation, as the transmembrane potential is not intended to include the electrostatic field produced by the fixed charges of the protein.[a] Figure 12.3(a) illustrates the transmembrane potential through the KcsA channel calculated using Equation 12.1. Note the difference between the open and close conformation of the main gate, which modifies the dielectric environment in the inner vestibule. Because of the irregular geometry of the channel with its high dielectric wide aqueous region

[a] The formalism of Equation 12.1 is implemented in the PBEQ module[57] of the CHARMM program.[36]

and its narrow selectivity filter, the calculated membrane potential differs markedly from the naive linear potential, which would arise across a planar membrane slab.

12.4.2 *Potential of Mean Force*

As mentioned above, studying ion permeation using brute force MD remains prohibitive. To circumvent those difficulties and characterize the ion conduction mechanism quantitatively, one can compute the free energy profile, or potential of mean force (PMF), governing the elementary microscopic steps of ion translocation in the pore. When the reaction coordinate is a simple Cartesian coordinate, e.g. the position z of an ion along the channel axis, the PMF may be calculated by integrating the reversible work done by the mean force $\langle F(z) \rangle$ acting on the ion in the z-direction:

$$W(z) = W(z_0) - \int_{z_0}^{z} dz' \langle F(z') \rangle \tag{12.2}$$

One advantage of this formulation is that the mean force can be decomposed linearly into a sum of contributions.[4,58] A related approach is the adaptive biasing force (ABF) method, which estimates the PMF along the reaction coordinate z by averaging the instantaneous force along z and canceling it by an adaptive biasing force (that will eventually correspond to the PMF). For a biasing force that corresponds exactly to the PMF, coordinate z has a purely diffusive motion.[59,60,b]

Another approach to compute the PMF is the "umbrella sampling" technique.[62] In this method, the microscopic system of interest is simulated in the presence of an artificial biasing window potential, $u_i(z)$, introduced to enhance the sampling in the vicinity of a chosen value z_i. Typically, the biasing potential serves to confine the variations of the coordinate z within a small interval around some prescribed value z_i, helping to achieve a more efficient configurational sampling in this region (this is the reason why the biasing potential is called a

[b] The ABF method is implemented in the NAMD software package.[60,61]

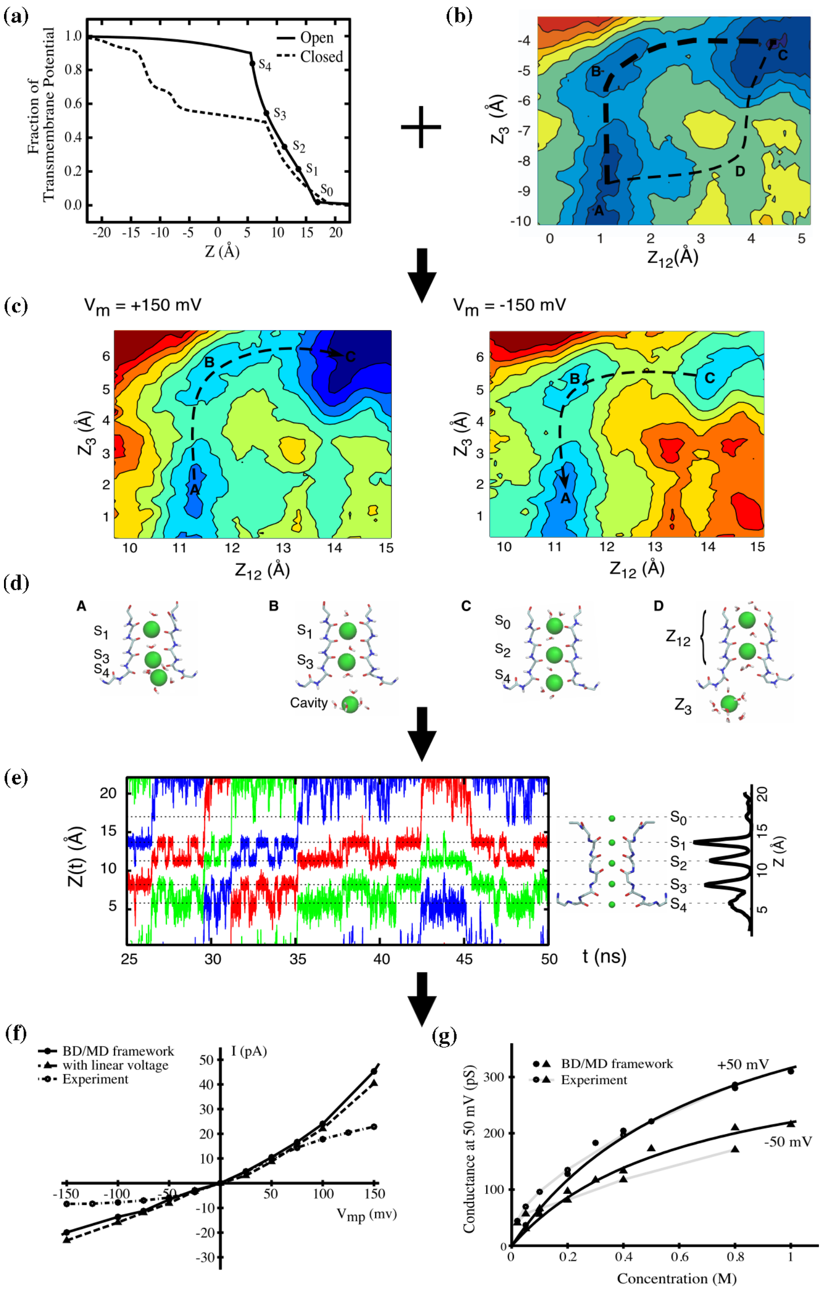


Fig. 12.3 Hierarchical PMF/BD simulation framework **(a)** Transmembrane potential profile along the pore of the KcsA channel **(b)** Equilibrium PMF describing

window potential). For example, a reasonable choice to produce the biased ensembles, though not the unique one, is to use harmonic functions of the form, $u_i(z) = \frac{1}{2}k(z - z_i)^2$, centered on successive values of z_i. The piece of unbiased PMF from the *i*th window is,

$$W_i(z) = F_i - k_B T \ln\left[\langle \rho(z) \rangle_{(i)}^{(biased)}\right] - u_i(z) \tag{12.3}$$

where $\langle \rho(z) \rangle_{(i)}^{(biased)}$ is the biased histogram from the i-th simulation, and F_i is an undetermined free energy constant. To obtain the complete PMF, the data from several windows have to be combined together and the bias introduced by the constraining potentials has to be removed (i.e. the different F_i have to be determined). The most efficient procedure to do this is the Weighted Histogram Analysis Method (WHAM).[63] One of the main advantages of WHAM is that it can be easily extended to treat the case of a PMF depending on more than one variable.[62]

The PMF governing ion permeation in the selectivity filter of the KcsA K^+ channel was calculated using an umbrella sampling simulation composed of 312 windows; a projection of the three-dimensional PMF (one dimension for each ion involved) is presented in Fig. 12.3(b). It shows that ion conduction involves transitions between two main states with, respectively, two and three K^+ ions occupying the selectivity filter, according to a process reminiscent of the "knock-on" mechanism proposed by Hodgkin and Keynes. The largest free energy barrier is on the order of 2 to 3 kcal/mol, implying that the ion conduction process is diffusion limited.[38]

ion permeation in the selectivity filter of KcsA. **(c)** The total multi-ion free-energy profile $W_{tot}(Z_1, Z_2, Z_3)$, including the equilibrium PMF calculated from MD and a transmembrane voltage of 150 mV. **(d)** Principal ion occupancy states identified on the different PMFs by the letters A, B, C, or D. **(e)** BD trajectory generated with an applied membrane potential +50 mV and under symmetric conditions of K^+ concentration. The Z(t) of ions in the system is alternatively plotted in blue, red, and green for the sake of clarity. The relative ion density along the pore is shown in relation to the different binding sites. **(f)** I–V relation calculated from BD simulations under symmetric conditions and K^+ concentration of 400 mM. **(g)** Conductance of the KcsA at 50 mV as a function of permeant ion concentration.

12.4.3 *Hierarchal PMF/BD Framework*

A very attractive computational approach for simulating ion permeation over long timescales, without having to treat a system in all atomic details explicitly, is Brownian dynamics (BD).[54] In its simplest one-dimensional form, the random BD trajectory of the permeating ions along the pore axis is generated by integrating the stochastic equations of motion:[64]

$$\dot{z}_i(t) = -\frac{D(z_i)}{k_B T}\frac{\partial W_{\mathrm{tot}}}{\partial z_i} + \zeta_i(t) \tag{12.4}$$

where z_i and $\dot{z}_i$ are the position and velocity of the ith ion along the channel axis, $D(z_i)$ is the space dependent diffusion constant, $\zeta_i(t)$ is a random Gaussian noise, and W_{tot} is an effective potential energy function. As embodied by Equation 12.4, the construction of a BD model requires very specific input quantities such as the ion diffusion constant $D(z)$ and the total PMF W_{tot}, which serve as the fundamental "microscopic ingredients" of the theory.

If the input quantities are treated as free adjustable parameters to fit experimental data, then BD can be utilized as a phenomenological framework to extract information about fundamental microscopic quantities from experimental data. Alternatively, the theory may be anchored tightly to the molecular reality of ion channels if all the fundamental input quantities are rigorously extracted from calculations based on detailed atomic models.[c]

In the case of the KcsA channel containing no more than three ions in its selectivity filter at a given time, the total PMF, W_{tot}, is expressed as:[56]

$$W_{\mathrm{tot}}(z_1, z_2, z_3) = W_{eq}(z_1, z_2, z_3) + \sum_{i=1}^{3} q\phi_{mp}(z_i) \tag{12.5}$$

[c] Such a strategy finds its roots in statistical mechanical theories of nonequilibrium transport phenomena in which dissipative equations of motion are derived for a reduced set of degrees of freedom while rigorously projecting out the dynamics of the rest of the system (see Ref. 65 and references therein).

where W_{eq} is the equilibrium PMF and $\phi_{mp}(z)$ is the transmembrane potential profile along the z-axis, as presented in the two previous sections [see Fig. 12.3(c)]. Additionally, the diffusion constant profile $D(z)$ can be extracted from the velocity autocorrelation function of the ions.[66,67]

The stochastic Brownian motion of the multi-ion system is efficiently implemented as a continuous time Markov chain, in which discrete states correspond to the ion positions and the state-to-state random walk depends on exponentially distributed random survival times.[d] The forward and backward transition rates are given by (e.g. for ion 1):

$$k_{[(z_1,z_2,z_3)\rightarrow(z_1\pm\delta z,z_2,z_3)]} = \frac{D(z_1)+D(z_1\pm\delta z)}{2\delta z^2} \times e^{-[W_{\text{tot}}(z_1\pm\delta z,z_2,z_3)-W_{\text{tot}}(z_1,z_2,z_3)]/2k_\text{B}T} \quad (12.6)$$

where δz is the grid spacing. Similarly, an ion can attempt to enter a two-ion occupied channel at any time with a rate of (e.g. on the intracellular side):

$$k_{entry} = [C_{\text{int}}]S\delta z \frac{D(z_{\text{min}})+D_{\text{bulk}}}{2\delta z^2} \times e^{-[W_{\text{tot}}(z_1,z_2,z_{\text{min}})-W_{\text{tot}}(z_1,z_2)]/2k_\text{B}T} \quad (12.7)$$

where $[C_{int}]$ is the ion concentration on the intracellular side and S is the cross-sectional area of the entrance vestibule. The equilibrium PMF of the two-ion occupied channel, $W(z_1, z_2)$, corresponds to the three-ion PMF with the third ion as far as possible from the selectivity filter, i.e. $W(z_1, z_2) \approx W(z_{\text{max}}, z_1, z_2)$.

[d] Such a Markov random walk satisfies the condition of detailed balance under equilibrium conditions in the absence of net flux, and the stochastic evolution of the system obeys a multi-dimensional Smulochowsky (Nernst-Planck) diffusion equation as δz becomes increasingly small.[68]

Following this scheme, ions in the channel undergo a random walk on the total free energy surface with the space-dependent diffusion constant $D(z)$, hopping from state to state [see Fig. 12.3(e)]. In the case of KcsA, the calculated diffusion constant varies weakly throughout the entire permeation pathway, decreasing to roughly 70% of its bulk value in the selectivity filter region.[54] Further analysis shows that dynamical memory and inertial effects are negligible and that using a non-inertial Markovian dynamics approximation (i.e. BD) is physically justified. For these reasons, the average structural character of the random ion movements governing the conduction mechanism (i.e. single-file diffusion of the ions with water in between) is largely determined by the multi-ion free energy surface W_{tot} rather than the dissipative and frictional forces.

The PMF/BD framework presented here enables one to simulate ion fluxes for various conditions of ion concentration and transmembrane potential. The calculated current-voltage and conductance-concentration relations for the KcsA K^+ channel are shown in Fig. 12.3(f–g). The experimental I–V is well reproduced at small and moderate voltages, but non-linearity becomes too pronounced at large voltages because access resistance is not included in the model. The values of G_{max}, the maximum conductance of the channel at saturating concentration, estimated from Fig. 12.3(g) is on the order of 550 pS and 360 pS for outward and inward ions flux, respectively. These values are in remarkable agreement with the experimental measurements.[69]

Although no parameters were specifically adjusted to reproduce the value of the maximum conductance of KcsA, the results are in excellent agreement with available observations. In view of the extreme sensitivity of calculations based on atomic models (a small increase of approximately $k_{\text{B}}T$ in the central energy barrier in the ion-conduction mechanism is sufficient to decrease the ion flux by a factor of three), even a semi-quantitative agreement with experimental results is very satisfying. The present effort demonstrates that the calculation of the conductance characteristics of a selective ion channel from first principles is possible. A similar approach could be generally useful in studies of slow biomolecular processes whenever there is a need to extend the information extracted from all-atom MD trajectories to long time scales, e.g. in the case of gating events.

12.5 Ion Selectivity

Questions about ion selectivity have fascinated researchers for decades. Many investigators, with many different ideas, have contributed to frame the current view of ion selectivity.[2] The availability of computations based on atomic models[1,34,38,70–73] now offers a "virtual route" for testing various ideas about the molecular mechanism of ion selectivity.

12.5.1 *Selectivity Concepts*

In its simplest terms, selectivity reflects the fact that the "wrong" ion encounters more difficulty in permeation than the "correct" ion, i.e. it experiences microscopic forces that makes its progress through the channel more difficult (barriers along its permeation PMF are higher). In this sense, ion selectivity is first and foremost about energetics. But selectivity may manifest itself in different ways, depending on whether it is experimentally probed using equilibrium binding measurements or non-equilibrium flux and ionic current measurements.[2] Some types of measurements are more sensitive to the free energy at the bottom of a binding site, whereas other types of experiments are more sensitive to the height of free energy barriers. Electrophysiological measurements with blockade relief are most convenient to quantitatively characterize the selectivity of K^+ channels, because the experimental conditions approach those of thermodynamic equilibrium (which is intrinsically easier to interpret than kinetic measurements). Historically, Ba^{2+} blockade experiments were used by Miller and Neyton,[74] and later by Latorre and co-workers,[75] to detect and characterize the ion binding sites in the pore of the large conductance BK channels. These results indicated that there is a binding site between the extracellular solution and the position of the Ba^{2+} blocker, called the "external lock-in site," which is highly selective for K^+ over Na^+ by 4–6 kcal/mol. The X-ray structure of

KcsA in the presence of Ba^{2+} shows that it binds near the site S_4 [see Fig. 12.3(d)], at the intracellular end of the selectivity filter,[76] which suggests that the external lock-in site is either the site S_1 or S_2 (it is unlikely that sites S_4 and S_3 are simultaneously occupied under those conditions). As pointed out by Neyton and Miller,[74] these results present a purely thermodynamic view of selectivity (rather than kinetic), as it is governed by the relative free energy of the ions at the bottom of the binding site that plays the dominant role (rather than at the top of the energy barrier). Fundamentally, selectivity for K^+ over Na^+ implies that the relative free energy $\Delta\Delta G$ of K^+ and Na^+ in the pore and in the bulk solution,

$$\begin{aligned}\Delta\Delta G(\mathrm{K}^+ \rightarrow \mathrm{Na}^+) = &\left[G_{\mathrm{pore}}(\mathrm{Na}^+) - G_{\mathrm{bulk}}(\mathrm{Na}^+)\right] \\ &- \left[G_{\mathrm{pore}}(\mathrm{K}^+) - G_{\mathrm{bulk}}(\mathrm{K}^+)\right]\end{aligned} \tag{12.8}$$

is larger than zero. According to electrophysiological measurements, $\Delta\Delta G$ is on the order of 4–6 kcal/mol for the highly selective BK channels.[74,75] The key question about the selectivity of K^+ channels is thus to identify the physical origin of the unfavorable free energy $\Delta\Delta G$. Because of its smaller radius, the hydration free energy of Na^+ is ~18 kcal/mol more negative than that of K^+, i.e. $G_{\mathrm{bulk}}(Na^+) \approx G_{\mathrm{bulk}}(K^+)$ –18 kcal/mol. This larger difference in the hydration free energies of Na^+ and K^+ ions, corresponding to ΔG_{bulk}, sets the fundamental "baseline" for the Na/K selectivity according to which all biological ion channels are carrying its function (whether it is specific for Na^+ or for K^+). One may also note that $G_{\mathrm{pore}}(Na^+) \approx G_{\mathrm{pore}}(K^+)$ – 12 kcal/mol, which implies that — in absolute terms — Na^+ interacts more strongly with the pore than K^+ (it is the difference in hydration free energy that shifts the balance).

The most intuitively appealing explanation of K^+ selectivity is the concept of the "snug-fit" proposed in the early 1970s.[77] The snug-fit mechanism postulates that the binding site is, for structural reasons, rigidly constrained in an optimal geometry so that a dehydrated K^+ fits with proper coordination, but that Na^+ is too small and is thus poorly coordinated by the host. Selectivity is then due to the difference in the interaction of the ions with the coordinating ligands (i.e. carbonyl oxygen atoms lining the pore) compared to the hydration free energy. Structurally, the snug-fit mechanism implies a significant structural inability to deform and adapt: the energetic cost upon collapsing to cradle a Na^+ (a structural distortion of about 0.38 Å) must give rise to a significant energy penalty (much larger than $k_\text{B}T$). This is obviously an idealization. In reality, molecules are flexible and may be able to structurally deform and adapt (to some extent) to a bound ion. To go beyond "verbal arguments" about selectivity, it is necessary to use computational approaches based on atomic models.

12.5.2 *Selectivity Calculations by Free Energy Perturbation*

Free energy perturbation (FEP) based on all-atom MD simulations[78,79] represents the most fundamental approach to elucidate the microscopic origin of "hidden" thermodynamic factors governing the function of biological systems. By carrying FEP simulations, it is possible to incorporate the effect of thermal fluctuations and the contributions from all the atomic coordinates into a computed free energy difference of interest. The difference in solvation free energy between K^+ and Na^+ can be expressed as:[80]

$$e^{-\left[G(\text{Na}^+)-G(\text{K}^+)\right]/k_\text{B}T} = \left\langle e^{-\left[E(\text{Na}^+)-E(\text{K}^+)\right]/k_\text{B}T} \right\rangle_{(\text{K}^+)} \tag{12.9}$$

where $E(\text{Na}^+)$ and $E(\text{K}^+)$ are, respectively, the potential energy with a Na^+ or a K^+ ion in the dynamical system (keeping all atomic coordinates unchanged). In the FEP expression, the bracket formally represents an average over configurations generated with a K^+ ion in the system. Using the FEP method, the free energy difference between Na^+ and K^+ in the bulk solution[48,49] as well as inside the channel[38,70,72] can be calculated from all-atom MD simulations.

Such FEP/MD simulations were performed for each of the five cation binding sites in the selectivity filter of the KcsA channel[38,70,72] and for the binding sites of the non-selective (but structurally similar) NaK channel.[81] The calculations indicate the most selective site in KcsA is located in the middle of the pore (site S_2). In contrast, none of the binding sites in the NaK channel is selective. Analysis showed that the variations in the free energy of selectivity in the various sites were associated with the differences in hydration of the cation.[81] A cation in the site S_2 of KcsA is almost dehydrated and coordinated by eight backbone carbonyl oxygens, while a cation in the corresponding site of NaK is well hydrated.

The calculations show that selectivity of the binding sites of KcsA and NaK is largely controlled by the dynamical interplay of local ion-ligand and ligand-ligand interactions. Ion-ligand interactions are obviously attractive, while ligand-ligand repulsion acts as a hidden "through-space" electrostatic strain energy. Strain energy (here extended to ligand repulsion) is a classic host-guest chemistry concept to describe processes involving an induced-fit of the receptor upon the binding of a subtrate. The key variables are the number of coordinating ligands, as well as their particular properties. The resulting dynamical interplay of ion-ligand and ligand-ligand interactions in a binding site is complex. For example, for an ion coordinated by N ligands, the magnitude of ion-ligand interactions and of the ligand-ligand strain grow like $\sim N$ and $\sim N^2$, respectively. Most importantly, coordinating oxygen atoms donated by a carbonyl or a water molecule are not equivalent. For this reason, the coordination number alone cannot predetermine ion selectivity because both ion-ligand and ligand-ligand interactions depend also on the electrostatic nature of the ligands. Different combinations of water and carbonyls can give rise to different K^+/Na^+ selectivity. A dynamical site with eight carbonyl groups is robustly selective for K^+ over Na^+, but the selectivity is lost as the carbonyl groups progressively are replaced by water molecules, or as the coordination number is decreased. A tightly controlled dehydration of permeating cations, as enforced by the long and narrow KcsA pore, is an essential aspect that enables the robust selectivity for K^+ over Na^+. Because of the widening at the level corresponding to the S_2 site, the NaK pore allows a minor increase in ion hydration and is permissive to Na^+.

12.6 Gating

Little is known about the structural features of ion channel gating. Even in the rare occasions for which both conductive and non-conductive structures are available, many uncertainties remain as to how they might relate to functional data. Elucidating the microscopic mechanisms underlying gating events remains a challenge that will require the combination of many different experimental and theoretical approaches. Molecular mechanics simulations can be useful at different stages of that process: they can potentially reveal small conformation changes that might be associated with gating, or serve as a theoretical framework to model larger conformational changes.

Any conformational change in the vicinity of the pore can potentially affect the conductance of a channel. The time-scale of the transition and the extent at which the current is hampered define whether the conformational change should be considered as part of a gating mechanism or not. A really fast transition on the nano- to micro-second time-scale would most probably affect the conductance of the channel and would not be resolved by usual electrophysiological measurements. A transition on the micro- to milli-second time-scale could be associated with flickering in single-channel recording and other physiologically fast gating events. Macroscopic gating events, as usually conceived, would obviously require slower transitions. Information on the time-scale of the conformational change can be evaluated by calculating the potential of mean force governing the transition from the open to the closed state. In the case of gating mechanisms involving only one or two residues, the reaction-coordinates of the PMF can usually be defined in terms of the internal coordinates of the backbone and side-chains of the residues. For more complex conformational changes involving less than 10 residues, the root mean squared deviation (RMSD) between the initial and final conformations represents an effective and useful order parameter to monitor the progress of the transformation.[82] The PMF provides information on the relative stability of the end states, as well as on the free energy barrier separating them. The PMF can be further interpreted in terms of transition rates by either using rate constant theory, or the BD/PMF framework described in the previous section.

To insure that a given conformational change corresponds to a gating process, one should also evaluate the conductance of the channel in the conducting and putative non-conducting states. Depending on the width of the pore at the level of the gate, one could either use the BD/PMF framework as described above, or alternative techniques based on continuum mean-field theories (see Section 12.7).

Using this approach, it was demonstrated that the simple reorientation of the amide plane of two residues in the selectivity filter of the KcsA K^+ channel could potentially act as a gate by reducing the ionic current by at least 100 fold.[83] By calculating the PMF governing the gating transition for different ion occupancy states, it was shown that this gate is most probably related to a physiologically important gating mechanism known as slow or C-type inactivation.[84]

12.7 Overview of Alternative Approaches

Approaches that are simpler and computationally less expensive than all-atom MD are very important tools in studies of ion channels. In particular, macroscopic continuum electrostatic calculations, in which the polar solvent is represented as a structureless dielectric medium can help reveal the dominant energetic factors related to ion permeation, and thus, serve to illustrate fundamental principles in a particularly clear fashion.[85] In this approach, the protein is typically kept in a fixed conformation. This implies that factors concerning the structural flexibility and thermal fluctuations of the protein are not taken directly into consideration. The calculation of the transmembrane potential profile through the pore of the KcsA channel as described in Section 12.4.1 is such an example. Continuum electrostatic calculations, either based on the Poisson-Boltzmann equation or the Generalized Born approximation, can be combined with molecular dynamics or stochastic algorithms to study the evolution of a system.[86,87]

12.7.1 *Grand Canonical Brownian Dynamics*

Brownian dynamics (BD), which consists of integrating stochastic equation of motions describing the displacement of the ions with some effective potential function, is an attractive computational approach for simulating the permeation process over long time-scales. As presented in Section 12.4, the effective potential function can be rigorously calculated from all-atom potential-of-mean-force simulations. Alternatively, a less computationally demanding approach consists of calculating the ion-protein and ion-ion interactions at each BD time-step on the basis of continuum electrostatic approximation, i.e. without treating all the solvent molecules explicitly.[88,89] The approach is particularly well-suited for the study of wide aqueous pores.[90,91]

To simulate ion fluxes on a long time-scale, the total number of ions in the system must be allowed to fluctuate under the influence of specific non-equilibrium boundary conditions. This can be accomplished by combining the BD stochastic dynamics with the Grand Canonical Monte Carlo (GCMC) algorithm.[92] The GCMC procedure has the effect of enforcing boundary conditions corresponding to constant electro-chemical potential in two buffer regions representing the bulk solution on either side of a membrane. Since the buffer regions cannot run out of particles or be filled by particles, they essentially act as infinite thermodynamic reservoirs and sinks for the particles with respect to the central inner region. The procedure can be used to simulate equilibrium as well as non-equilibrium conditions of ion diffusion and permeation.[92] One cycle of GCMC/BD corresponds to one step of BD followed by a few steps of GCMC (typically one to 10) to maintain the buffer regions in equilibrium. No ion creation or destruction is taking place in the inner region, and the time-course of the ions in the inner system evolves dynamically according to BD. When the system is at the equilibrium, the electro-chemical potential of any ion is the same in all the regions of the system and there is no net flow. However, when non-equilibrium conditions are imposed at the boundaries, a stationary state is simulated as particles flow from the regions with a high value of electro-chemical potential to the regions with lower values.

12.7.2 *Poisson-Nernst-Planck*

It is also possible to simplify the details of a system even further by treating the average ion fluxes in terms of concentration gradient and average electric field. This is the goal of the Poisson-Nernst-Planck (PNP) continuum electrodiffusion theory.[93,94] Rather than an actual simulation of atomic movements, PNP requires a numerical solution to a set of differential equations. PNP is often described as a "mean-field theory" because the average electrostatic potential and average concentration gradients determine the average fluxes. In the absence of any net flux PNP becomes equivalent to the equilibrium non-linear Poisson-Boltzmann equation. As highlighted by several authors,[95,96] the underlying approximations can lead to serious problems and the theory must be used with caution.

12.8 Future Outlook

The recent progress and achievement are encouraging and illustrate that, although computer simulations can be improved, they are able to provide results of semi-quantitative accuracy. If one is allowed to dream a little, it shall one day be possible to use sophisticated computer algorithms, exploiting all the available information in structural and genomic databases, to construct a reliable atomic model of any channel, and then characterize fully and accurately its functional phenotype (conductance, selectivity, gating, inactivation, etc …) *in silico* with computations. Such a virtual model could also serve to help rationally design novel drugs and molecules specifically to alter the function of the channel in a desired way for the ultimate purpose of curing a given pathological physiological condition. This may sound almost like science fiction, and to a certain extent, one must admit that we are very far from having such incredible abilities at the present time. Nonetheless, as a long-term goal this makes perfect sense. Developing the required skills and techniques to progress toward such a goal, of course, will require tremendous progress on several fronts. With this long-term perspective in mind, it is useful to elaborate more specifically on the directions that are likely to be very active in the near future.

12.8.1 *Force Field Development*

Most simulations of ion channels to date have been based on additive force fields that treat the influence of induced polarization in an effective average way. For this reason, the result of most simulations is only of semi-quantitative accuracy. This is, of course, a severe limitation, and at the present time, computational chemists and theoreticians are actively pursuing the development of a new generation of force fields that will include induced polarization for computational studies of biological systems.[97,98] However, more work is needed before such potential functions are ready to be used reliably in simulations of biological ion channels. Meanwhile, there are reasons to believe that MD studies of ion channels can still yield meaningful results, as long as they are based on effective potential functions that have been calibrated to correctly reproduce solvation free energies. A recent simulation study of K^+ permeation through the gA channel illustrates this point clearly.[51]

12.8.2 *Studying Macroscopic Conformational Changes Involved in Transduction Events*

Traditionally, it has been possible to monitor important microscopic processes by mapping the free energy landscape along some prechosen reaction coordinates. As described above, ion permeation through the KcsA channel can be monitored by computing the free energy profile as a function of the position of the ions along the pore. However, the large conformational changes that underlie the fundamental transduction events in membrane proteins (voltage or ligand gating, activity of transporters or pumps …) are expected to occur via concerted motions involving a large number of atoms. Attempts at describing those complex transitions using a simple reaction coordinate are clearly going to fail. It is therefore necessary to develop computational methodologies able to determine a plausible reaction coordinate for conformational changes of macromolecules. To make progress on this issue, it is useful to imagine that the macromolecule is undergoing random Brownian diffusive motions in the multidimensional space of all its degrees of freedom during a large conformational change. The issue is thus to refine a "path," defined as a sequence of states, linking the end-points of a conformational change. Several methods have been developed to seek a solution to this problem, see, for example, Elber *et al.*[99] Interestingly, the transition path that has the highest likelihood, among all the possible paths, corresponds to systems evolving along the so-called minimum free energy path (MFEP). Determining the MFEP is the goal of the "string method" of Maragliano *et al.*[100] The knowledge of the MFEP reveals explicitly the mechanism of the conformational change, and enables one to compute the transition rates between the two conformations at the end-points. This area of research is expected to become extremely active in the next years, with the rising need to describe increasingly complex conformational changes.

12.8.3 *Bridging the Gap between Atomic Simulations and Physiology*

This chapter presented an overview of the computational methods used for modeling ion channels at the atomic level. While there is no doubt that the behavior of macromolecules such as ion channels can be understood from the fundamental laws of physics, there is obviously a long, long way from the atoms to the physiology of entire organisms. Given the level of complexity of these systems, we do not think that a straight brute-force "bottom-up" approach is viable, nor desirable. A more realistic strategy is to adopt a progressive multilevel representation of these complex systems, whereby the details of a finer level are absorbed into the effective parameters of the next, coarser level, and so on and so forth. Although this has not been achieved yet, some elements are already in place. For example, the work of Voth and collaborators provides good clues as to how a detailed atomic representation can be optimally reduced to a smaller number of degrees of freedom.[101] Such reduced coarse-grained models can then be utilized to explore the mechanistic and functional consequences, as done beautifully by Oster and collaborators in the case of F1-ATPase, for example.[102] The statistical behavior of large population of different channels, can then be understood with approaches similar to those developed by Rudy to model the cardiac tissue.[103] What is important is to have rigorous computational algorithms enabling us to trace a phenomenon, all the way from the molecular level to the physiological level. Bridging this gap rigorously will allow the possibility to understand pathological mutations and the action of specific drugs.

References

1. Roux B, Allen T, Bernèche S, Im W. (2004) Theoretical and computational models of biological ion channels. *Q Rev Biophys* **37**: 15–103.
2. Hille B. (2001) *Ion Channels of Excitable Membranes,* 3rd ed., Sinauer, Sunderland, Massachusettes.
3. Doyle DA, Morais Cabral JH, Pfuetzner RA, *et al.* (1998) The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* **280**: 69–77.
4. Roux B, Karplus M. (1991) Ion transport in a gramicidin-like channel: structure and thermodynamics. *Biophys J* **59**: 961–981.
5. Tieleman DP, Berendsen HJC, Sansom MSP. (1999) An alamethicin channel in a lipid bilayer: molecular dynamics simulations. *Biophys J* **76**: 1757–1769.
6. Zhou Y, Morais Cabral JH, Kaufman A, MacKinnon R. (2001) Chemistry of ion coordination and hydration revealed by a K^+ channel-fab complex at 2.0 a resolution. *Nature* **414**: 43–48.
7. Shi N, Ye S, Alam A, Chen L, Jiang Y. (2006) Atomic structure of a Na^+- and K^+-conducting channel. *Nature* **440**: 570–574.
8. Jiang Y, Lee A, Chen J, *et al.* (2002) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**: 515-522.
9. Jiang Y, Lee A, Chen J, *et al.* (2003) X-ray structure of a voltage-dependent K^+ channel. *Nature* **423**: 33–41.
10. Kuo A, Gulbis JM, Antcliff JF, *et al.* (2003) Crystal structure of the potassium channel kirbac1.1 in the closed state. *Science* **300**: 1922–1926.
11. Nishida M, Cadene M, Chait BT, Mackinnon R. (2007) Crystal structure of a kir3.1-prokaryotic kir channel chimera. *EMBO J*, **in press**.
12. Long SB, Campbell EB, Mackinnon R. (2005) Crystal structure of a mammalian voltage-dependent shaker family K^+ channel. *Science* **309**: 897–903.

12a. Long SB, Tao X, Campbell EB, MacKinnon R. (2007) Atomic structure of a voltage-dependent K$^+$ channel in a lipid membrane-like environment. *Nature* **450**: 376–382.

12b. Jasti J, Furukawa H, Gonzales EB, Gouaux E. (2007) Structure of acid-sensing ion channel 1 at 1.9 A resolution and low pH. *Nature* **449**: 316–323.

13. Miyazawa A, Fujiyoshi Y, Unwin N. (2003) Structure and gating mechanism of the acetylcholine receptor pore. *Nature* **423**: 949–955.

14. Unwin N. (2005) Refined structure of the nicotinic acetylcholine receptor at 4a resolution. *J Mol Biol* **346**: 967–989.

15. Chang G, Spencer RH, Lee AT, Barclay MT, Rees DC. (1998) Structure of the mscl homolog from mycobacterium tuberculosis: a gated mechanosensitive ion channel. *Science* **282**: 2220–2226.

16. Bass RB, Strop P, Barclay M, Rees DC. (2002) Crystal structure of *Escherichia coli* mscs, a voltage-modulated and mechanosensitive channel. *Science* **298**: 1582–1587.

16a. Lunin VV, Dobrovetsky E, Khutoreskaya G, Zhang R, Joachimiak A, Doyle DA, Bochkarev A, Maguire ME, Edwards AM, Koth CM. (2006) Crystal structure of the CorA Mg^{2+} transporter. *Nature* **440**: 833–837.

16b. Eshaghi S, Neigowski D, Kohl A, Martinez Molina D, Lesley SA, Nordlund P. (2006) Crystal structure of a divalent metal ion transporter CorA at 2.9 Å resolution. *Science* **313**: 354–357 Erratum: *Science* **313**: 1389.

16c. Hattori M, Tanaka Y, Fukai S, Ishitani R, Nureki O. (2007) Crystal structure of the MgtE Mg^{2+} transporter. *Nature* **448**: 1072–1075.

17. Dutzler R, Campbell EB, Cadene M, Chait BT, MacKinnon R. (2002) X-ray structure of a CLC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. *Nature* **415**: 287–294.

18. Dutzler R, Campbell EB, MacKinnon R. (2003) Gating the selectivity filter in CLC chloride channels. *Science* **300**: 108–112.

19. Accardi A, Walden M, Nguitragool W, *et al.* (2005) Separate ion pathways in a Cl–/H+ exchanger. *J Gen Physiol* **126**: 563–570.

20. Hunte C, Screpanti E, Venturi M, *et al.* (2005) Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH. *Nature* **435**: 1197–1202.

20a. Toyoshima C, Nakasako M, Nomura H, Ogawa H. (2000) Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **405**: 647–655.

20b. Morth JP, Pedersen BP, Toustrup-Jensen MS, Sørensen TL, Petersen J, Andersen JP, Vilsen B, Nissen P. (2007) Crystal structure of the sodium-potassium pump. *Nature* **450**: 1043–1049.

20c. Pedersen BP, Buch-Pedersen MJ, Morth JP, Palmgren MG, Nissen P. (2007) Crystal structure of the plasma membrane proton pump. *Nature* **450**: 1111–1114.

21. Baker D, Sali A. (2001) Protein structure prediction and structural genomics. *Science* **294**: 93–96.
22. White SH, Wimley WC. (1999) Membrane protein folding and stability: physical principles. *Ann Rev Biophys Biomol Struct* **28**: 319–365.
23. Laine M, Lin MCA, Bannister JPA, *et al.* (2003) Atomic proximity between S4 segment and pore domain in shaker potassium channels. *Neuron* **39**: 467–481.
24. Silverman WR, Roux B, Papazian DM. (2003) Structural basis of two-stage voltage-dependent activation in K^+ channels. *Proc Natl Acad Sci USA* **100**: 2935–2240.
25. Hong KH, Miller C. (2000) The lipid-protein interface of a shaker K(+) channel. *J Gen Physiol* **115**: 51–58.
26. Li-Smerin Y, Swartz KJ. (2001) Helical structure of the COOH terminus of S3 and its contribution to the gating modifier toxin receptor in voltage-gated ion channels. *J Gen Physiol* **117**: 205–218.
27. Chanda B, Asamoah OK, Blunck R, Roux B, Bezanilla F. (2005) Gating charge displacement in voltage-gated ion channels involves limited transmembrane movement. *Nature* **436**: 852–856.
28. Long SB, Campbell EB, Mackinnon R. (2005) Crystal structure of a mammalian voltage-dependent shaker family K^+ channel. *Science* **309**: 897–903.
29. Kandt C, Ash WL, Tieleman DP. (2007) Setting up and running molecular dynamics simulations of membrane proteins. *Methods* **41**: 475–488.
30. Woolf TB, Roux B. (1994) Molecular dynamics simulation of the gramicidin channel in a phospholipid bilayer. *Proc Natl Acad Sci USA* **91**: 11631–11635.
31. Nagle JF, Tristram-Nagle S. (2000) Structure of lipid bilayers. *Biochem Biophys Acta* **1469**: 159–195.
32. Bernèche S, Roux B. (2000) Molecular dynamics of the KcsA K(+) channel in a bilayer membrane. *Biophys J* **78**: 2900–2917.
33. Sachs JN, Crozier PS, Woolf TB. (2004) Atomistic simulations of biologically realistic transmembrane potential gradients. *J Chem Phys* **121**: 10847–10851.
34. Guidoni L, Torre V, Carloni P. (1999) Potassium and sodium binding to the outer mouth of the K^+ channel. *Biochemistry* **38**: 8599–8604.
35. Bond PJ, Sansom MS. (2003) Membrane protein dynamics versus environment: simulations of OmpA in a micelle and in a bilayer. *J Mol Biol* **329**: 1035–1053.
36. MacKerell AD, Brooks B, III CLB, *et al.* (1998) CHARMM: the energy function and the program. In *The Encyclopedia of Computational Chemistry.* Chichester, John Wiley & Sons.
37. Darden T, York D, Pedersen L. (1993) Particle mesh ewald — an n.Log(n) method for ewald sums in large systems. *J Chem Phys* **98**: 10089–10092.
38. Bernèche S, Roux B. (2001) Energetics of ion conduction through the K^+ channel. *Nature* **414**: 73–77.

39. Anezo C, de Vries AH, Holtje HD, Tieleman DP, Marrink SJ. (2003) Methodological issues in lipid bilayer simulations. *J Phys Chem B* **107**: 9424–9433.
40. Feller SE, Pastor RW. (1996) On simulating lipid bilayers with an applied surface tension: periodic boundary conditions and undulations. *Biophys J* **71**: 1350–1355.
41. Feller SE, MacKerell AD. (2000) An improved empirical potential energy function for molecular simulations of phospholipids. *J Phys Chem B* **104**: 7510–7515.
42. Benz RW, Castro-Roman F, Tobias DJ, White SH. (2005) Experimental validation of molecular dynamics simulations of lipid bilayers: a new approach. *Biophys J* **88**: 805–817.
43. de Vries AH, Mark AE, Marrink SJ. (2004) The binary mixing behavior of phospholipids in a bilayer: a molecular dynamics study. *J Phys Chem B* **108**: 2454–2463.
44. Murzyn K, Rog T, Pasenkiewicz-Gierula M. (2005) Phosphatidylethanolamine-phosphatidylglycerol bilayer as a model of the inner bacterial membrane. *Biophys J* **88**: 1091–1103.
45. Leekumjorn S, Sum AK. (2006) Molecular simulation study of structural and dynamic properties of mixed DPPC/DPPE bilayers. *Biophys J* **90**: 3951–3965.
46. Klassen J, Anderson S, Blades A, Kebarle P. (1996) Reaction enthalpies for m(+)l = m(+)+l, where m(+) = Na^+ and K^+ and l equals acetamide, n-methylacetamide, n, n-dimethylacetamide, glycine, and glycylglycine, from determinations of the collision-induced dissociation thresholds. *J Phys Chem* **100**: 14218–14227.
47. Roux B, Karplus M. (1995) Potential energy function for cations-peptides interactions: an *ab initio* study. *J Comput Chem* **16**: 690–704.
48. Åqvist J. (1990) Ion water interaction potential derived from free energy perturbation simulations. *J Phys Chem* **94**: 8021–8024.
49. Beglov D, Roux B. (1994) Finite representation of an infinite bulk system: solvent boundary potential for computer simulations. *J Chem Phys* **100**: 9050–9063.
50. Roux B, Bernèche S. (2002) On the potential functions used in molecular dynamics simulations of ion channels. *Biophys J* **82**: 1681–1684.
51. Allen TW, Andersen OS, Roux B. (2006) Molecular dynamics — potential of mean force calculations as a tool for understanding ion permeation and selectivity in narrow channels. *Biophys Chem* **124**: 251–267.
52. Cox BG, Hedwig GR, Parker AJ, Watts DW. (1974) Solvation of ions xix. Thermodynamic properties for transfer of single ions between protic and dipolar aprotic solvents. *Aust J Chem* **27**: 477–501.
53. Khalili-Araghi F, Tajkhorshid E, Schulten K. (2006) Dynamics of K^+ ion conduction through kv1.2. *Biophys J* **91**: L72–L74.

54. Bernèche S, Roux B. (2003) A microscopic view of ion conduction through the K+ channel. *Proc Natl Acad Sci USA* **100**: 8644–8648.
55. Tieleman DP, Leontiadou H, Mark AE, Marrink SJ. (2003) Simulation of pore formation in lipid bilayers by mechanical stress and electric fields. *J Am Chem Soc* **125**: 6382–6383.
56. Roux B. (1999) Statistical mechanical equilibrium theory of selective ion channels. *Biophys J* 77: 139–153.
57. Im W, Beglov D, Roux B. (1998) Continuum solvation model: electrostatic forces from numerical solutions to the Poisson-Bolztmann equation. *Comput Phys Commun* **111**: 59–75.
58. Allen TW, Andersen OS, Roux B. (2004) Energetics of ion conduction through the gramicidin channel. *Proc Natl Acad Sci USA* **101**: 117–122.
59. Darve E, Pohorille A. (2001) Calculating free energies using average force. *J Chem Phys* **115**: 9169–9183.
60. Hénin J, Chipot C. (2004) Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J Chem Phys* **121**: 2904–2914.
61. Kale L, Skeel R, Bhandarkar M, *et al.* (1999) Namd2: greater scalability for parallel molecular dynamics. *J Comput Phys* **151**: 283–312.
62. Roux B. (1995) The calculation of the potential of mean force using computer-simulations. *Comput Phys Commun* **91**: 275–282.
63. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM. (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. *J Comput Chem* **13**: 1011–1021.
64. Ermak DL, McCammon JA. (1978) Brownian dynamics with hydrodynamic interactions. *J Chem Phys* **69**: 1352–1360.
65. Zwanzig R. (2001) *Nonequilibrium Statistical Mechanics.* Oxford University Press, New York.
66. Berne BJ, Borkovec M, Straub JE. (1988) Classical and modern methods in reaction-rate theory. *J Phys Chem* **92**: 3711–3725.
67. Crouzy S, Woolf TB, Roux B. (1994) A molecular dynamics study of gating in dioxolane-linked gramicidin a channels. *Biophys J* **67**: 1370–1386.
68. Schumaker MF, Pomes R, Roux B. (2001) Framework model for single proton conduction through gramicidin. *Biophys J* **80**: 12–30.
69. LeMasurier M, Heginbotham L, Miller C. (2001) Kcsa: it's a potassium channel. *J Gen Physiol* **118**: 303–314.
70. Luzhkov VB, Åqvist, J. (2001) K+/Na+ selectivity of the KcsA potassium channel from microscopic free energy perturbation calculations. *Biochim Biophys Acta* **1548**: 194–202.
71. Shrivastava IH, Tieleman DP, Biggin PC, Sansom MSP. (2002) K+ versus Na+ ions in a K channel selectivity filter: a simulation study. *Biophys J* **83**: 633–645.

72. Noskov SY, Bernèche S, Roux B. (2004) Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands. *Nature* **431**: 830–834.
73. Chung SH, Corry, B. (2005) Three computational methods for studying permeation, selectivity, and dynamics in biological ion channels. *Soft Matter* **1**: 417–427.
74. Neyton J, Miller, C. (1988) Discrete Ba^{2+} block as a probe of ion occupancy and pore structure in the high-conductance Ca^{2+} — activated K^+ channel. *J Gen Physiol* **1988**: 569–596.
75. Vergara C, Alvarez O, Latorre R. (1999) Localization of the K^+ lock-in and the Ba^{2+} binding sites in a voltage-gated calcium-modulated channel. Implications for survival of K^+ permeability. *J Gen Physiol* **114**: 365–376.
76. Jiang Y, MacKinnon R. (2000) The barium site in a potassium channel by X-ray crystallography. *J Gen Physiol* **115**: 269–272.
77. Bezanilla F, Armstrong CM. (1972) Negative conductance caused by entry of sodium and cesium ions into the potassium channels of squid axons. *J Gen Physiol* **60**: 588–608.
78. McCammon J, Straatsma T. (1992) Alchemical free energy simulation. *Ann Rev Phys Chem* **43**: 407.
79. Kollman PA. (1993) Free energy calculations: applications to chemical and biochemical phenomena. *Chem Rev* **93**: 2395–2417.
80. Zwanzig RW. (1954) High temperature equation of state by a perturbation method. *J Chem Phys* **22**: 1420–1426.
81. Noskov SY, Roux B. (2007) Importance of hydration and dynamics on the selectivity of the KcsA and NaK channels. *J Gen Physiol* **129**: 135–143.
82. Banavali NK, Roux B. (2005) Free energy landscape of α-DNA to β-DNA conversion in aqueous solution. *J Am Chem Soc* **127**: 6866–6876.
83. Bernèche S, Roux B. (2005) A gate in the selectivity filter of potassium channels. *Structure (Camb)* **13**: 591–600.
84. Yellen G. (1998) The moving parts of voltage-gated ion channels. *Q Rev Biophys* **31**: 239–295.
85. Roux B, Bernèche S, Im W. (2000) Ion channels, permeation and electrostatics: insight into the function of KcsA. *Biochemistry* **39**: 13295–13306.
86. Im W, Roux B. (2001) Brownian dynamics simulations of ions channels: a general treatment of electrostatic reaction fields for molecular pores of arbitrary geometry. *J Chem Phys* **115**: 4850–4861.
87. Im W, Brooks CL. (2005) Interfacial folding and membrane insertion of designed peptides studied by molecular dynamics simulations. *Proc Natl Acad Sci USA* **102**: 6771–6776.
88. Cooper KE, Jakobsson E, Wolynes PG. (1985) The theory of ion transport through membrane channels. *Prog Biophys Mol Biol* **46**: 51–96.
89. Chung S, Hoyles M, Allen T, Kuyucak S. (1998) Study of ionic currents across a model membrane channel using Brownian dynamics. *Biophys J* **75**: 793–809.

90. Allen TW, Chung SH. (2001) Brownian dynamics of an open-state KcsA potassium channel. *Biophys Biochim Acta* **1515**: 83–91.
91. Burykin A, Schutz C, Villa J, Warshel A. (2002) Simulations of ion current in realistic models of ion channels: the KcsA potassium channel. *Protein Struct Funct Gen* **47**: 265–280.
92. Im W, Seefeld S, Roux B. (2000) A grand canonical Monte Carlo-Brownian dynamics algorithm for simulating ion channels. *Biophys J* **79**: 788–801.
93. Onsager L. (1926) Zur theorie der elektrolyte (1). *Phys Z* **27**: 388–392.
94. Schuss Z, Nadler B, Eisenberg R. (2001) Derivation of Poisson and Nernst-Planck equations in a bath and channel from a molecular model. *Phys Rev E* **64**: 036116.
95. Moy G, Corry B, Kuyucak S, Chung S. (2000) Tests of continuum theories as models of ion channels. I. Poisson-Boltzmann theory versus Brownian dynamics. *Biophys J* **78**: 2349–2363.
96. Corry B, Kuyucak S, Chung S. (2000) Tests of continuum theories as models of ion channels. II. Poisson-Nernst-Planck theory versus Brownian dynamics. *Biophys J* **78**: 2364–2381.
97. Halgren TA, Damm W. (2001) Polarizable force fields. *Curr Opin Struct Biol* **11**: 236–242.
98. Ponder JW, Case DA. (2003) Force fields for protein simulations. *Adv Protein Chem* **66**: 27.
99. Elber R, Ghosh A, Cardenas A. (2002) Long time dynamics of complex systems. *Acc Chem Res* **35**: 396–403.
100. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G. (2006) String method in collective variables: minimum free energy paths and isocommittor surfaces. *J Chem Phys* **125**: 024105.
101. Ayton GS, Noid WG, Voth GA. (2007) Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr Opin Struct Biol* **17**: 192–198.
102. Oster G, Wang HY. (2003) Rotary protein motors. *Trends Cell Biol* **13**: 114–121.
103. Faber GM, Rudy Y. (2007) Calsequestrin mutation and catecholaminergic polymorphic ventricular tachycardia: a simulation study of cellular mechanism. *Cardiovasc Res* **75**: 79–88.

This page intentionally left blank

Chapter 13

Milestones in Molecular Dynamics Simulations of RNA Systems

Y. Hashem[†], E. Westhof[†], and P. Auffinger[*,†]

13.1 Introduction

The first MD simulation of a complete protein (Bovine Pancreatic Trypsin Inhibitor or BPTI; 58 residues; 9.2 ps of simulation time) was published in 1976[1–3] and provided significant insight on the shortest biomolecular motions.[4] Hence, in 2007, the history of molecular dynamics (MD) simulations of biomolecular systems covers a timespan of 31 years. Nowadays, MD simulations of aqueous protein systems have become quite popular, and numerous methods have been developed to address a large variety of issues of interest to structural biochemists[5] on time scales reaching the microsecond or 10^6 ps.[6]

The first MD simulations of DNA duplexes were published seven years later, in 1983.[4,7,8] At the same time, the first MD simulations of an RNA system were reported.[4,9] The tRNA$^{\mathrm{Phe}}$ molecule that was chosen for these initial theoretical investigations comprised 76 nucleotides, and was, consequently, much larger than most biomolecular systems studied by then. Given the very limited

[*] Corresponding author: Email: p.auffinger@ibmc.u-strasbg.fr.

[†] Architecture et Réactivité de l'ARN, Université Louis Pasteur de Strasbourg, CNRS, IBMC, 15 rue René Descartes, 67084 Strasbourg Cedex, France.

computational means available during these pioneering days, drastic approximations had to be made in order to achieve 12 ps of simulated time. For instance, no solvent particles could be taken into consideration. Rather unfortunately, in such *in vacuo* conditions, nucleic acid systems revealed a strong propensity towards structural degradation. In order to improve structural stability, several strategies were developed, such as: setting all electrostatic charges to zero; scaling the electrostatic charges and/or the dielectric constant in order to mimic the effects of the solvent and of counterion condensation; including explicit hydrogen bonds; and even using periodic longitudinal boundary conditions that make a DNA oligomer effectively a segment of an infinite double helix (see Ref. 4). Evidently, these methods were of a transient kind, and it became soon obvious that real stability improvements were only possible through the inclusion of explicit solvent particles, especially if one aims to investigate molecular motions on longer time scales. Indeed, explicit solvent techniques were applied with some success to protein and DNA systems.[10,11] Unfortunately, for RNA molecules, no real improvement could be achieved since this class of nucleic acids was shown to be much more sensitive than DNA systems to the accuracy with which their environment was modeled. This observation resulted in an almost complete absence of RNA simulations between 1983 and 1995. It was only in 1995, with the development of cost-efficient methods for the treatment of long-range electrostatic interactions based on the Ewald summation techniques, that the field came "back to life."

In the following, we will evoke the pre- and contemporary Ewald "era" by describing some salient results gathered during this 24-year-long journey of MD simulations of RNA systems. This report includes also a table that regroups all MD simulations of RNA systems reported so far that use state-of-the-art Ewald summation techniques and explicit solvent models (Table 13.1).

We will not address issues related to the use of implicit solvation methods that are described in other publications.[5,12–14] The interested reader will find in the following reviews more information on general[5,15,16] and RNA[3,4,17–21] MD simulations.

Table 13.1 List of MD Simulations of RNA Systems (up to September 2007) Using an Explicit Representation of the Solvent and Ewald Summation Methods for the Treatment of the Long-Range Electrostatic Interactions (Simulations Using Truncation[37–42] or Density Functional Methods[86] are not Listed in the Table). This Table is an Update of Data Published Elsewhere.[18,19]

Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Modified Nucleotides	Force-field[d]	Water Model[e]	References	Year
Single strands									
r(A_3), r(U_3), r(A_6), r(U_6)	m	6	65.0	K$^+$		AMBER	TIP3P	77	2004
r(GA_4C)	m	6	1.5	Na$^+$		AMBER	TIP3P	87	2004
r($CGCU_4GCG$)	m	10	100.0	Na$^+$		CHARMM	TIP3P	88	2007
9 ≠ single strands	m	13	2.1	Na$^+$	Am; Cm; Um	AMBER	TIP3P	89	2003
Hybrids									
DNA.RNA	m	20	2.0	Na$^+$		AMBER	TIP3P	90	1997
	m/x	24	11.0	Na$^+$		AMBER	TIP3P	91	2005
	m	20	2.0	Na$^+$		AMBER	TIP3P	92	2003
	x	20	10.0	Na$^+$		AMBER$_{new}$	TIP3P	46	2007
HNA.RNA	m	16	1.1	Na$^+$	HNA	AMBER	TIP3P	93	1998
MOE.RNA	m	20	1.3	Na$^+$	MOE	AMBER	TIP3P	94	1998
PNA.RNA	m	12	2.5	Na$^+$	PNA	AMBER	TIP3P	95	2000
Amide-3 DNA.RNA	m	12	10.0	NH$_4^+$	Amide-3 linkage; Am; Gm	CHARMM	TIP3P	96	2005
5-propynyl DNA.RNA	m	20	2.0	Na$^+$	5-propynyl	AMBER	TIP3P	92	2003

(Continued)

Table 13.1 (*Continued*)

Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Modified Nucleotides	Force-field[d]	Water Model[e]	References	Year
Duplexes									
ApU and GpC steps (in crystal)	x	8	2.0	Na^+		AMBER	TIP3P	97	1995
$r\{(ApU)_{12}\}_2$	m	48	2.4	0.2M KCl		$AMBER_{ion}$	SPC/E	98	2001
$r\{(CpG)_{12}\}_2$	m	48	2.4	0.2M KCl		$AMBER_{ion}$	SPC/E	99,98–101	2000/1/2
$r\{(CmpGm)_{12}\}_2$	m	48	4.4	0.2M KCl	Cm; Gm	$AMBER_{ion}$	SPC/E	100	2001
$r\{(CmpGm)_3\}_2$	x	12	0.7	Na^+	Cm; Gm	CHARMM	SPC_f	102	2000
$r\{(CpG)_3\}_2$	m								
$r(CCAACGUUGG)_2$	m	20	2.0	Na^+		AMBER	TIP3P	90	1997
$r(CGCGAAUUCGCG)_2$	m	24	11.0	Na^+		AMBER	TIP3P	103	2004
$r(CGCGGAUUCGCG)_2$	m	24	30.0	0.1M NaCl		AMBER	TIP3P	104	2004
$r(GGACUUCGGUCC)_2$	x	24	4.0	0.1M NaCl		AMBER	TIP3P	105	2001
$r(UAAGGAGGUGUA)_2$ (in crystal)	m	24	2.0	Na^+		CHARMM	$TIP3P_m$	47,48	2000
$r(GCCAGUUCGCUGGC)_2$	x	28	3.0	Na^+, 0.1M NaCl	Br^5C	AMBER	TIP3P	106	2002
$r(CGCUGCG)_2$	m	16	5.0	NaCl	F (fluorinated); P (phenyl)	AMBER	TIP3P	107	2004
$r(CGUUACG)_2$									
$r(GAGUACUC)_2$	m	16							
$r(GCGAGUACUCGC)_2$	m	24	5.0	0.3, 1.0M NaCl		CHARMM	$TIP3P_m$	108	2003

(*Continued*)

Table 13.1 (*Continued*)

Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Modified Nucleotides	Force-field[d]	Water Model[e]	References	Year
r(CGCGAUCGCG)$_2$	m	20							
r(CCUUUCGAAAGG)$_2$	m	24							
r(UAAGGAGGUGAU)$_2$	x	24							
r(GGCUGGCC)$_2$									
r(GGCGUGCC)$_2$									
r(GACUGGUC)$_2$									
r(GACGUGUC)$_2$									
r(GGAUGUCC)$_2$	m	16	5.0	0.3, 1.0M NaCl		CHARMM	TIP3P$_m$	109	2005
r(GGAGUUCC)$_2$									
r(GGCUAGCC)$_2$									
r(GGCAUGCC)$_2$									
r(CGCU$_4$GCG)$_2$	m	20	5.0	Na$^+$		CHARMM	TIP3P	88	2007
6 RNA duplexes	m	32	50.0	Na$^+$	I^{2}A/I^{5}C/I^{5}U-phosphoramidite	AMBER	TIP3P	110	2007
G = C rich duplex	m	74	40.0	Na$^+$		AMBER	TIP3P	111	2006
r(GGCGAGCC)$_2$	n	16	0.3	Na$^+$		AMBER	TIP3P	112	2006
r(GCGGACGC)$_2$									
r(CGCGAAUUCGCG)$_2$	m	24	10.0	Na$^+$		AMBER$_{new}$	TIP3P	46	2007
r(GCGAGAGUAGG)/ r(CCGAUGGUAGU)	x	22	10.0	Na$^+$		AMBER$_{new}$	TIP3P	46	2007

(*Continued*)

Table 13.1 (*Continued*)

Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Modified Nucleotides	Force-field[d]	Water Model[e]	References	Year
Bulges									

r(CGCGACGCG/ CGCGCGCG)
r(CGCGUCGCG/ CGCGCGCG)
	m	17	4.0	Na^+		AMBER	TIP3P	113	2006

Hairpins									
r(GGGC[GCAA]GCCU) tetraloop	n	12	0.2	Na^+		OPLS	SPC/E	43	1995
r[GCAA] tetraloops	m	26	3.0	0.1M NaCl		AMBER	TIP3P	114	2000
r[GCAA] tetraloop (folding)	m	12	*NC*	Na^+, Mg^{2+}		AMBER	TIP3P TIP4P	115	2005
r(GC[GAAG]GC) tetraloop (folding)	m	8	2.0	0.3, 0.5M NaCl		AMBER	TIP3P	116	2001
r(GGC[NCAA]GCC) tetraloop	m	10	1.4	Na^+	I	CHARMM	TIP3P	117	2003
r(GGAC[UUCG]GUCC) tetraloop	n	12	2.0	Na^+		AMBER	TIP3P	33,118	1995/7
r(GGCAC[UUCG] GUGCC) tetraloop	m	14	50.0	Na^+		AMBER	TIP3P	119	2006

(*Continued*)

Table 13.1 (*Continued*)

Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Modified Nucleotides	Force-field[d]	Water Model[e]	References	Year
r(CGC[U_4]GCG) tetraloop	n	10	12.0	Na^+		AMBER	TIP3P	120	2005
	m	10	70.0	Na^+		CHARMM	TIP3P	88	2007
r(GAGGUC[O_6] GAUCUC) hexaloop	m	18	23.0	Na^+	O = abasic	AMBER	TIP3P	121	2006
Ribozymes and ribozyme fragments									
Hammerhead ribozyme	x	41	0.8	0.1M NaCl, Mg^{2+}		AMBER	SPC/E	122,123	1997/8
	x	41	1.1	Na^+, Mg^{2+}		AMBER	TIP3P	124,125	1998; 2000
	x	NC	13.0	Na^+, Mg^{2+}		AMBER	TIP3P	126	2005
	x	NC	12.0	0.1M NaCl, Mg^{2+}		CHARMM	TIP3P	127	2007
HDV ribozyme	x	78	17.0	Na^+, Mg^{2+}	C^+	AMBER	TIP3P	61,128	2005
	x	61	20.0	Na^+	C^+	AMBER	TIP3P	129	2007
	x	62	17.0	Na^+, Mg^{2+}	C^+	AMBER	TIP3P	130	2007
HCV IRES IIId domain	n	28	2.6	Na^+, Mg^{2+}		AMBER	TIP3P	131	2004
Hairpin ribozyme	x	*NC*	30.0	Na^+, Mg^{2+}	d(A), d(G)	AMBER	TIP3P	132	2006

(*Continued*)

Table 13.1 (*Continued*)

Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Modified Nucleotides	Force-field[d]	Water Model[e]	References	Year
snRNA and snRNA protein complexes									
U1A heptaloop	n	21	5.0	Na^+, Cl^-		AMBER	TIP3P	133	2002
	n	30	1.0	Na^+		AMBER	TIP3P	134	1999
U4 Kink-turn	x	*NC*	10.0	Na^+		AMBER	*NC*	135	2005
U4 Human snRNA	x	43	10.0	Na^+		AMBER	TIP3P	136	2005
	x	47	74.0	Na^+		AMBER	TIP3P	137, 138	2005/6
U1A RNA/protein complex	x	21	1.0	Na^+		AMBER	TIP3P	134	1999
	x	21	1.8	0.1, 1.0M NaCl		AMBER	SPC/E	139	1999
	x	20	5.0	*NC*		AMBER	TIP3P	140	2005
U1A/SL2 RNA/protein complex	x	21	10.0	Na^+, 0.3M NaCl		AMBER	TIP3P	141	2007
U1A/U1hpII RNA/ protein complex	x	28	2.0	Na^+		AMBER	*NC*	142	2006
U2/hairpin IV RNA/ protein complex	x	23	2.2	K^+		AMBER	TIP3P	143	2001
U4 Human snRNA/ sm binding site	m	62	3.0	K^+		AMBER	TIP3P	144	2000

(*Continued*)

Table 13.1 (*Continued*)

					Modified		Water		
Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Nucleotides	Force-field[d]	Model[e]	References	Year
Telomerase fragments									
RNA hairpin	n	30	20.0	0.1M KCl		AMBER	TIP3P	145	2005
Pseudoknot	m	48	4.0	Na$^+$, 0.1M NaCl		AMBER	NC	146	2006
tRNA, tRNA fragments and tRNA protein complexes									
tRNA$^{\mathrm{Gln}}$ & var-AGGU-tRNA$^{\mathrm{Gln}}$	m	76	6.5	NH$_4^+$		AMBER	TIP3P	147	2007
tRNA$^{\mathrm{Asp}}$	x	76	0.5	NH$_4^+$	D; Ψ; m^{1}G; m^{5}C; m^{5}U	AMBER$_{\mathrm{ion}}$	SPC/E	53,54	1996/9
tRNA$^{\mathrm{Asp}}$ anticodon hairpin	x	17	0.5	NH$_4^+$	Ψ; m^{1}G	AMBER$_{\mathrm{ion}}$	SPC/E	44,55,148	1996/7
tRNA$^{\mathrm{lys,3}}$ anticodon hairpin	x	17	6.0	Na$^+$	mcm^{5}s^{2}U, ms^{2}t^{6}A, ψ	AMBER	TIP3P	149	2006
tRNA$^{\mathrm{Ala}}$ acceptor stem hairpin	m	22	2.5	NH$_4^+$		CHARMM	TIP3P$_{\mathrm{m}}$	150	2002
	m	22	2.0	Na$^+$	I; 2AA; 2AP; IsoC; dU; Z; M; 7DAA	AMBER	TIP3P	151–153 154	1999; 2000/2
tRNA$^{\mathrm{Gln}}$/synthetase	x	76	6.5	NH$_4^+$		AMBER	TIP3P	147	2007

(*Continued*)

Table 13.1 (*Continued*)

Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Modified Nucleotides	Force-field[d]	Water Model[e]	References	Year
Ribosome and ribosomal fragments									
30S *Thermus thermophilus*	x	1.5k	10.0	0.1M KCl,		AMBER	TIP3P	155	2006
70S *Thermus thermophilus*	m	4.5k	4.0	7 mM $MgCl_2$					
16S rRNA core	x	81	5.5	0.1M NaCl		AMBER	TIP3P	156	2003
Helix 44 of 16S rRNA	x	*NC*	30.0	Na^+, Mg^{2+}		AMBER	TIP3P	111	2006
SRD 23S rRNA *E.coli*	x	27	25.0	Na^+		AMBER	TIP3P	157	2006
SRD 28S rRNA rat		29	35.0	Na^+		AMBER	TIP3P		
8 duplexes with G•U mismatches	x	48	10.0	Na^+		AMBER	TIP3P	158	2006
Kink-turns 38, 42, 58	x	38	43.0	Na^+		AMBER	TIP3P	137,159	2004/5
Kink-turns 38, 42, 42+FBS, 58	x	84	79.0	Na^+, K^+, Mg^{2+}		AMBER	TIP3P	138	2006
5S rRNA loop E	x	24	10.0	Na^+, Mg^{2+}		AMBER	TIP3P	160	2003
	x	24	11.5	0.2, 1.0M KCl, Mg^{2+}		AMBER	SPC/E	64,161, 162	2003/4

(*Continued*)

Table 13.1 (*Continued*)

Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Modified Nucleotides	Force-field[d]	Water Model[e]	References	Year
rRNA/protein									
5S rRNA loopE/L25	x	32	24.0	Na^+ & Mg^{2+}		AMBER	TIP3P	163	2004
S15/16S binding site	x	57	15.0	Na^+, & Mg^{2+}		AMBER	TIP3P	164	2007
L11/rRNA *Thermotoga maritima*	x	58	16.0	Na^+, 0.1M NaCl		AMBER	TIP3P	165	2006
rRNA/ligand									
A site/neomycin B	x	21	10.0	Na^+		AMBER	TIP3P	166	2002
A site/neamin, paromomycin & synthetic antibiotic	x	46	1.8	Na^+		AMBER	TIP3P	167,168	2006/7
A site/paromomycin	x	42	25.0	0.2M KCl		$AMBER_{ion}$	SPC/E	67	2006
Viral particles and fragments									
Satellite tobacco mosaic virus (STMV)	m	949	13.0	Cl^-, Mg^{2+}		CHARMM	TIP3P	70	2006
Beet western yellow virus pseudoknot	x	26	5.0	Na^+	C^+	AMBER	TIP3P	60	2001
	x	26	10.0	Na^+		$AMBER_{new}$	TIP3P	46	2007

(*Continued*)

Table 13.1 (*Continued*)

Starting Structures	Type[a]	nt.[b]	Length (ns)[c]	Ions	Modified Nucleotides	Force-field[d]	Water Model[e]	References	Year
H3 kissing loop (murine leukemia virus)	n	28	16.0	Na^+		$AMBER_{ion}$	TIP3P	169	2003
BIV tat-TAR complex	n	28	1.2	Na^+		AMBER	TIP3P	170	2001
HIV fragments									
LAI SL1 hairpin	n	23	10.0	Na^+		AMBER	TIP3P	171,172	2002/3
LAI SL1 extended dimer	x/n	46	NC	Na^+		AMBER	TIP3P	173	2005
LAI SL1 kissing loop	m	70	31.0	*NC*		AMBER	*NC*	174	2007
LAI & MAL kissing loops	m	46	0.4	Na^+		AMBER	TIP3P	175	2002
	x/n	46	20.0	Na^+, Mg^{2+}		AMBER	TIP3P	176	2004
	x	46	7.5	Na^+, Mg^{2+}		$AMBER_{ion}$	TIP3P	169	2003
TAR RNA hairpin	n	14	1.6	Na^+		CHARMM	TIP3P	177	2003
	n	29	20.0	Na^+		CHARMM	TIP3P	178	2006
	n	30	2.0	Mg^{2+}		AMBER	TIP3P	179	2004
TAR RNA/KkN ligand	m	29	20.0	Na^+		AMBER	TIP3P	180	2006
TAR aptamer complex	n	29	3.0	Na^+		CHARMM	$TIP3P_m$	181	2003

(*Continued*)

Table 13.1 (*Continued*)

					Modified		Water		
Starting Structures	**Type**[a]	**nt.**[b]	**Length (ns)**[c]	**Ions**	**Nucleotides**	**Force-field**[d]	**Model**[e]	**References**	**Year**
Others									
FMN aptamer	n	35	1.7	Na^+		AMBER	TIP3P	182	1999
RNA nanotube	m	21	13.0	0.1M NaCl		AMBER	TIP3P	183	2007
dsRBD/dsRNA RNA/ protein complex	n	30	2.0	Na^+, Cl^-		AMBER	TIP3P	184	2002

[a] Indicates the starting structure type: "m" for model, "n" for NMR and "x" for X-ray. Note that sometimes it is difficult to distinguish model from NMR or X-ray structures since models might sometimes include experimentally solved structural fragments.

[b] Indicative size (in nucleotides) of the largest simulated RNA fragment.

[c] Indicative length (in ns) of the longest described simulation.

[d] $AMBER_{ion}$ indicates that different ion parameters were used; $AMBER_{new}$ indicates that the most recent AMBER force field was used.[49]

[e] $TIP3P_m$ corresponds to a modified TIP3P model; SPC_f to a flexible SPC water model.

13.2 Pre-Ewald Times

Despite its large size, the tRNA^{Phe} molecule became the first target for RNA molecular dynamics simulations, since in 1983,[4] it was the only RNA molecule of significant size for which crystallographic coordinates were available. These pioneering simulations never exceeded the 24 ps time scale under *in vacuo* conditions. They were followed by a few studies using MD techniques to help solve NMR structures,[22–24] and by a remarkable work describing the dynamics of a model of the domain II of the HIV Rev responsive element (RRE) containing GG and GA non-Watson-Crick base pairs[25] under both *in vacuo* and *in aquo* conditions. In 1989, an *in vacuo* study trying to elucidate the catalytic mechanism of the hammerhead ribozyme based on a model-built structure was also published.[26] Note that the first hammerhead crystal structure was published in 1994.[27] Similarly, a model-based MD simulation of a bacteriophage intron fragment of 112 nucleotides was published in 1990.[28] Unfortunately, these simulations failed to provide useful insight into the dynamics of the "real" systems since the starting model structures were by far too imprecise.[19]

Moreover, in order to limit the computational involvement of such simulations, long-range electrostatic interactions were usually truncated around 8Å. With this approximation, some attempts made in 1995 to simulate the dynamics of the tRNA^{Asp} anticodon hairpin failed despite the inclusion in the model of water molecules and a neutralizing atmosphere of NH_4^+ cations. The hallmarks of this failure were related to a rapid degradation, on the 100 ps time scale, of the important tertiary interactions of the anticodon loop present in the crystallographic structure.[29] A subsequent study that used a 16 Å truncation value, revealed an increased stability of the anticodon hairpin on the 100 ps time frame associated with attenuated structural degradations. Despite remaining structural instabilities, these simulations illustrated the unexpected contribution of solvent molecules located at long distances from the solute to the stability of biomolecular structural motifs, and pointed to the importance of long-range solvation forces.[30,31] Further simulations using Ewald techniques (see next section) led to very stable simulations demonstrating that, for generating stable trajectories of folded RNA molecules, both the solvent and the long-range electrostatic interactions had to be taken into account in the calculations.[30,32] These findings stigmatized the limitations of some of the major approximations used at that time.

13.3 Ewald Summation Methods

As noted in the introduction, Ewald summation methods revolutionized the MD field at numerous levels.[14] Specifically, these methods were important for simulations of RNA systems that were shown to be especially sensitive to the treatment of the long-range electrostatic forces.[32,33] Although Ewald methods were marginally used in simulations of DNA systems,[34] the development and implementation, in 1995, of the cost-efficient particle mesh Ewald (PME) summation method[35] in the widely used AMBER molecular dynamics simulation package[36] boosted the field and led to the generation of stable trajectories of RNA systems on the nanosecond time-scale.[30,32,33]

Despite the popularity of Ewald summation methods, some alternative techniques for the treatment of long-range electrostatic interactions such as shifted truncation methods[37–42] are still in use. Other promising methods for the efficient treatment of long-range electrostatic interactions are currently being developed[14] but have not yet been used in MD simulations of RNA systems.

13.4 The Ewald Era

Nowadays, the largest number of MD simulations of RNA systems makes use of Ewald summation techniques. Yet, it is almost impossible to detail each of the about 110 publications referenced in Table 13.1. Since the first MD simulations, which had as their main purpose to establish the applicability of the Ewald methods for simulating RNA molecules,[33,43,44] some records have been established both in terms of system size (from dinucleotides to ribosome's) and time scales (from 0.1 to over 100 ns). The diversity of the investigated systems is important and comprises single stranded RNA particles, synthetic hybrids, duplexes, hairpins, kissing loops, kink-turns, ribozymes, pseudoknots, complete tRNAs and tRNA fragments, ribosomes and ribosomal motifs, RNA/antibiotic and RNA/protein complexes, and a complete viral particle. Most of the starting structures used by these studies are derived from crystallographic or NMR investigations in their original or mutated form. Other starting structures are models. Some of these simulations include synthetic or naturally modified nucleotides and are conducted either in a cationic neutralizing aqueous atmosphere (minimal salt) or in an aqueous environment mimicking various ionic conditions.[3] The largest part (≈80%) of these trajectories has been generated by using the AMBER simulation package. Since 1995, the number of publications related to this field is increasing steadily (Fig. 13.1). This raise in production is determined by the combination of three facts: the implementation of the fast Ewald methods for

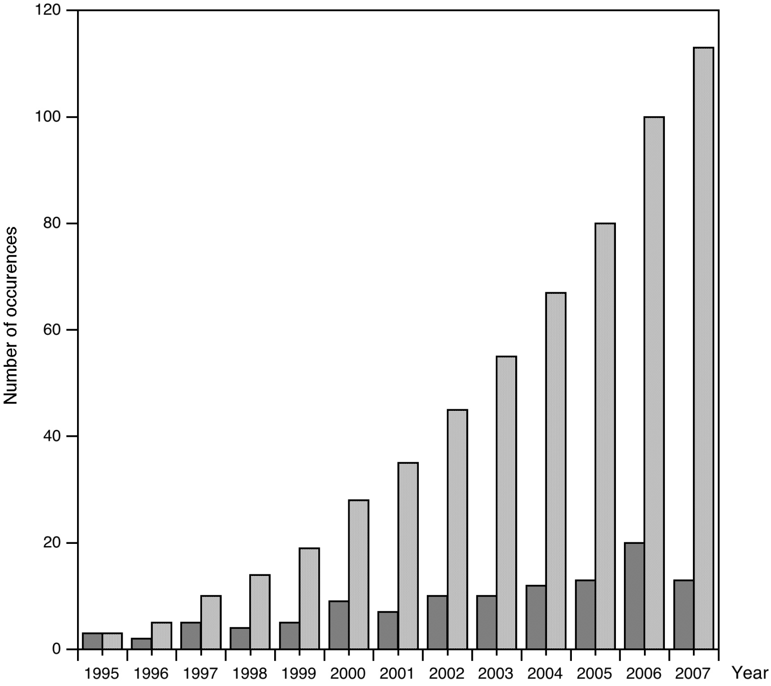


Fig. 13.1 Number of MD studies using an explicit representation of the solvent and Ewald summation methods from 1995 (origin) to September 2007 (see Table 13.1). The light histogram is a cumulative view of the yearly number of published MD studies of RNA systems (darker histogram).

the treatment of the long-range electrostatic interactions; the increase of available crystal structures (in 2007, more than 800 biomolecular structures including RNA nucleotides were deposited in the Nucleic acid DataBase or NDB); and the growing interest of the scientific community for this class of nucleic acids. Some successes and failures of these techniques will be described next.

13.4.1 *Structural Stability and Force Fields*

The history of MD simulations is marked by the requirement to constantly improve existing methodologies and force fields in order to generate dynamically stable biomolecular trajectories on the longer time scales. If, in 1995, the 1 ns time limit could be reached for RNA systems, the field is now facing other limitations associated with backbone parameters. For DNA trajectories generated by using the AMBER parm99 force-field[45] and the particle-mesh-Ewald summation method, it has been shown that the investigated duplexes started to lose their structure in a stepwise manner after 10 ns of simulated time. These structural alterations were characterized by an important increase in α/γ transitions, most of them being irreversible.[46] A new version of the AMBER force field has been developed in order to address this issue, and preliminary data seem to indicate that this new force field behaves well for RNA systems. Other force fields for nucleic acids are currently being developed and will have to be thoroughly tested on these longer time scales.[47–49] These studies remind us that great caution must be exerted in the interpretation of spontaneous structural transitions occurring on the longer time scales in MD trajectories.

13.4.2 *Solvent and Ion Parameters*

One of the greatest achievements of MD studies resides in the assessment that the solvent, composed of water molecules, mono- and di-valent cations, as well as anions, plays a determining role in the structure of RNA systems.[3] All the studies referenced in Table 13.1

use at least a minimal ionic atmosphere composed of Na^+, K^+, or NH_4^+ cations in order to neutralize the charges carried by the anionic RNA backbone, while others include in their system divalent cations (Mg^{2+}) and/or an excess of salt (NaCl or KCl) at concentrations ranging from 0.1 to 1.0 M.

Hence, it is of great importance to have reliable parameter sets available for the solvent particles surrounding nucleic acids. Numerous water models have been developed (TIP3P, TIP4P, SPC/E, etc.) and some of the most popular ones have been repeatedly tested.[50] Similarly, various parameter sets for ions have been developed and used in MD simulations. Parameters developed by Åqvist[51] have been integrated in the AMBER force field. Unfortunately, recent investigations demonstrated that, when the solute was surrounded by an excess of (> 0.2M) salt, KCl or NaCl aggregates formed spontaneously, pointing to an imbalance of the parameters solvent (water and ion) implemented in the AMBER force field. Simulations conducted with parameters developed by Dang[52] did not exhibit such unphysical behavior and were therefore considered better candidates for realistically simulating the ionic environment around nucleic acids. Simulations using the AMBER force field and the latter, or different ion parameters, are rare and are noted $\text{AMBER}_{\text{ion}}$ in Table 13.1. Again, one has to be aware of the limitations of current force fields in order not to over interpret data.

13.4.3 *tRNA and Modified Nucleotides*

tRNA molecules and molecular fragments have been historically very important in the development of MD simulations of RNA systems. The first *in vacuo* MD simulations were associated with tRNA^{Phe} molecules while some of the first MD simulations using Ewald summation methods were devoted to the entire tRNA^{Asp} molecule[53,54] and its anticodon hairpin.[30,55] Recently, the binding affinity of *E. coli* tRNA^{Gln} to glutaminyl-tRNA synthetase was investigated. Other studies focused on the tRNA^{Ala} acceptor stem. Most of these simulations took into consideration the modified nucleotides that are part of the maturated tRNAs. Among those, the most commonly occurring modification is pseudouridine[56] followed by 2′-O-Me or Nm nucleotides.[57] MD simulations were able to provide some clues confirming experimental findings stating that these modified residues are stabilizing RNA structures through encaging water molecules with long-residency times.[55,56]

Since the interest for tRNA molecules is still vivid, and since modified nucleotides play a role in numerous other RNA structures including the largest ones (ribosomes), a study devoted to the parameterization of the 107 currently known naturally modified nucleotides has been published[58]; see also.[59] In addition, one has to be aware of a very specific type of "modification" that is associated with protonated residues, often not detected in crystallographic structures and for which parameters have rarely been developed.[60,61] Other studies dealing with modified backbones resulted in the parameterization of these unusual residues that are of importance in the development of antisense strategies (Table 13.1).

13.4.4 *Ribozymes*

The first crystallographic structures of a ribozyme (catalytic RNA), a minimal hammerhead construct, led to the generation of several MD simulations having the aim to elucidate its catalytic mechanism without, however, much success for two main reasons. First, these crystallized structures were different from the active form of the hammerhead ribozyme that was solved recently.[62] Second, and not least importantly, the crystal structure on which most studies were based[63] suffered from a wrong solvent density interpretation. Indeed, some Mg^{2+} cations were assigned to electropositive locations close to nucleic acid bases that were shown to correspond to SO_4^{2-} binding sites.[64] Note that these hammerheads were crystallized in 1.8M Li_2SO_4 salt conditions. Hence, it is most probable that the participation of Mg^{2+} cations (if any) is much smaller than initially thought based on the above-mentioned interpretations of the crystal structures. Consequently, great caution should be taken in choosing the starting structures on which subsequent MD studies will be based.[19]

Other ribozymes or ribozyme fragments (Table 13.1), like those of the hepatitis delta virus (HDV), hepatitis C virus (HCV), and hairpin ribozymes, are currently being investigated, and some clues related to their internal dynamics and possible catalytic implications are being presented.

13.4.5 *Ribosomes and Ribosomal Fragments*

Ribosome systems are currently the biggest and most complex systems that have been simulated today, and this represents a considerable achievement for simulation techniques.[65] The published simulations are based on crystallographic structures of the small ribosomal subunit or on model structures of the 70S particle. This last system comprises approximately 2.64 million atoms if one counts its RNA, protein, and solvent parts, and it gave birth, to the best of our knowledge, to the largest all-atom biomolecular simulation published to date.

Of course, in order to better understand the whole ribosomal picture, the dynamics of important fragments such as kink-turns, bulges, helices, and proteins have also to be better understood. In this respect, besides classical MD studies,[21] the energy landscape of the ribosomal decoding center has been investigated by using replica MD methods[65,66] and the role of water molecules in aminoglycoside binding has been described.[67] Besides, the reaction mechanism of the peptide bond formation,[68] as well as the energetics of the codon-anticodon recognition,[69] have been studied, though without using Ewald techniques.

13.4.6 *Viral Particles*

Another main achievement of the use of theoretical methods, which goes far beyond the somewhat "narrow" but expanding field of RNA simulations, is related to the first all-atom simulation of a complete viral particle.[70] This satellite tobacco mosaic virus (STMV) consists of an icosahedral capsid composed of 60 identical copies of a single protein and a 1.058 kb RNA genome coding for that protein and a second protein of unknown function (note that the simulated

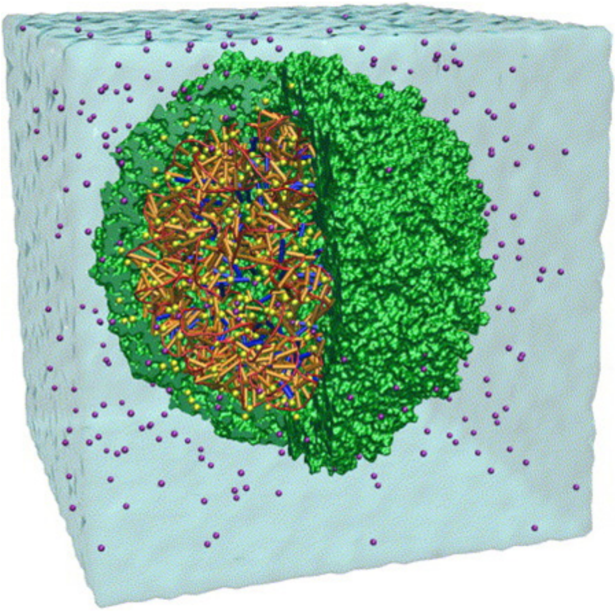


Fig. 13.2 Schematic representation of the STMV system. The protein capsid (green) is enveloping the RNA (red and orange). The ions are drawn in yellow (magnesium) and purple (chloride). Reproduced with permission from Freddolino PL, Arkhipov AS, Larson SB, McPherson A, Schulten K. (2006) *Structure* **14**: 437–449. © Cell Press.

nucleotide sequence is artificial). This system was solvated by ~300.000 water molecules and 773 ions that represent about 84% of the 1.06 million simulated atoms (Fig. 13.2). Among other details, the simulation revealed that although the virus looks symmetrical, it pulses in and out asymmetrically in breathing-like motion. This simulation represents, to the best of our knowledge, the first all-atom simulation of a "complete" life form. Besides, significant efforts are made to investigate the dynamical behavior of some important HIV motifs (Table 13.1).

13.5 Dynamic Models

Next to classical MD simulations, a large number of MD simulations make use of "targeting" strategies in order to induce specific conformational changes[71] or local enhanced sampling (LES)[72] techniques and replica methods (REMD)[72] to increase conformational sampling.[73] Such simulations were not integrated in Table 13.1 since they do not, in our opinion, correspond to the definition of an MD simulation. In MD simulations, one defines specific conditions comprising, among other things, a starting structure (model, X-ray, NMR or mixed); a representation of the solvent, and a representation of the interatomic forces at play in biomolecular systems.[19] On the opposite, targeting strategies voluntarily impose artificial forces that have the aim to drive the system from one state to another. During such a process, it is difficult to observe spontaneous conformational changes. Moreover, the time-scales of the simulated processes are certainly far from the "real" ones. Hence, they should rather be called "dynamic models." It is indeed important to establish such a distinction in order to avoid bringing additional confusion to this already complex field.[74]

13.6 Future Outlook

If one looks back at the brief history of MD simulations, it appears that evolution occurred mostly stepwise through the development of new techniques that helped in eliminating approximations thanks to the steady increase of available computational power. Indeed, the main purpose of approximations in computational science is to be able to artificially "overcome" size and time scale limits. Yet, approximations often considerably degrade the reliability of the data that are gathered through the simulation techniques. Nevertheless, approximations are currently needed to reach the longer time scales that are currently in the 100 ns range, and to create dynamic models for systems of the size of a ribosome.[75]

In the future, the field will have to deal with "complex" issues, but at the same time, also with less "impressive" or less obviously "complex" ones. Among the most impressive ones, we will probably create more models on RNA folding, simulations of large size particles,[65] and gain better insight into molecular recognition issues related to RNA/drug interactions.[67,76] For example, the electrophoretic transport of single-stranded RNA molecules through 1.5 nm-wide pores of carbon nanotube membranes has been investigated by using MD techniques.[77] Such studies are mainly driven by experimentalist's demands.[21]

"Less" spectacular issues are associated with uncovering some of the basic principles associated with molecular recognition phenomena. Besides hydrogen-bonds and the less-well defined stacking interactions, numerous other intermolecular interactions remain to be uncovered in the biomolecular field and integrated in current force fields. For instance, the unfrequent halogen bonds involving an interaction between an electron acceptor C–X (X = Cl, Br, I) group and an electron donor O = C partner were only recently described in biomolecular systems.[78] Although rare, one has to take these interactions into account if he or she wants to understand recognition phenomena involving halogenated drugs or substrates.

In that perspective, polarizable force fields will certainly become more and more popular since they allow for improved adaptation towards "environmental changes." Promising results on DNA systems have been published recently.[79] However, new and probably tedious parameterization studies will have to be undertaken.

At a more "elaborate" level, density function theory (DFT)-based Car-Parrinello MD techniques or other hybrid quantum mechanical/molecular mechanical (QM/MM) methods will allow us to describe the time evolution of molecular systems without resorting to a predefined potential energy surface[80] and also allow the addressing of more subtle issues such as those related to a better understanding of ribozyme catalytic mechanisms.[81–84]

Last, but not least importantly, it is essential to develop tools for comparing MD simulation results with experimental data. The SwS or "Solvation web Service for nucleic acids"[85,3] has been designed to provide an exhaustive overview of the solvation of nucleic acid structural elements through the generation of 3D solvent density maps. It is only through such confrontations (Fig. 13.3) and through the detection of possible "discrepancies" that available force fields and methodologies will be able to evolve towards new levels of "realisms". For

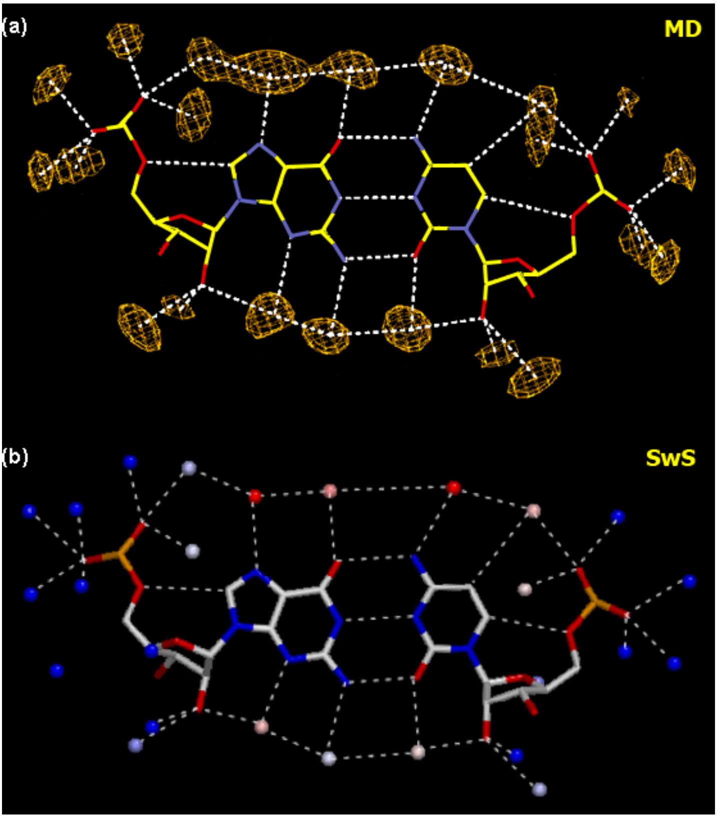


Fig. 13.3 Comparison of water-binding sites derived from **(a)** MD simulations of a solvated r(GC)$_{12}$ duplex and **(b)** a statistical analysis of r(G = C) pairs extracted from all NDB nucleic acid structures with resolutions equal to or below 3.0 Å using the SwS web service, available at http://www-ibmc.u-strasbg.fr/arn/sws.html.

achieving such a purpose, a good choice of the MD starting structures and conditions is crucial.[19]

Acknowledgments

We wish to thank all the authors of RNA molecular simulations that shared their knowledge with us. Y. Hashem was supported by a grant from the French ministry of research (MRT).

References

1. McCammon JA. (1976) Molecular dynamics study of the bovine pancreatic trypsin inhibitor. In *Models for Protein Dynamics*, pp. 137. CECAM, Orsay, France.
2. McCammon JA, Gelin BR, Karplus M. (1977) Dynamics of folded proteins. *Nature* **267**: 585–590.
3. Auffinger P, Hashem Y. (2007) Nucleic acid solvation: from outside to insight. *Curr Opin Struct Biol* **17**: 325–333.
4. McCammon JA, Harvey SC. (1987) *Dynamics of Proteins and Nucleic Acids.* Cambridge University Press, New York.
5. Adcock SA, McCammon JA. (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev* **106**: 1589–1615.
6. Duan Y, Kollman PA. (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**: 740–744.
7. Levitt M. (1983) Computer simulations of DNA double helix dynamics. *Cold Spring Harb Symp Quant Biol* **47**: 251–262.
8. Tidor B, Irikura KK, Brooks BR, Karplus M. (1983) Dynamics of DNA oligomers. *J Biomol Struct Dyn* **1**: 231–252.
9. Prabhakaran M, Harvey SC, Mao B, McCammon JA. (1983) Molecular dynamics of phenylalanine transfer RNA. *J Biomol Struct Dyn* **1**: 357–369.
10. Seibel GL, Singh UC, Kollman PA. (1985) A molecular dynamics simulation of double-helical B-DNA including counterions and water. *Proc Natl Acad Sci USA* **82**: 6537–6540.
11. van Gunsteren WF, Berendsen HJC, Geurtsen RG, Zwinderman HRJ. (1986) A molecular dynamics computer simulation of an eight-base-pair DNA fragment in aqueous solution: comparison with experimental two-dimensional NMR data. *Proc Natl Acad Sci USA* **482**: 287–303.
12. Orozco M, Luque FJ. (2000) Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem Rev* **100**: 4187–4226.
13. Feig M, Chocholousova J, Tanizaki S. (2006) Extending the horizon: towards the efficient modeling of large biomolecular complexes in atomic detail. *Theoret Chem Acc* **116**: 194–205.
14. Koehl P. (2006) Electrostatics calculations: latest methodological advances. *Curr Opin Struct Biol* **16**: 142–151.
15. Norberg J, Nilsson L. (2003) Advances in biomolecular simulations: methodology and recent applications. *Q Rev Biophys* **36**: 257–306.
16. van Gunsteren WF, Bakowies D, Baron R, *et al.* (2006) Biomolecular modeling: goals, problems, perspectives. *Angew Chem Int Ed Engl* **45**: 4064–4092.
17. Zacharias M. (2000) Simulation of the structure and dynamics of nonhelical RNA motifs. *Curr Opin Struct Biol* **10**: 311–317.

18. Auffinger P, Vaiana AC. (2005) Molecular dynamics simulations of RNA systems. In RK Hartmann, A Bindereif, A Schön, E Westhof (eds.), *Handbook of RNA Biochemistry*, pp. 560–576. Willey-VCH, Manheim.
19. Auffinger P. (2006) Molecular dynamics simulations of RNA systems: importance of the initial conditions. In J Sponer, F Lankas (eds.), *Computational Studies of DNA and RNA*, pp. 283–300. Springer Verlag.
20. Sponer J, Lankas F. (2006) *Computational Studies of RNA and DNA*. Springer, The Netherlands.
21. McDowell SE, Spackova N, Sponer J, Walter NG. (2007) Molecular dynamics simulations of RNA: an *in silico* single molecule approach. *Biopolymers* **85**: 169–184.
22. Happ CS, Happ E, Nilges M, Gronenborn AM, Core GM. (1988) Refinement of the solution structure of the ribonucleotide 5′(GCAUGC)$_2$: combined use of nuclear magnetic resonance and restrained molecular dynamics. *Biochemistry* **27**: 1735–1743.
23. Davis PW, Thurmes W, Tinoco I. (1993) Structure of a small RNA hairpin. *Nucl Acids Res* **21**: 537–545.
24. Agback P, Sandstrom A, Yamakage S, *et al.* (1993) Solution structure of lariat RNA by 500 MHz NMR spectroscopy and molecular dynamics studies in water. *J Biochem Biophys Meth* **27**: 229–259.
25. Le S, Pattabiraman N, Maizel JV. (1994) RNA tertiary structure of the HIV RRE domain II containing non-Watson-Crick base pairs GG and GA: molecular modeling studies. *Nucl Acids Res* **22**: 3966–3976.
26. Mei HY, Kaaret TW, Bruice TC. (1989) A computational approach to the mechanism of self-cleavage of hammerhead RNA. *Proc Natl Acad Sci USA* **86**: 9727–9731.
27. Pley HM, Flaherty KM, McKay DB. (1994) Three-dimensional structure of a hammerhead ribozyme. *Nature* **372**: 68–74.
28. Nilsson L, Ahgren-Stalhandske A, Sjogren AS, Hahne S, Sjoberg BM. (1990) Three-dimensional model and molecular dynamics simulation of the active site of the self-splicing intervening sequence of the bacteriophage T4 nrdB messenger RNA. *Biochemistry* **29**: 10317–10322.
29. Auffinger P, Louise-May S, Westhof E. (1995) Multiple molecular dynamics simulations of the anticodon loop of tRNA$^{\mathrm{Asp}}$ in aqueous solution with counterions. *J Am Chem Soc* **117**: 6720–6726.
30. Auffinger P, Louise-May S, Westhof E. (1996) Molecular dynamics simulations of the anticodon hairpin of tRNA(asp): structuring effects of C–H···O hydrogen bonds and of long-range hydration forces. *J Am Chem Soc* **118**: 1181–1189.
31. Leckband D, Israelachvili J. (2001) Intermolecular forces in biology. *Q Rev Biophys* **34**: 105–267.
32. Louise-May S, Auffinger P, Westhof E. (1996) Calculation of nucleic acid conformation. *Curr Opin Struct Biol* **6**: 289–298.

33. Cheatham TE, Miller JL, Fox T, Darden TA, Kollman PA. (1995) Molecular dynamics simulations on solvated biomolecular systems: the particle mesh Ewald method leads to stable trajectories of DNA, RNA, and proteins. *J Am Chem Soc* **117**: 4193–4194.
34. Mohan V, Smith PE, Pettitt BM. (1993) Evidence for a new spine of hydration: solvation of DNA triple helices. *J Am Chem Soc* **115**: 9297–9298.
35. Sagui C, Darden TA. (1999) Molecular dynamics simulations of biomolecules: long-range electrostatic effects. *Ann Rev Biophys Struct* **28**: 155–179.
36. Case DA, Cheatham TE, Darden T, *et al.* (2005) The Amber biomolecular simulation programs. *J Comput Chem* **26**: 1668–1688.
37. Norberg J, Nilsson L. (1996) Constant pressure molecular dynamics simulations of the dodecamers d(GCGCGCGCGCGC)$_2$ and r(GCGCGCGCGCGC)$_2$. *J Chem Phys* **104**: 6052–6057.
38. Tang Y, Nilsson L. (1999) Molecular dynamics simulations of the complex between human U1A protein and hairpin II of U1 small nuclear RNA and of free RNA in solution. *Biophys J* **77**: 1284–1305.
39. Lahiri A, Nilsson L. (2000) Molecular dynamics of the anticodon domain of yeast tRNA(Phe): codon-anticodon interaction. *Biophys J* **79**: 2276–2289.
40. Sarzynska J, Kulinski T, Nilsson L. (2000) Conformational dynamics of a 5S rRNA hairpin domain containing loop D and a single nucleotide bulge. *Biophys J* **79**: 1213–1227.
41. Hart K, Nystrom B, Ohman M, Nilsson L. (2005) Molecular dynamics simulations and free energy calculations of base flipping in dsRNA. *RNA* **11**: 609–618.
42. Nystrom B, Nilsson L. (2007) Molecular dynamics study of intrinsic stability in six RNA terminal loop motifs. *J Biomol Struct Dyn* **24**: 525–536.
43. Zichi DA. (1995) Molecular dynamics of RNA with the OPLS force field. Aqueous simulation of a hairpin containing a tetranucleotide loop. *J Am Chem Soc* **117**: 2957–2969.
44. Auffinger P, Westhof E. (1996) H-bond stability in the tRNA^{Asp} anticodon hairpin: 3 ns of multiple molecular dynamics simulations. *Biophys J* **71**: 940–954.
45. Cornell WD, Cieplak P, Bayly CI, *et al.* (1995) A second-generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* **117**: 5179–5197.
46. Perez A, Marchan I, Svozil D, *et al.* (2007) Refinement of the amber force field for nucleic acids. Improving the description of {alpha}/{gamma} conformers. *Biophys J* **92**: 3817–3829.
47. Foloppe N, MacKerell AD. (2000) All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J Comput Chem* **21**: 88–104.
48. MacKerell AD, Banavali N. (2000) All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solutions. *J Comput Chem* **21**: 105–120.

49. Soares TA, Hunenberger PH, Kastenholz MA, *et al.* (2005) An improved nucleic acid parameter set for the GROMOS force field. *J Comput Chem* **26**: 725–737.
50. Nutt DR, Smith JC. (2007) Molecular dynamics simulations of proteins: can the water explicit water model be varied? *J Chem Theory Comput* **3**: 1550–1560.
51. Åqvist J. (1990) Ion-water interaction potentials derived from free energy perturbation simulations. *J Phys Chem* **94**: 8021–8024.
52. Vaiana AC, Cheatham TE, Auffinger P. (2007) Spontaneous formation of KCl aggregates in biomolecular simulations: a force field issue? *J Chem Theory Comput* **3**: 1851–1859.
53. Auffinger P, Louise-May S, Westhof E. (1996) Hydration of C–H groups in tRNA. *Farad Discuss* **103**: 151–174.
54. Auffinger P, Louise-May S, Westhof E. (1999) Molecular dynamics simulations of the solvated yeast tRNA(Asp). *Biophys J* **76**: 50–64.
55. Auffinger P, Westhof E. (1997) RNA hydration: three nanoseconds of multiple molecular dynamics simulations of the solvated tRNA$^{\mathrm{Asp}}$ anticodon hairpin. *J Mol Biol* **269**: 326–341.
56. Auffinger P, Westhof E. (1998) Effects of pseudouridylation on tRNA hydration and dynamics: a theoretical approach. In H Grosjean, R Benne (eds.), *Modification and Editing of RNA*, pp. 103–112. American Society for Microbiology, Washington, DC (2005).
57. Auffinger P, Westhof E. (1998) Location and distribution of modified nucleotides in tRNA. In H Grosjean, R Benne (eds.), *Modification and Editing of RNA*, pp. 569–576. American Society for Microbiology, Washington, DC (2005).
58. Aduri R, Psciuk BT, Saro P, *et al.* (2007) AMBER force field parameters for the naturally occurring modified nucleosides in RNA. *J Chem Theory Comput* **3**: 1464–1475.
59. Mayaan E, Moser A, MacKerell AD, Jr, York DM. (2007) CHARMM force field parameters for simulation of reactive intermediates in native and thio-substituted ribozymes. *J Comput Chem* **28**: 495–507.
60. Csaszar K, Spackova N, Stefl R, Sponer J, Leontis NB. (2001) Molecular dynamics of the frame-shifting pseudoknot from beet western yellow virus: the role of non-Watson-Crick base-pairing, ordered hydration, cation binding, and base mutations on stability and unfolding. *J Mol Biol* **313**: 1073–1091.
61. Krasovska MV, Sefcikova J, Spackova N, Sponer J, Walter NG. (2005) Structural dynamics of precursor and product of the RNA enzyme from the hepatitis delta virus as revealed by molecular dynamics simulations. *J Mol Biol* **351**: 731–748.
62. Martick M, Scott WG. (2006) Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* **126**: 309–320.

63. Scott WG, Finch JT, Klug A. (1995) The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell* **81**: 991–1002.
64. Auffinger P, Bielecki L, Westhof E. (2004) Anion binding to nucleic acids. *Structure* **12**: 379–388.
65. Sanbonmatsu KY. (2006) Using computer simulations to study decoding by the ribosome. In J Sponer, F Lankas (eds.), *Computational Studies of DNA and RNA*, pp. 327–342. Springer Verlag.
66. Sanbonmatsu KY. (2006) Energy landscape of the ribosomal decoding center. *Biochimie* **88**: 1053–1059.
67. Vaiana AC, Westhof E, Auffinger P. (2006) A molecular dynamics simulation study of an aminoglycoside/A-site RNA complex: conformational and hydration patterns. *Biochimie* **88**: 1061–1073.
68. Trobro S, Aqvist J. (2006) Analysis of predictions for the catalytic mechanism of ribosomal peptidyl transfer. *Biochemistry* **45**: 7049–7056.
69. Almlof M, Ander M, Aqvist J. (2007) Energetics of codon-anticodon recognition on the small ribosomal subunit. *Biochemistry* **46**: 200–209.
70. Freddolino PL, Arkhipov AS, Larson SB, McPherson A, Schulten K. (2006) Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **14**: 437–449.
71. Sanbonmatsu KY, Joseph S, Tung CS. (2005) Simulating movement of tRNA into the ribosome during decoding. *Proc Natl Acad Sci USA* **102**: 15854–15859.
72. Cheng X, Cui G, Hornak V, Simmerling C. (2005) Modified replica exchange simulation methods for local structure refinement. *J Phys Chem B* **109**: 8220–8230.
73. Christen M, van Gunsteren WF. (2008) On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: a review. *J Comput Chem* **29**: 157–166.
74. Coveney PV, Fowler PW. (2005) Modeling biological complexity: a physical scientist's perspective. *J R Soc Interface* **2**: 267–280.
75. Feig M, Brooks CL, 3rd. (2004) Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr Opin Struct Biol* **14**: 217–224.
76. Jorgensen WL. (2007) The many roles of computation in drug discovery. *Science* **303**: 1813–1818.
77. Yeh IC, Hummer G. (2004) Diffusion and electrophoretic mobility of single-stranded RNA from molecular dynamics simulations. *Biophys J* **86**: 681–689.
78. Auffinger P, Hays FA, Westhof E, Ho PS. (2004) Halogen bonds in biological molecules. *Proc Natl Acad Sci USA* **101**: 16789–16794.
79. Babin V, Baucom J, Darden TA, Sagui C. (2006) Molecular dynamics simulations of DNA with polarizable force fields: convergence of an ideal

B-DNA structure to the crystallographic structure. *J Phys Chem B* **110**: 11571–11581.

80. Dal Peraro M, Ruggerone P, Raugei S, Gervasio FL, Carloni P. (2007) Investigating biological systems using first principles Car-Parrinello molecular dynamics simulations. *Curr Opin Struct Biol* **17**: 149–156.
81. Boero M, Terakura K, Tateno M. (2002) Catalytic role of metal ion in the selection of competing reaction paths: a first principles molecular dynamics study of the enzymatic reaction in ribozyme. *J Am Chem Soc* **124**: 8949–8957.
82. Leclerc F, Karplus M. (2006) Two-metal-ion mechanism for hammerhead-ribozyme catalysis. *J Phys Chem B* **110**: 3395–3409.
83. Lopez CS, Faza ON, Gregersen BA, *et al.* (2004) Pseudorotation of natural and chemically modified biological phosphoranes: implications for RNA catalysis. *Chemphyschemistry* **5**: 1045–1049.
84. Min D, Xue S, Li H, Yang W. (2007) "In-line attack" conformational effect plays a modest role in an enzyme-catalyzed RNA cleavage: a free energy simulation study. *Nucl Acids Res* **35**: 4001–4006.
85. Auffinger P, Hashem Y. (2007) SwS: a solvation web service for nucleic acids. *Bioinformatics* **23**: 1035–1037, http://www-ibmc.u-strasbg.fr/arn/sws.html.
86. Hutter J, Carloni P, Parrinello M. (1996) Nonempirical calculations of a hydrated RNA duplex. *J Am Chem Soc* **118**: 8710–8712.
87. Isaksson J, Acharya S, Barman J, Cheruku P, Chattopadhyaya J. (2004) Single-stranded adenine-rich DNA and RNA retain structural characteristics of their respective double-stranded conformations and show directional differences in stacking pattern. *Biochemistry* **43**: 15996–16010.
88. Deng NJ, Cieplak P. (2007) Molecular dynamics and free energy study of the conformational equilibria in the UUUU RNA hairpin. *J Chem Theory Comput* **3**: 1435–1450.
89. Kaukinen U, Lonnberg H, Perakyla M. (2003) Stabilisation of the transition state of phosphodiester bond cleavage within linear single-stranded oligoribonucleotides. *Org Biomol Chem* **2**: 66–73.
90. Cheatham TE, Kollman PA. (1997) Molecular dynamics simulations highlights the structural differences among DNA:DNA, RNA:RNA, and DNA:RNA hybrid duplexes. *J Am Chem Soc* **119**: 4805–4825.
91. Noy A, Perez A, Marquez M, Luque FJ, Orozco M. (2005) Structure, recognition properties, and flexibility of the DNA.RNA hybrid. *J Am Chem Soc* **127**: 4910–4920.
92. Gyi JI, Gao D, Conn GL, *et al.* (2003) The solution structure of a DNA*RNA duplex containing 5-propynyl U and C; comparison with 5-Me modifications. *Nucl Acids Res* **31**: 2683–2693.
93. De Winter H, Lescrinier E, Van Aerschot A, Herdewijn PC. (1998) Molecular dynamics simulation to investigate differences in minor groove hydration of

HNA/RNA hybrids as compared to HNA/DNA complexes. *J Am Chem Soc* **120**: 5381–5394.

94. Lind KE, Mohan V, Manoharan M, Ferguson DM. (1998) Structural characteristics of 2′-O-(2-methoxyethyl)-modified nucleic acids from molecular dynamics simulations. *Nucl Acids Res* **26**: 3694–3699.
95. Soliva R, Sherer E, Luque FJ, Laughton CA, Orozco M. (2000) Molecular dynamics simulations of PNA.DNA and PNA.RNA duplexes in aqueous solution. *J Am Chem Soc* **122**: 5997–6008.
96. Nina M, Fonne-Pfister R, Beaudegnies R, *et al.* (2005) Recognition of RNA by amide modified backbone nucleic acids: molecular dynamics simulations of DNA-RNA hybrids in aqueous solution. *J Am Chem Soc* **127**: 6027–6038.
97. Lee H, Darden T, Pedersen L. (1995) Accurate crystal molecular dynamics simulations using particle-mesh-Ewald: RNA dinucleotides — ApU and GpC. *Chem Phys Lett* **243**: 229–235.
98. Auffinger P, Westhof E. (2001) Water and ion binding around r(UpA)$_{12}$ and d(TpA)$_{12}$ oligomers — comparison with RNA and DNA (CpG)$_{12}$ duplexes. *J Mol Biol* **305**: 1057–1072.
99. Auffinger P, Westhof E. (2000) Water and ion binding around RNA and DNA (C,G)-oligomers. *J Mol Biol* **300**: 1113–1131.
100. Auffinger P, Westhof E. (2001) Hydrophobic groups stabilize the hydration shell of 2′-O-methylated RNA duplexes. *Angew Chem Int Ed* **40**: 4648–4650.
101. Auffinger P, Westhof E. (2002) Melting of the solvent structure around a RNA duplex: a molecular dynamics simulation study. *Biophys Chem* **95**: 203–210.
102. Kulinska K, Kulinski T, Lyubartsev A, Laaksonen A, Adamiak RW. (2000) Spatial distribution functions as a tool in the analysis of ribonucleic acids hydration — molecular dynamics studies. *Comput Chem* **24**: 451–457.
103. Noy A, Perez A, Lankas F, Javier Luque F, Orozco M. (2004) Relative flexibility of DNA and RNA: a molecular dynamics study. *J Mol Biol* **343**: 627–638.
104. Varnai P, Canalia M, Leroy JL. (2004) Opening mechanism of G.T/U pairs in DNA and RNA duplexes: a combined study of imino proton exchange and molecular dynamics simulation. *J Am Chem Soc* **126**: 14659–14667.
105. Schneider C, Brandl M, Sühnel J. (2001) Molecular dynamics simulation reveals conformational switching of water-mediated uracil-cytosine base pairs in an RNA duplex. *J Mol Biol* **305**: 659–667.
106. Sherer EC, Cramer CJ. (2002) Internal loop-helix coupling in the dynamics of the RNA duplex (GC*C*AGUUCGCUGGC)$_2$. *J Phys Chem B* **106**: 5075–5085.
107. Zacharias M, Engels JW. (2004) Influence of a fluorobenzene nucleobase analogue on the conformational flexibility of RNA studied by molecular dynamics simulations. *Nucl Acids Res* **32**: 6304–6311.
108. Pan Y, MacKerell AD. (2003) Altered structural fluctuations in duplex RNA versus RNA: a conformational switch involving base pair opening. *Nucl Acids Res* **31**: 7131–7140.

109. Pan Y, Priyakumar UD, MacKerell AD, Jr. (2005) Conformational determinants of tandem GU mismatches in RNA: insights from molecular dynamics simulations and quantum mechanical calculations. *Biochemistry* **44**: 1433–1443.
110. Piton N, Mu Y, Stock G, *et al.* (2007) Base-specific spin-labeling of RNA for structure determination. *Nucl Acids Res* **35**: 3128–3143.
111. Reblova K, Lankas F, Razga F, *et al.* (2006) Structure, dynamics, and elasticity of free 16S rRNA helix 44 studied by molecular dynamics simulations. *Biopolymers* **82**: 504–520.
112. Beckman RA, Moreland D, Louise-May S, Humblet C. (2006) RNA unrestrained molecular dynamics ensemble improves agreement with experimental NMR data compared to single static structure: a test case. *J Comput Aided Mol Des* **20**: 263–279.
113. Barthel A, Zacharias M. (2006) Conformational transitions in RNA single uridine and adenosine bulge structures: a molecular dynamics free energy simulation study. *Biophys J* **90**: 2450–2462.
114. Schneider C, Sühnel J. (2000) A molecular dynamics simulation study of coaxial stacking in RNA. *J Biomol Struct Dyn* **18**: 345–352.
115. Sorin EJ, Rhee YM, Pande VS. (2005) Does water play a structural role in the folding of small nucleic acids? *Biophys J* **88**: 2516–2524.
116. Li W, Ma B, Shapiro B. (2001) Molecular dynamics simulations of the denaturation and refolding of an RNA tetraloop. *J Biomol Struct Dyn* **19**: 381–396.
117. Sarzynska J, Nilsson L, Kulinski T. (2003) Effects of base substitutions in an RNA hairpin from molecular dynamics and free energy simulations. *Biophys J* **85**: 3445–3459.
118. Miller J, Kollman PA. (1997) Theoretical studies of an exceptionally stable RNA tetraloop: observation of convergence from an incorrect NMR structure to the correct one using unrestrained molecular dynamics. *J Mol Biol* **270**: 436–450.
119. Villa S, Stock G. (2006) What NMR relaxation can tell us about the internal motion of an RNA hairpin: a molecular dynamics simulation study. *J Chem Theory Comput* **2**: 1228–1236.
120. Koplin J, Mu Y, Richter C, Schwalbe H, Stock G. (2005) Structure and dynamics of an RNA tetraloop: a joint moleular dynamics and NMR study. *Structure* **13**: 1255–1267.
121. Joli F, Hantz E, Hartmann B. (2006) Structure and dynamics of phosphate linkages and sugars in an abasic hexaloop RNA hairpin. *Biophys J* **90**: 1480–1488.
122. Hermann T, Auffinger P, Scott WG, Westhof E. (1997) Evidence for a hydroxide ion bridging two magnesium ions at the active site of the hammerhead ribozyme. *Nucl Acids Res* **25**: 3421–3427.
123. Hermann T, Auffinger P, Westhof E. (1998) Molecular dynamics investigations of the hammerhead ribozyme RNA. *Eur J Biophys* **27**: 153–165.

124. Torres RA, Bruice TC. (1998) Molecular dynamics study displays near in-line attack conformations in the hammerhead ribozyme self-cleavage reaction. *Proc Natl Acad Sci USA* **95**: 11077–11082.
125. Torres RA, Bruice TC. (2000) The mechanism of phosphodiester hydrolysis – near in-line attack conformations in the hammerhead ribozyme. *J Am Chem Soc* **122**: 781–791.
126. Van Wynsberghe AW, Cui Q. (2005) Comparison of mode analyses at different resolutions applied to nucleic acid systems. *Biophys J* **89**: 2939–2949.
127. Lee TS, Silva-Lopez C, Martick M, Scott WG, York DM. (2007) Insight into the role of Mg^{2+} in hammerhead ribozyme catalysis from X-ray crystallography and molecular dynamics simulation. *J Chem Theory Comput* **3**: 325–327.
128. Krasovska MV, Sefcikova J, Reblova K, *et al.* (2006) Cations and hydration in catalytic RNA: molecular dynamics of the hepatitis delta virus ribozyme. *Biophys J* **91**: 626–638.
129. Sefcikova J, Krasovska MV, Sponer J, Walter NG. (2007) The genomic HDV ribozyme utilizes a previously unnoticed U-turn motif to accomplish fast site-specific catalysis. *Nucl Acids Res* **35**: 1933–1946.
130. Sefcikova J, Krasovska MV, Spackova N, Sponer J, Walter NG. (2007) Impact of an extruded nucleotide on cleavage activity and dynamic catalytic core conformation of the hepatitis delta virus ribozyme. *Biopolymers* **85**: 392–406.
131. Golebiowski J, Antonczak S, Di-Giorgio A, Condom R, Cabrol-Bass D. (2004) Molecular dynamics simulation of hepatitis C virus IRES IIId domain: structural behavior, electrostatic and energetic analysis. *J Mol Model* **10**: 60–68.
132. Rhodes MM, Reblova K, Sponer J, Walter NG. (2006) Trapped water molecules are essential to structural dynamics and function of a ribozyme. *Proc Natl Acad Sci USA* **103**: 13380–13385.
133. Pitici F, Beveridge DL, Baranger AM. (2002) Molecular dynamics simulation studies of induced fit and conformational capture in U1A-RNA binding: do molecular substates code for specificity? *Biopolymers* **65**: 424–435.
134. Reyes CM, Kollman PA. (1999) Molecular dynamics studies of U1A-RNA complexes. *RNA* **5**: 235–244.
135. Cojocaru V, Klement R, Jovin TM. (2005) Loss of G-A base pairs is insufficient for achieving a large opening of U4 snRNA K-turn motif. *Nucl Acids Res* **33**: 3435–3446.
136. Cojocaru V, Nottrott S, Klement R, Jovin TM. (2005) The snRNP 15.5K protein folds its cognate K-turn RNA: a combined theoretical and biochemical study. *RNA* **11**: 197–209.
137. Razga F, Koca J, Sponer J, Leontis NB. (2005) Hinge-like motions in RNA kink-turns: the role of the second a-minor motif and nominally unpaired bases. *Biophys J* **88**: 3466–3485.

138. Razga F, Zacharias M, Reblova K, Koca J, Sponer J. (2006) RNA kink-turns as molecular elbows: hydration, cation binding, and large-scale dynamics. *Structure* **14**: 825–835.
139. Hermann T, Westhof E. (1999) Simulations of the dynamics at an RNA-protein interface. *Nat Struct Biol* **6**: 540–544.
140. Showalter SA, Hall KB. (2005) Correlated motions in the U1 snRNA stem/loop 2: U1A RBD1 complex. *Biophys J* **89**: 2046–2058.
141. Kormos BL, Baranger AM, Beveridge DL. (2007) A study of collective atomic fluctuations and cooperativity in the U1A-RNA complex based on molecular dynamics simulations. *J Struct Biol* **157**: 500–513.
142. Law MJ, Linde ME, Chambers EJ, *et al.* (2006) The role of positively charged amino acids and electrostatic interactions in the complex of U1A protein and U1 hairpin II RNA. *Nucl Acids Res* **34**: 275–285.
143. Guo J, Gmeiner WH. (2001) Molecular dynamics simulation of the human U2B′ protein complex with U2 snRNA hairpin IV in aqueous solution. *Biophys J* **81**: 630–642.
144. Guo J, Daizadeh I, Gmeiner WH. (2000) Structure of the Sm binding site from human U4 snRNA derived from a 3 ns PME molecular dynamics simulation. *J Biomol Struct Dyn* **18**: 335–344.
145. Yingling YG, Shapiro BA. (2005) Dynamic behavior of the telomerase RNA hairpin structure and its relationship to dyskeratosis congenita. *J Mol Biol* **348**: 27–42.
146. Yingling YG, Shapiro BA. (2006) The prediction of the wild-type telomerase RNA pseudoknot structure and the pivotal role of the bulge in its formation. *J Mol Graph Model* **25**: 261–274.
147. Yamasaki S, Nakamura S, Terada T, Shimizu K. (2007) Mechanism of the difference in the binding affinity of *E. coli* tRNA(Gln) to glutaminyl-tRNA synthetase caused by noninterface nucleotides in variable loop. *Biophys J* **92**: 192–200.
148. Auffinger P, Westhof E. (1997) Rules governing the orientation of the 2′-hydroxyl group in RNA. *J Mol Biol* **274**: 54–63.
149. McCrate NE, Varner ME, Kim KI, Nagan MC. (2006) Molecular dynamics simulations of human tRNA Lys,3 UUU: the role of modified bases in mRNA recognition. *Nucl Acids Res* **34**: 5361–5368.
150. Nina M, Simonson T. (2002) Molecular dynamics of the tRNA(Ala) acceptor stem: comparison between continuum reaction field and particle-mesh Ewald electrostatic treatments. *J Phys Chem B* **106**: 3696–3705.
151. Nagan MC, Kerimo SS, Musierforsyth K, Cramer CJ. (1999) Wild-type tRNA microhelix(Ala) and 3:70 variants: molecular dynamics analysis of local helical structure, and tightly bound water. *J Am Chem Soc* **121**: 7310–7317.
152. Nagan MC, Beuning P, Musier-Forsyth K, Cramer CJ. (2000) Importance of discriminator base stacking interactions: molecular dynamics analysis of A73 microhelix$^{\text{Ala}}$ variants. *Nucl Acids Res* **28**: 2527–2534.

153. Beuning PJ, Nagan MC, Cramer CJ, *et al.* (2002) Efficient aminoacylation of the tRNA(Ala) acceptor stem: dependence on the 2:71 base pair. *RNA* **8**: 659–670.
154. Kallick DA, Nagan MC, Beuning PJ, *et al.* (2002) Discrimination of C1: G72 microhelix^{Ala} by AlaRS is based on specific atomic groups rather than conformational effects: an NMR and MD analysis. *J Phys Chem B* **106**: 8878–8884.
155. Sanbonmatsu KY, Tung CS. (2006) High performance computing in biology: multimillion atom simulations of nanoscale systems. *J Struct Biol* **157**: 470–480.
156. Li W, Ma B, Shapiro BA. (2003) Binding interactions between the core central domain of 16S rRNA and the ribosomal protein S15 determined by molecular dynamics simulations. *Nucl Acids Res* **31**: 629–638.
157. Spackova N, Sponer J. (2006) Molecular dynamics simulations of sarcin-ricin rRNA motif. *Nucl Acids Res* **34**: 697–708.
158. Mokdad A, Krasovska MV, Sponer J, Leontis NB. (2006) Structural and evolutionary classification of G/U wobble base pairs in the ribosome. *Nucl Acids Res* **34**: 1326–1341.
159. Razga F, Spackova N, Reblova K, *et al.* (2004) Ribosomal RNA kink-turn motif — a flexible molecular hinge. *J Biomol Struct Dyn* **22**: 183–194.
160. Reblova K, Spackova N, Stefl R, *et al.* (2003) Non-Watson-Crick base pairing and hydration in RNA motifs: molecular dynamics of 5S rRNA loop E. *Biophys J* **84**: 3564–3582.
161. Auffinger P, Bielecki L, Westhof E. (2003) The Mg^{2+} binding sites of the 5S rRNA loop E motif as investigated by molecular dynamics simulations. *Chem Biol* **10**: 551–561.
162. Auffinger P, Bielecki L, Westhof E. (2004) Symmetric K^+ and Mg^{2+} ion binding sites in the 5S rRNA loop E inferred from molecular dynamics simulations. *J Mol Biol* **335**: 555–571.
163. Reblova K, Spackova N, Koca J, Leontis NB, Sponer J. (2004) Long-residency hydration, cation binding, and dynamics of loop E/helix IV rRNA-L25 protein complex. *Biophys J* **87**: 3397–3412.
164. Crety T, Malliavin TE. (2007) The conformational landscape of the ribosomal protein S15 and its influence on the protein interaction with 16S RNA. *Biophys J* **92**: 2647–2665.
165. Li W, Sengupta J, Rath BK, Frank J. (2006) Functional conformations of the L11-ribosomal RNA complex revealed by correlative analysis of cryo-EM and molecular dynamics simulations. *RNA* **12**: 1240–1253.
166. Asensio JL, Hidalgo A, Cuesta I, *et al.* (2002) Experimental evidence for the existence of non-exo-anomeric conformations in branched oligosaccharides: NMR analysis of the structure and dynamics of aminoglycosides of the neomycin family. *Chemistry* **8**: 5228–5240.

167. Murray JB, Merouch SO, Russell RJ, Lentzen G, Haddad J, Mobashery S. (2006) Interactions of designer antibiotics and the bacterial ribosomal aminoacyl-tRNA site. *Chem Biol* **13**: 129–138.
168. Merouch SO, Mobashery S. (2007) Conformational transition in the aminoacyl t-RNA site of the bacterial ribosome both in the presence and absence of an aminoglycoside antibiotic. *Chem Biol Drug Des* **69**: 291–297.
169. Reblova K, Spackova N, Sponer JE, Koca J, Sponer J. (2003) Molecular dynamics simulations of RNA kissing-loop motifs reveal structural dynamics and formation of cation-binding pockets. *Nucl Acids Res* **31**: 6942–6952.
170. Reyes CM, Nifosi R, Frankel AD, Kollman PA. (2001) Molecular dynamics and binding specificity analysis of the bovine immunodeficiency virus biv tat-tar complex. *Biophys J* **80**: 2833–2842.
171. Kieken F, Arnoult E, Barbault F, *et al.* (2002) HIV-1(Lai) genomic RNA: combined use of NMR and molecular dynamics simulation for studying the structure and internal dynamics of a mutated SL1 hairpin. *Eur Biophys J* **31**: 521–531.
172. La Penna G, Genest D, Perico A. (2003) Modeling the dynamics of the solvated SL1 domain of HIV-1 genomic RNA. *Biopolymers* **69**: 1–14.
173. Aci S, Mazier S, Genest D. (2005) Conformational pathway for the kissing complex — extended dimer transition of the SL1 stem-loop from genomic HIV-1 RNA as monitored by targeted molecular dynamics techniques. *J Mol Biol* **351**: 520–530.
174. Mazier S, Genest D. (2007) Molecular dynamics simulation for probing the flexibility of the 35 nucleotide SL1 sequence kissing complex from HIV-1Lai genomic RNA. *J Biomol Struct Dyn* **24**: 471–479.
175. Pattabiraman N, Martinez HM, Shapiro BA. (2002) Molecular modeling and dynamics studies of HIV-1 kissing loop structures. *J Biomol Struct Dyn* **20**: 397–412.
176. Aci S, Gangneux L, Paoletti J, Genest D. (2004) On the stability of different experimental dimeric structures of the SL1 sequence from the genomic RNA of HIV-1 in solution: a molecular dynamics simulation and electrophoresis study. *Biopolymers* **74**: 177–188.
177. Kulinski T, Olejniczak M, Huthoff H, *et al.* (2003) The apical loop of the HIV-1 TAR RNA hairpin is stabilized by a cross-loop base pair. *J Biol Chem* **278**: 38892–38901.
178. Musselman C, Pitt SW, Gulati K, *et al.* (2006) Impact of static and dynamic A-form heterogeneity on the determination of RNA global structural dynamics using NMR residual dipolar couplings. *J Biomol NMR* **36**: 235–249.
179. Golebiowski J, Antonczak S, Fernandez-Carmona J, Condom R, Cabrol-Bass D. (2004) Closing loop base pairs in RNA loop-loop complexes: structural behavior, interaction energy and solvation analysis through molecular dynamics simulations. *J Mol Model* **10**: 408–417.

180. Mu Y, Stock G. (2006) Conformational dynamics of RNA-peptide binding: a molecular dynamics simulation study. *Biophys J* **90**: 391–399.
181. Beaurain F, Di Primo C, Toulme JJ, Laguerre M. (2003) Molecular dynamics reveals the stabilizing role of loop closing residues in kissing interactions: comparison between TAR-TAR* and TAR-aptamer. *Nucl Acids Res* **31**: 4275–4284.
182. Schneider C, Sühnel J. (1999) A molecular-dynamics simulation of the flavin mononucleotide-RNA aptamer complex. *Biopolymers* **50**: 287–302.
183. Yingling YG, Shapiro BA. (2007) Computational design of an RNA hexagonal nanoring and an RNA nanotube. *Nano Lett* **7**: 2328–2334.
184. Castrignano T, Chillemi G, Varani G, Desideri A. (2002) Molecular dynamics simulation of the RNA complex of a double-stranded RNA-binding domain reveals dynamic features of the intermolecular interface and its hydration. *Biophys J* **83**: 3542–3552.

This page intentionally left blank

Chapter 14

Computational Protein Design

J. G. Saven*

14.1 Introduction

Proteins have been evolutionarily engineered to perform a variety of functions that involve biomolecular structure, catalysis, and recognition. Hence, obtaining a quantitative, predictive molecular understanding of the structure and function of proteins is central to understanding the molecular basis of many life processes. Because the function of a protein is often closely tied to its structure, protein structure determination is critical for understanding the interplay between protein sequence, structure, and function. While determining the sequence of a protein is straightforward, accurately predicting the three-dimensional structure of a protein based on its sequence alone remains a challenging task. Although great advances are being made in this field of structure prediction, limitations with regard to the ability to reliably predict the three-dimensional structure of a protein suggest other types of studies involving sequence and structure may be used to further investigate (and engineer) proteins.

While structure prediction attempts to determine a three-dimensional structure from a protein's sequence of amino acids, protein design involves the identification of sequences having a target folded structure and desired molecular properties. Efforts addressing

*University of Pennsylvania, Department of Chemistry, 231 South 34th Street, Philadelphia, PA 19104-6323. Email: saven@sas.upenn.edu.

this “inverse protein folding” problem[1] can test our understanding of the key features of well-folded proteins. Predictions from protein design are best verified experimentally, by synthesizing the proposed sequences and examining their structural and biophysical properties. *De novo* designed proteins have potential applications as novel therapeutics, catalysts, biomaterials, and molecular scaffolds. Moreover, protein design tests our understanding of the determinants of well-folded structures and provides insight on structure-function relationships, since the biological functions of proteins are usually contingent on their forming unique, well-defined three-dimensional structures.

Proteins are complex, however, and designing sequences can be nontrivial. These polymeric macromolecules have large numbers of backbone and side-chain degrees-of-freedom. The stabilizing interactions that guide a protein to its native state are largely noncovalent, and such interactions can be difficult to quantify accurately. Although proteins may be imperfect or partially “frustrated” with regard to interactions between residues,[2] the interiors of proteins are typically well-packed with a large degree of shape and chemical complementarity among the residues. Moreover, in protein design, the identities of the amino acids can vary, leading to exponentially large numbers of possible sequences: if 20 amino acids are available at each variable site in a protein comprising 100 amino acids, more than 10^{130} sequences are possible. To surmount some of these difficulties, computational methods are being developed that open new frontiers in the design and study of proteins and other molecular systems. Most design efforts select or create a target backbone structure and then use energetic criteria to identify individual sequences or the properties of an ensemble of sequences consistent with the target.

Despite the complexities involved in identifying sequences, protein design is somewhat simpler than protein structure prediction. In part, this is due to the fact that there are often multiple solutions, and only one need be found in a successful design effort. In nature, there are many examples of two or more proteins with essentially the same structure and function having very different sequences. The presence of multiple sequences consistent with a particular structure can complicate the use of protein design to understand sequence variability, since sequences folding to very similar structures may be broadly distributed in sequence space. Characterizing the ensemble of viable solutions may require extensive sampling or the development of methods for characterizing such ensembles statistically. In addition, protein structure prediction is difficult due to the covalent connectivity of the peptide backbone, which maintains a specific amino acid sequence. As the protein folds, distant residues in the sequence may have stabilizing interactions when they are nearby in space, e.g. due to the formation of hydrogen bonds or hydrophobic contacts. Since the sequence is fixed, residues close in sequence are also close in space, leading to possible situations where not all noncovalent interactions may be simultaneously satisfied — an effect referred to as "frustration".[3] In protein design, however, such "frustration" may be alleviated by changing the sequence of the amino acids, while in protein structure prediction, alternate conformations of the backbone must be sampled in the search for lower energy structures.

While this chapter focuses on computational approaches to protein design, such methods are not always necessary. From the properties observed in experimentally determined protein structures, structural motifs have been identified that are common to many proteins. Such motifs may be assembled to form whole proteins or protein complexes. This hierarchical protein design[4,5] has been successful in designing proteins such as helix-bundles and coiled-coils.[6,7] On the other hand, partially random protein libraries with diversities greater than 10^5 may be generated, from which variants with desired characteristics may be selected.[8,9] Catalytic antibodies and phage display demonstrate the power and versatility of combinatorial approaches to protein engineering,[10–12] which are appropriate for cases where we have incomplete knowledge about the determinants of structure and/or function. Despite their notable successes, such "non-computational" approaches to protein design can become problematic as protein structures become complex and asymmetric, and as the number of variable positions increases. In addition, such methods often yield proteins that do not have the well-defined tertiary structures of natural proteins.[4]

Herein, common elements of computational protein design are discussed. Computational techniques for identifying and characterizing the properties of sequences compatible with a particular structure are presented, as are examples of the application of such methods to the design of particular proteins that have been subsequently studied experimentally.

14.2 Methodology of Computational Protein Design

Computational protein design involves several fundamental elements.[13] These elements include information necessary prior to quantitative design, such as how structural features and inter-residue interactions are parameterized.

Target structure. Many design efforts start with a high-resolution structure obtained through X-ray crystallography or solution NMR, resulting in the redesign of a known protein structure. The reuse of an existing structure need not limit functional diversity, since different functionalities may be obtained using the same protein fold.[14] Tertiary structures may also be computationally modeled so as to obtain novel structures and topologies.[15,16] The target structure need not be rigid and can fluctuate about a desired overall fold topology.[17,18]

Degrees of freedom. In protein design, degrees of freedom associated with the residues are varied in the search for sequences consistent with a particular target structure. Two types of degrees of freedom are often simultaneously varied: the amino acid identities, which specify the sequence, and their side-chain conformations. Not all amino acids are required to create functional proteins,[19] and the use of prepatterning or a reduced alphabet can vastly simplify design by reducing the number of such degrees of freedom. Statistical analyses of high resolution structures have yielded discrete sets of side-chain conformations, rotamers, that are preferentially occupied.[20] Such rotamer approximations reduce the side-chain degrees of freedom available to each amino acid, and thus, facilitate design calculations.

Energy function. The compatibility between sequence and structure is evaluated using effective energy functions, which represent the physico-chemical interactions present in the folded structure. Often, atomistic potentials are used that are also applied to molecular simulations of proteins (e.g. Amber,[21] CHARMM,[22] Gromos[23]). Most potentials have terms involving bond lengths, bond angle, and dihedral angles, as well as terms accounting for the van der Waals, electrostatic, and hydrogen bonding interactions. Often, only the noncovalent terms are explicitly evaluated, since bond lengths, bond angles, and dihedral angles do not vary appreciably and are determined by the backbone structure and allowed rotamers. Simplified (coarse-grain) database-derived potentials that address structural propensities and do not include atomistic detail may also be implemented in sequence design.[24] The quantity that specifies consistency between structure and sequence is often the energy of a particular sequence-rotamer configuration for the template structure. For models and systems where alternate structures may have energies comparable to that of the template, these criteria may be extended to include explicitly the energetic separation of the template structure from other competing structures.[24–26]

Solvation and patterning. Hydrophobic effects help stabilize the compact structures of folded proteins,[27] and it is important to include terms that account for these effects and quantify the solvation propensities of the amino acids. Such solvation effects are often modeled with an effective free energy term that quantifies the hydrophobicity of the side-chain.[28] The "microphase separation" typically observed in proteins may also be realized via hydrophobic patterning[8] to ensure that nonpolar residues are buried in the interior while polar residues are exposed to the solvent. The patterning of the sequence may also be used to preferentially place amino acids that are consistent with secondary structures present in the template structure, i.e. amino acids having preferences for α-helix[29–31] and β-sheet structures.[32–34]

Search methods. Once a template structure and degree of sequence and structure variability have been specified, it remains necessary to search or characterize the space of possible sequences that fulfill energetic "foldability" criteria. Sequence search methods based on Monte Carlo algorithms have been widely used.[35–37] These algorithms may be made more efficient through the use of appropriate biasing in the variation of sequence and side-chain conformation.[38–40] Simulated annealing, a variant of Monte Carlo methods, may be used to identify low energy states,[41] though in some cases these may not be global optima.[42] Other stochastic methods, such as genetic algorithms, may also be useful for identifying low energy configurations on a rugged fitness landscape, such as that encountered during protein design.[17,18,43,44] Pruning and elimination methods such as dead end elimination (DEE) identify the global minimum energy sequence for pair-wise potentials.[45] The algorithm systematically discards local sequence-rotamer states that cannot be part of the global minimum. This leads to a narrowing of the search space during the computation. While significant challenges remain with regard to computation time as protein size and diversity increases, the method has been instrumental in many design projects.

Rather than identify particular sequences, statistical methods directly estimate the site-specific amino acid probabilities for sequences folding to a target structure.[25,46] Modeled on the concept of entropy maximization in statistical mechanics, the algorithm defines an effective entropy as a function of the individual amino acid probabilities. Maximization subject to desired energetic and functional constraints yields the site-specific probabilities of the amino acids. The sequence space can be characterized using this method, and the probabilistic approach to protein design can easily address systems that may be too large for direct sequence sampling. Furthermore, the method is versatile enough to identify optimal and sub-optimal sequences, and provides information that may be readily used to guide the construction of combinatorial experiments.[25,46] For sufficiently small degrees of sequence diversity, the sequence space may be sampled using efficient Monte Carlo methods to yield similar information concerning the likelihood of the amino acids at variable residue positions.[40]

14.3 Computationally Designed Proteins

Predictive tools for protein design have a variety of applications, which include the development of new biotechnological therapeutics, materials, and nanoscale systems as well as studies of protein structure, function, and stability. As proteins are complex and many factors contribute to their stability and folding, sequence design efforts often face tough challenges. Computational protein design methods have been created to tackle some of these challenges, including the large numbers of degrees of freedom and simultaneous consideration of a large number of inter-residue interactions. Broadly, design efforts have emphasized: (i) the redesign of existing proteins so as to impart novel properties or explore the effects of sequence variation, and (ii) the design of well-structured folded states using large-scale sequence variability (*de novo* design).

14.3.1 *Protein Re-engineering*

Many natural proteins have been redesigned so as to modulate function and stability and to introduce new catalytic activity. Such studies often make use of high-resolution structures and knowledge of the mechanisms of a protein's particular activity. Computationally-guided mutation has been used to stabilize yeast cytosine deaminase.[47] Three out of five computationally-designed mutants of chorismate mutase were found to have enzymatic activity comparable to wild type, with one exhibiting greater activity and efficiency.[48] Designed mutants of procarboxypeptidase have been characterized in terms of both their structures and their thermostabilities.[49] The homing endonuclease I-MsoI has been redesigned so as to bind and cleave at a novel DNA recognition site, and high-resolution structures confirm binding via the targeted protein-DNA interface.[50] The constellation of residues designed to confer triose phosphate isomerase activity to a bacterial ribose binding protein[51] may be transferred to homologous proteins with little loss in activity, illustrating the robustness and transferability of such centers.[52] Such ribose binding proteins have also been computationally designed to yield sensitive biosensors.[53]

Computational methods for the design of enzymes continue to be refined.[54,55]

Protein-protein interfaces have also recently been subjects of computational protein design. A novel hydrogen bond network has been engineered into the interface between DNase and an immunity protein to yield specific recognition between the cognate partners of the complex.[56] An important adhesion protein interaction involved in inflammatory response has been stabilized, which involves the interface between integrin lymphocyte function-associated antigen-1 (LFA-1) and its ligand intercellular adhesion molecule-1 (ICAM-1).[57] A calcium-binding site has been designed into the cell adhesion protein CD2, while retaining the ability to bind CD48. Such systems can be useful for studying the engineering of Ca^{2+} responsive sensors, switches, and signaling mechanisms.[58] Structure-based computational design has identified tetrapeptides that efficiently depolymerize serine-protease inhibitor (serpin) aggregates often associated with cirrhosis and emphysema.[59]

Miniproteins provide useful model systems for investigating protein folding and design. Their small size and ultrafast folding kinetics also facilitate their study in the context of folding theory and molecular dynamics simulations. Thus, these small molecules are excellent systems for both experimental and theoretical studies. The "speed limit" for folding has be estimated to be on the order of one microsecond.[60] Computationally-designed mutants of the 47-residue GA module of an albumin binding domain and the 20-residue Trp-cage miniprotein, Trp2-cage, fold on this time scale.[61,62] These studies illustrate how design may be used to explore the relationships between structure, sequence, stability, kinetics, and folding.

Complex quaternary structures have also been subjects of design. The DNA protection protein, Dps, is a ferritin-like dodecamer. This high symmetry protein complex has been redesigned by designing up to 120 mutations per protein so as to present a large hydrophobic interior surface (Fig. 14.1). The resulting proteins fold and the iron mineralization rates are comparable to that of the wild type. These studies illustrate the versatility of some ferritin scaffolds for engineering proteins containing large cavities (4.5 nm or more in diameter),

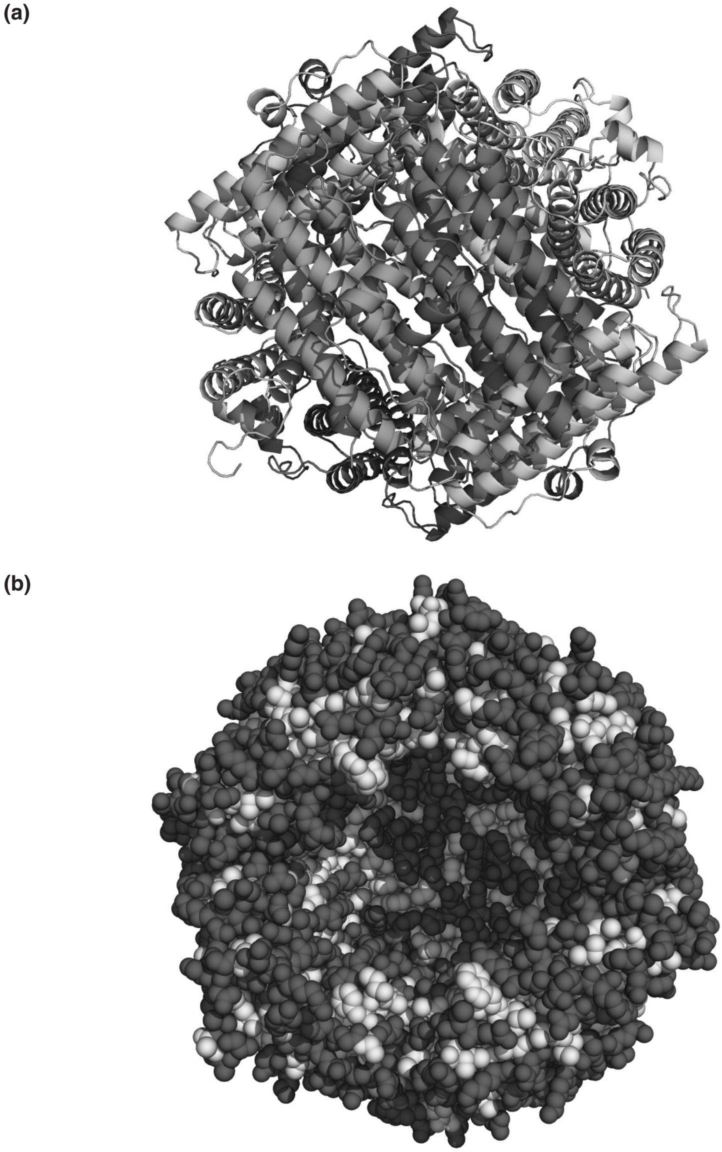


Fig. 14.1 Redesign of Dps to obtain hydrophobic interior nano-cavity.[63] **(a)** Wild type Dps dodecamer. **(b)** Wild type Dps, with two subunits deleted to expose interior cavity. Hydrophobic residues (A, V, I, L, F, M, W) are rendered in lighter shade of gray. **(c)** Model of protein with 120 computationally redesigned interior hydrophobic surface.

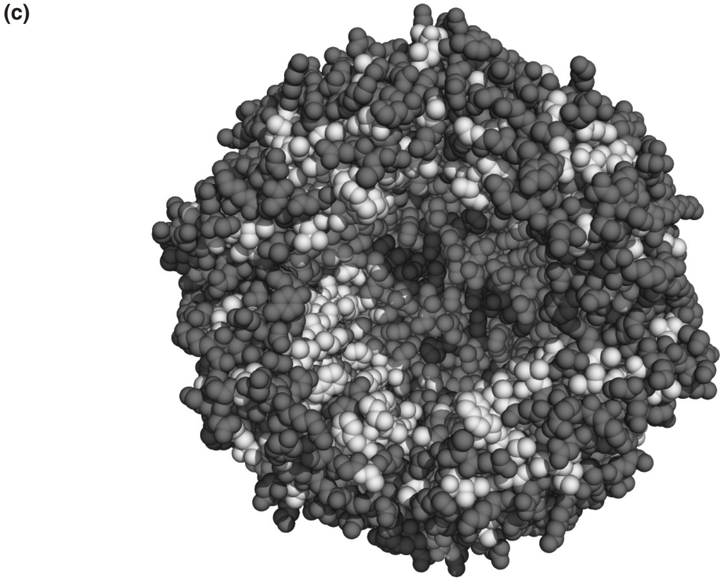


Fig. 14.1 (*Continued*)

proteins which have potential applications as nanoscale containers, and "reaction vessels" for hydrophobic solutes.[63]

Water-soluble analogues of the membrane-bound potassium channel KcsA have been computationally designed.[64] A version of the tetrameric membrane-bound protein with an engineered toxin-binding site was subjected to computational analysis (Fig. 14.2). Exposed, transmembrane hydrophobic residues on the exterior of the protein were targeted for mutation. In addition to considering inter-atomic interactions, the value of a database-derived solvation energy ("environmental energy")[46] was constrained to be that of a soluble protein with the same number of residues as KcsA. For the tetrameric KcsA complex, 140 exposed residues were initially targeted for variation. A computationally redesigned variant, WSK-3, expresses in high yield and shares structural and functional signature properties with its membrane soluble counterpart: it is predominantly tetrameric in solution; it binds the toxin specifically with the same affinity and stoichiometry as wild type KcsA; and the toxin binding may be inhibited

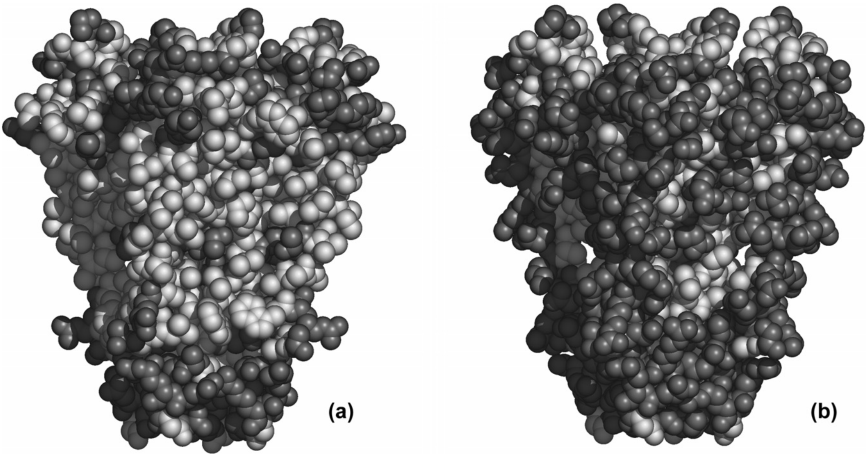


Fig. 14.2 Computationally designed water soluble variant of the integral membrane protein KcsA.[66] **(a)** Structure of wild type KcsA. **(b)** Computationally designed water soluble KcsA variant, WSK-3. Hydrophobic and aliphatic residues (A, G, L, I, V, W, M, F) are colored in lighter shade of gray.

with a small molecule blocker, tetraethyl ammonium. Simulations of the solubilized bacterial potassium ion channel[65] are consistent with experimental studies suggesting that the designed variant maintains its membrane-bound structure and binding properties in aqueous solution.[66] This study reveals how protein design may be used to facilitate characterization of the structure and functional properties of membrane proteins, which are notoriously difficult to work with due to their low expression levels and poor solubilities. In related work, sequence search algorithms have been used to identify solubilized variants of the integral membrane protein phospholamban.[67]

14.3.2 *De novo Designed Proteins*

De novo design often refers to the complete design of a novel sequence and can also include the design of new structures. One of the first examples of a computationally *de novo* designed protein was a $\beta\alpha\beta$-motif resembling a zinc finger DNA binding module. The designed protein folded stably without the Zn^{2+} metal ion of the parent structure.[68] A 73-residue three-helix bundle protein has also been designed with the aid of computational methods to identify core residues.[69]

Structure may also be included as part of the design process. A 97-residue α/β protein was designed to have a nonnatural fold and its structure determined.[16] Other designed nonnatural protein structures include a right-handed helical coiled-coil.[70] The structures of helix bundle motifs may be parameterized using a few global variables that describe the global structure and superhelical coiling of the protein.[71,72] Such methods have been used to arrive at model di-iron and di-manganese proteins, which provide useful platforms for engineering and investigating the versatile range of binding and catalytic properties exhibited by this metalloprotein motif.[73–76] A protein has been designed that can switch from a zinc finger-like fold to a trimeric coiled-coil in response to changes in pH or transition metal ion concentration.[77] Heterotetrameric structures have been designed that have mixed alpha-beta secondary structures, where the individual subunits are 21-residue miniproteins. High resolution crystal structures are consistent with the target structures, and these small proteins may be useful for exploring protein-protein interactions.[78]

A 114-residue four-helix bundle (DFsc) with a dinuclear metal center[15] has been computationally designed (Fig. 14.3). Two di-iron proteins, a heterotetrameric protein,[79] and a helix-loop-helix DF1 protein,[80] had been previously designed and characterized.[81] The target template for the monomeric variant DFsc was generated by designing a single chain that properly positioned the four helices. After constraining the identities of residues that confer metal binding and substrate accessibility to the active site, computational methods were used to determine the identities of the remaining 88 residues. Despite the presence of six ionizable residues within the interior, the designed apo protein folds in the absence of metal ions. The protein stoichiometrically binds two equivalents of Fe(III), Co(II), Mn(II), and Zn(II), and has increased thermal stability upon metal binding.

While metal binding sites have been engineered into proteins,[82,83] the design of functional metalloproteins containing beta structure

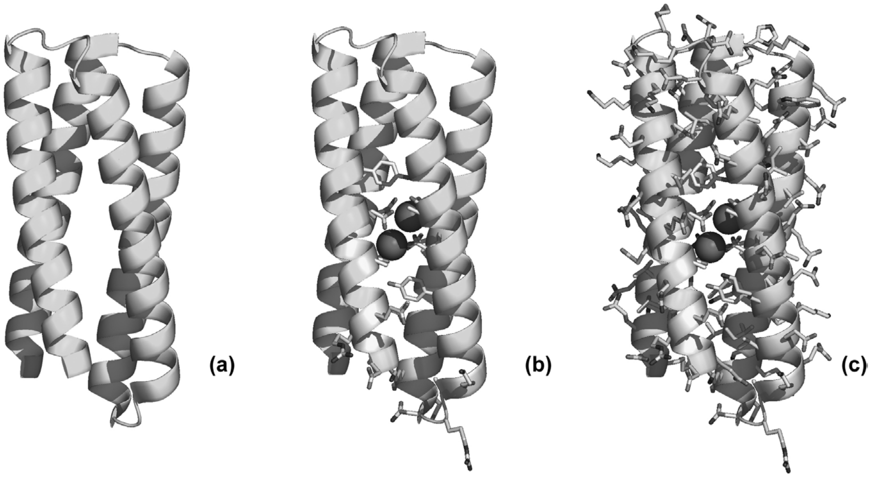


Fig. 14.3 Computational design of 114-residue dinuclear metalloprotein, DFsc.[15] **(a)** Computationally designed four-helix bundle suitable for dinuclear metal coordination. **(b)** Keystone residues (explicit side-chains) comprise primary and secondary ligands for the metal ions, a helix initiation sequence, a suitable turn sequence, and small side-chain alanine residues that confer accessibility to the active site. **(c)** Computational design is used to determine the identities of the remaining 88 amino acids.

is less well-developed than that of helical proteins, but such beta conformations occur frequently in natural metal-binding proteins. A structure and sequence has been designed for a beta protein whose metal binding site mimics that of rubredoxin and recovers the binding of Fe(II/III). The resulting 40-residue protein folds into a beta-structure, in the presence and absence of metal ions and binds Fe(II/III) to form a redox-active site that is stable to more than 16 cycles of oxidation and reduction, even in an aerobic environment.[84]

In addition to the study of natural proteins, computational *de novo* design facilitates the construction of novel biomaterials and electrochemical devices, which may not have analogs seen in nature. The use of non-biological cofactors may potentially be used to create proteins with new properties not accessible with naturally occurring amino acids or biological cofactors. Computational design has been applied to arrive at a protein framework that encapsulates synthetic porphyrin cofactors. Selective cofactor recognition is a hallmark of natural systems and a significant challenge. Previously redesigned heme proteins have bound synthetic metalloporphyrins with relatively low specificity. A native-like protein has been computationally designed that selectively binds a non-biological cofactor, diphenyl porphyrin (DPP-Fe) (Fig. 14.4). The protein encapsulates a pair of the DPP-Fe units through biaxial histidine coordination to Fe, resulting in a 34-residue peptide assembled to form a tetrahelical bundle. Binding of the cofactor in a bis-His fashion was observed for DPP but not for other Fe-containing porphyrins.[85] More recently, this approach has been extended to the design of a 108-residue protein that binds the same cofactor but in a lower symmetry environment.[86]

The folding of a protein to a unique, stable structure requires a sufficiently large free energy difference between native and nonnative states. In order for the folded state to be unique and thermodynamically stable for a particular sequence, the structure must be separated energetically from competing structures. Designing against such competing structures explicitly in the design process is referred to as negative design. Similar concepts form the basis of the free energy landscape theory of protein folding, which postulates that naturally occurring proteins have a smooth funnel-like energy landscape to guide the folding of a protein to a well-defined minimum energy conformation.[2] In arriving at a completely redesigned helical protein, energy landscape theory has been applied to the design of a three-helix bundle.[24] Using simplified representation of the side-chains, an ensemble of denatured decoy states was generated from folding simulations. Optimization of weighted Z-scores of candidate sequences were used to guide sequence design, where the Z-score is an energy difference between the target and denatured states relative to the size of the energy fluctuations among denatured states. NMR and circular dichroism studies of one of the designed sequences were consistent with both the expected α-helix content and a well-defined three-dimensional structure. More commonly, atomistic energy functions are used to optimize the energy of a sequence in a target structure relative to its energies in competing misfolded conformations. Incorporating negative design is important when degnerate structures

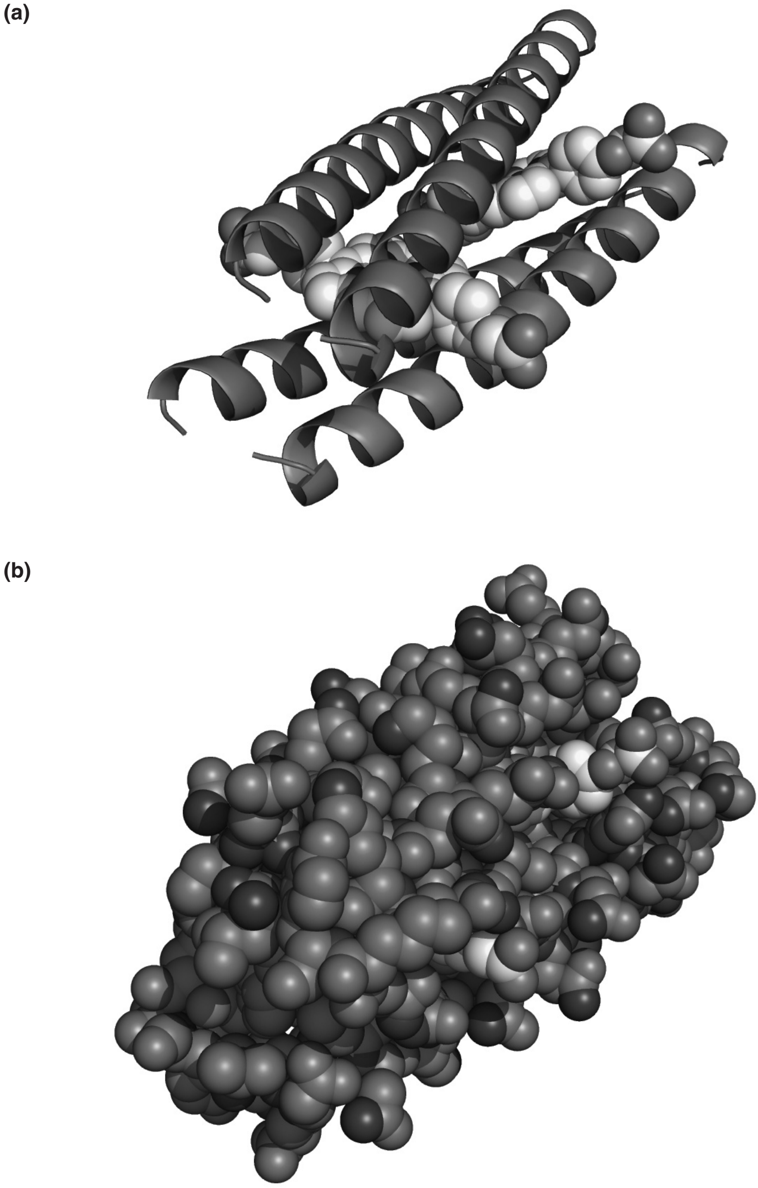


Fig. 14.4 Computational *de novo* design of four-helix bundle containing the non-biological cofactor, iron diphenyl porphyrin (DPP-Fe).[85] **(a)** Computationally designed structure containing two DPP-Fe cofactors. **(b)** Computationally designed sequence and structure of complex.

for a given sequence are likely, e.g. in designing protein-protein interfaces, where the subtlety of specific interactions and the presence of multiple low energy configurations present a significant challenge, or when simplified representations of the amino acids are used. In designing an A_2B_2 helical heterotetrameric protein with a dinuclear metal center, negative design was used successfully to select for sequences with charge patterned exterior positions such that the peptides do not form non-target homotetramers or heterotetramers.[79] Havranek and Harbury describe an algorithm that uses explicit negative design to engineer coiled-coil interfaces that favor the formation of either homodimers or heterodimers.[26]

Membrane proteins are vital to a variety of cellular processes and also are the targets of a large number of drugs and therapeutics. Only recently, however, have they been examined in the context structure-based design and engineering. Studies of a model helical protein having both transmembrane and aqueous domains have highlighted the roles of Asn-mediated interactions in both domains to conferring a particular fold and oligomerization state.[87] The role of cooperative, interhelix interactions in a serine-zipper transmembrane helix motif has been examined in computationally designed systems. The designed protein exhibits parallel dimerization of the helices, but mutation of the central serine residues to alanine yields dimers of comparable stability, suggesting that complementrary packing interactions rather than hydrogen bonding play a dominant role in stabilizing the dimmer.[88] Building upon these findings, peptides have been computationally designed that target transmembrane helices in a sequence specific manner. The designed peptides are able to discriminate transmembrane helices of two closely related integrins, where the specificity is determined largely through complementary peptide-helix steric interactions.[89]

14.4 Future Outlook

Recent achievements of computational protein design are striking. The structures and sequences of proteins having more than 100 variable residues are now consistently being realized. The experimentally determined structures of many designed proteins agree well with those that are targeted. Increasingly, desired properties and functionalities are being introduced into proteins with the aid of computational methods. Genetic methods and mutagenesis are widely used techniques for probing protein structure and function, and computational protein design can make such methods more informative and efficient. The design of proteins that mimic the efficiency, selectivity, and regulation of natural proteins is likely to be difficult. A synthesis of computational design and library-based methods is likely to yield important advances in both the identification of novel proteins and in furthering our understanding of proteins and their biological functions. Lastly, incorporating abiotic components, e.g. non-biological amino acids, cofactors, and monomers, will yield protein systems poised to present new functions that have not previously been accessible — functions that go beyond those observed in natural systems.

Acknowledgments

The author gratefully acknowledges support from the National Institutes of Health (GM61267, GM71628), the US Department of Energy (DE-FG02-04ER46156), the University of Pennsylvania's Nano/Bio Interface Center through the National Science Foundation (NSF) NSEC DMR-0425780 and Laboratory for Research on the Structure of Matter through NSF MRSEC DMR05-20020, and the NSF (NIRT BES 0210777). The figures were rendered using PyMol (DeLano Scientific LLC).

References

1. Pabo C. (1983) Molecular technology. Designing proteins and peptides. *Nature* **301**: 200.
2. Onuchic JN, Luthey-Schulten Z, Wolynes PG. (1997) Theory of protein folding: the energy landscape perspective. *Ann Rev Phys Chem* **48**: 545–600.
3. Bryngelson JD, Wolynes PG. (1987) Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA* **84**: 7524–7528.

4. Bryson, JW, Betz SF, Lu HS, *et al.* (1995) Protein design: a hierarchic approach. *Science* **270**: 935–941.
5. DeGrado WF, Summa CM, Pavone V, *et al.* (1999) *De novo* design and structural characterization of proteins and metalloproteins. *Ann Rev Biochem* **68**: 779–819.
6. Regan L, DeGrado WF. (1988) Characterization of a helical protein designed from first principles. *Science* **241**: 976–978.
7. Harbury PB, Zhang T, Kim PS, Alber T. (1993) A switch between two-, three-, and four-stranded coiled coils in gcn4 leucine zipper mutants. *Science* **262**: 1401–1407.
8. Kamtekar S, Schiffer JM, Xiong HY, Babik JM, Hecht MH. (1993) Protein design by binary patterning of polar and nonpolar amino-acids. *Science* **262**: 1680–1685.
9. MacBeath G, Kast P, Hilvert D. (1998) Redesigning enzyme topology by directed evolution. *Science* **279**: 1958–1961.
10. Pollack SJ, Jacobs JW, Schultz PG. (1986) Selective chemical catalysis by an antibody. *Science* **234**: 1570–1573.
11. Hoess RH. (2001) Protein design and phage display. *Chem Rev* **101**: 3205–3218.
12. Weiss GA, Watanabe CK, Zhong A, Goddard A, Sidhu SS. (2000) Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc Natl Acad Sci USA* **97**: 8950–8954.
13. Saven JG. (2001) Designing protein energy landscapes. *Chem Rev* **101**: 3113–3130.
14. Nagano N, Orengo CA, Thornton JM. (2002) One fold with many functions: the evolutionary relationships between tim barrel families based on their sequences, structures, and functions. *J Mol Biol* **321**: 741–765.
15. Calhoun JR, Kono H, Lahr S, *et al.* (2003) Computational design and characterization of a monomeric helical dinuclear metalloprotein. *J Mol Biol* **334**: 1101–1115.
16. Kuhlman B, Dantas G, Ireton GC, *et al.* (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**: 1364–1368.
17. Desjarlais JR and TM Handel. (1999) Side-chain and backbone flexibility in protein core design. *J Mol Biol* **290**: 305–318.
18. Kraemer-Pecore, CM, Lecomte JT, Desjarlais JR. (2003) A *de novo* redesign of the ww domain. *Protein Sci* **12**: 2194–2205.
19. Akanuma S, Kigawa T, Yokoyama S. (2002) Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. *Proc Natl Acad Sci USA* **99**: 13549–13553.
20. Dunbrack RL. (2002) Rotamer libraries in the 21st century. *Curr Opin Struct Biol* **12**: 431–440.

21. Weiner SJ, Kollman PA, Case DA, *et al.* (1984) A new force-field for molecular mechanical simulation of nucleic-acids and proteins. *J Am Chem Soc* **106**: 765–784.
22. Brooks BR, Bruccoleri RE, Olafson BD, *et al.* (1983) Charmm — a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* **4**: 187–217.
23. Hermans J, Berendsen HJC, Vangunsteren WF, Postma JPM. (1984) A consistent empirical potential for water-protein interactions. *Biopolymers* **23**: 1513–1518.
24. Jin W, Kambara O, Sasakawa H, Tamura A, Takada S. (2003) *De novo* design of foldable proteins with smooth folding funnel: automated negative design and experimental verification. *Struct Fold Des* **11**: 581–590.
25. Zou JM, Saven JG. (2000) Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. *J Mol Biol* **296**: 281–294.
26. Havranek JJ, Harbury PB. (2003) Automated design of specificity in molecular recognition. *Nat Struct Biol* **10**: 45–52.
27. Fersht AR. (1999) *Structure and Mechanism in Protein Science*. Freeman, New York.
28. Eisenberg D, McLachlan A. (1986) Solvation energy in protein folding and binding. *Nature* **319**: 199–203.
29. O'Neil KT, Degrado WF. (1990) A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* **250**: 646–651.
30. Chakrabartty A, Kortemme T, Baldwin RL. (1994) Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci* **3**: 843–852.
31. Lyu PC, Liff MI, Marky LA, Kallenbach NR. (1990) Side-chain contributions to the stability of alpha-helical structure in peptides. *Science* **250**: 669–673.
32. Smith CK, Withka JM, Regan L. (1994) A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry* **33**: 5510–5517.
33. Kim CA, Berg JM. (1993) Thermodynamic beta-sheet propensities measured using a zinc-finger host peptide. *Nature* **362**: 267–270.
34. Minor DL, Kim PS. (1994) Measurement of the beta-sheet-forming propensities of amino acids. *Nature* **367**: 660–663.
35. Shakhnovich EI, Gutin AM. (1993) A new approach to the design of stable proteins. *Protein Eng* **6**: 793–800.
36. Hellinga H. (1999) Automated design of functional metal-binding sites in proteins. *FASEB J* **13**: A1430–A1430.
37. Kuhlman B, Baker D. (2000) Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA* **97**: 10383–10388.
38. Cootes AP, Curmi PMG, Torda AE. (2000) Biased Monte Carlo optimization of protein sequences. *J Chem Phys* **113**: 2489–2496.

39. Zou JM, Saven JG. (2003) Using self-consistent fields to bias Monte Carlo methods with applications to designing and sampling protein sequences. *J Chem Phys* **118**: 3843–3854.
40. Yang X, Saven JG. (2005) Computational methods for protein design and protein sequence variability: biased Monte Carlo and replica exchange. *Chem Phys Lett* **401**: 205–210.
41. Kirkpatrick S, Gelatt CD, Vecchi MP. (1983) Optimization by simulated annealing. *Science* **220**: 671–680.
42. Voigt CA, Gordon DB, Mayo SL. (2000) Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* **299**: 789–803.
43. Holland JH. (1975) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, pp. 183. University of Michigan Press.
44. Lazar GA, Desjarlais JR, Handel TM. (1997) *De novo* design of the hydrophobic core of ubiquitin. *Protein Sci* **6**: 1167–1178.
45. Desmet J, De Maeyer M, Hazes B, Lasters I. (1992) The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**: 539–542.
46. Kono H, Saven JG. (2001) Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J Mol Biol* **306**: 607–628.
47. Korkegian A, Black ME, Baker D, Stoddard BL. (2005) Computational thermostabilization of an enzyme. *Science* **308**: 857–860.
48. Lassila JK, Keeffe JR, Oelschlaeger P, Mayo SL. (2005) Computationally designed variants of *Escherichia coli* chorismate mutase show altered catalytic activity. *Protein Eng Des Sel* **18**: 161–163.
49. Dantas G, Corrent C, Reichow SL, *et al.* (2007) High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. *J Mol Biol* **366**: 1209–1221.
50. Ashworth J, Havranek JJ, Duarte CM, *et al.* (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**: 656–659.
51. Dwyer MA, Looger LL, Hellinga HW. (2004) Computational design of a biologically active enzyme. *Science* **304**: 1967–1971.
52. Allert M, Dwyer MA, Hellinga HW. (2007) Local encoding of computationally designed enzyme activity. *J Mol Biol* **366**: 945–953.
53. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. (2003) Computational design of receptor and sensor proteins with novel functions. *Nature* **423**: 185–190.
54. Zanghellini A, Jiang L, Wollacott AM, *et al.* (2006) New algorithms and an *in silico* benchmark for computational enzyme design. *Protein Sci* **15**: 2785–2794.

55. Lassila JK, Privett HK, Allen BD, Mayo SL. (2006) Combinatorial methods for small-molecule placement in computational enzyme design. *Proc Natl Acad Sci USA* **103**: 16710–16715.
56. Joachimiak LA, Kortemme T, Stoddard BL, Baker D. (2006) Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. *J Mol Biol* **361**: 195–208.
57. Song G, Lazar GA, Kortemme T, *et al.* (2006) Rational design of intercellular adhesion molecule-1 (icam-1) variants for antagonizing integrin lymphocyte function-associated antigen-1-dependent adhesion. *J Biol Chem* **281**: 5042–5049.
58. Yang W, Wilkins AL, Ye YM, *et al.* (2005) Design of a calcium-binding protein with desired structure in a cell adhesion molecule. *J Am Chem Soc* **127**: 2085–2093.
59. Chowdhury P, Wang W, Lavender S, *et al.* (2007) Fluorescence correlation spectroscopic study of serpin depolymerization by computationally designed peptides. *J Mol Biol*
60. Kubelka J, Hofrichter J, Eaton WA. (2004) The protein folding "speed limit". *Curr Opin Struct Biol* **14**: 76–88.
61. Zhu Y, Fu X, Wang T, *et al.* (2004) Guiding the search for a protein's maximum rate of folding. *Chem Phys* **307**: 99–109.
62. Bunagan MR, X Yang, JG Saven, F Gai. (2006) Ultrafast folding of a computationally designed trp-cage mutant: Trp(2)-cage. *J Phys Chem B* **110**: 3759–3763.
63. Swift J, Wehbi WA, Kelly BD, *et al.* (2006) Design of functional ferritin-like proteins with hydrophobic cavities. *J Am Chem Soc* **128**: 6611–6619.
64. Slovic AM, Kono H, Lear JD, Saven JG, DeGrado WF. (2004) Computational design of water-soluble analogues of the potassium channel KcsA. *Proc Natl Acad Sci USA* **101**: 1828–1833.
65. Bronson J, Lee OS, Saven JG. (2006) Molecular dynamics simulation of WSK-3, a computationally designed, water-soluble variant of the integral membrane protein KcsA. *Biophys J* **90**: 1156–1163.
66. Slovic AM, Kono H, Lear JD, Saven JG, DeGrado WF. (2004) Computational design of water-soluble analogues of the potassium channel KcsA. *Proc Natl Acad Sci USA* **101**: 1828–1833.
67. Slovic AM, Summa CM, Lear JD, DeGrado WF. (2003) Computational design of a water-soluble analog of phospholamban. *Protein Sci* **12**: 337–348.
68. Dahiyat BI, Mayo SL. (1997) *De novo* protein design: fully automated sequence selection. *Science* **278**: 82–87.
69. Walsh ST, Cheng H, Bryson JW, Roder H, DeGrado WF. (1999) Solution structure and dynamics of a *de novo* designed three-helix bundle protein. *Proc Natl Acad Sci USA* **96**: 5486–5491.

70. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. (1998) High-resolution protein design with backbone freedom. *Science* **282**: 1462–1467.
71. North B, Summa CM, Ghirlanda G, DeGrado WF. (2001) D-n-symmetrical tertiary templates for the design of tubular proteins. *J Mol Biol* **311**: 1081–1090.
72. Plecs JJ, Harbury PB, Kim PS, Alber T. (2004) Structural test of the parameterized-backbone method for protein design. *J Mol Biol* **342**: 289–297.
73. Maglio O, Nastri F, Calhoun JR, *et al.* (2005) Artificial diiron proteins: Solution characterization of four helix bundles containing two distinct types of inter-helical loops. *J Biol Inorg Chem* **10**: 539–549.
74. Calhoun JR, Nastri F, Maglio O, *et al.* (2005) Artificial diiron proteins: From structure to function. *Biopolymers* **80**: 264–278.
75. Wei PP, Skulan AJ, Wade H, DeGrado WF, Solomon EI. (2005) Spectroscopic and computational studies of the *de novo* designed protein df2t: correlation to the biferrous active site of ribonucleotide reductase and factors that affect o-2 reactivity. *J Am Chem Soc* **127**: 16098–16106.
76. Wade H, Stayrook SE, DeGrado WF. (2006) The structure of a designed diiron(iii) protein: Implications for cofactor stabilization and catalysis. *Angew Chem Int Ed* **45**: 4951–4954.
77. Ambroggio XI, Kuhlman B. (2006) Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J Am Chem Soc* **128**: 1154–1161.
78. Ali MH, Taylor CM, Grigoryan G, *et al.* (2005) Design of a heterospecific, tetrameric, 21-residue miniprotein with mixed alpha/beta structure. *Structure* **13**: 225–234.
79. Summa CM, Rosenblatt MM, Hong JK, Lear JD, DeGrado WF. (2002) Computational *de novo* design, and characterization of an a(2)b(2) diiron protein. *J Mol Biol* **321**: 923–938.
80. Lombardi A, Summa CM, Geremia S, *et al.* (2000) Retrostructural analysis of metalloproteins: application to the design of a minimal model for diiron proteins. *Proc Natl Acad Sci USA* **97**: 6298–6305.
81. Kaplan J, DeGrado WF. (2004) *De novo* design of catalytic proteins. *Proc Natl Acad Sci USA* **101**: 11566–11570.
82. Klemba M, Gardner KH, Marino S, Clarke ND, Regan L. (1995) Novel metal-binding proteins by design. *Nat Struct Biol* **2**: 368–373.
83. Benson DE, Wisz MS, Liu W, Hellinga HW. (1998) Construction of a novel redox protein by rational design: Conversion of a disulfide bridge into a mononuclear iron-sulfur center. *Biochemistry* **37**: 7070–7076.
84. Nanda V, Rosenblatt MM, Osyczka A, *et al.* (2005) *De novo* design of a redox-active minimal rubredoxin mimic. *J Am Chem Soc* **127**: 5804–5805.
85. Cochran FV, Wu SP, Wang W, *et al.* (2005) Computational *de novo* design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. *J Am Chem Soc* **127**: 1346–1347.

86. Bender GM, Lehmann A, Zou H, *et al.* (2007) *De novo* design of a single-chain diphenylporphyrin metalloprotein. *J Am Chem Soc* **129**: 10732–10740.
87. Cristian L, Nanda V, Lear JD, DeGrado WF. (2005) Synergistic interactions between aqueous and membrane domains of a designed protein determine its fold and stability. *J Mol Biol* **348**: 1225–1233.
88. North B, Cristian L, Stowell XF, *et al.* (2006) Characterization of a membrane protein folding motif the ser zipper, using designed peptides. *J Mol Biol* **359**: 930–939.
89. Yin H, Slusky JS, Berger BW, *et al.* (2007) Computational design of peptides that target transmembrane helices. *Science* **315**: 1817–1822.

This page intentionally left blank

Chapter 15

Prediction and Identification of B Cell Epitopes Using Protein Sequence and Structure Information

P. Andersen*,† D. Mkhailov‡ and O. Lund†

15.1 Introduction

Recognition of antigens by antibodies is an essential mechanism in the immune system. A general challenge in immunological research is to identify the molecular entities recognized by antibodies (B cell epitopes). This identification can lead to a better understanding of the mechanisms involved in host-pathogen interactions, and can facilitate the development of novel vaccines and antibody therapeutics. In this chapter, we give an introduction to B cell epitopes and describe methods for prediction of B cell epitopes.

The immune system is the body's defense against foreign agents, such as infectious organisms, toxins, and other molecules that are not part of the body. The defense can be divided into the innate and adaptive response.[1] The innate immune response is rapid and nonspecific; it is mediated by the recognition of conserved structural patterns found primarily in microorganisms. The adaptive immune response is

*Corresponding author.

†Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark, building 208, DK-2800 Lyngby Denmark. Email: pan@cbs.dtu.dk.

‡Center for Proteomic Chemistry, Novartis Institutes for Biomedical Research, Inc., 250 Massachusetts Ave, Cambridge, MA 02139, USA.

directed specifically to target individual infectious agents and is developed over time. After the infection is resolved, memory cells persist in the body and can induce a rapid and effective response if the harmful agent or infectious organism is encountered again.

The adaptive immune system has two major branches: the cellular immune response mediated by the T lymphocytes (T cells) and the humoral immune response mediated by antibody-secreting B lymphocytes (B cells). The responses of both systems are based on receptors that specifically bind minor parts of the foreign agent, called epitopes. The molecules containing the epitopes are called antigens, and the receptors of the humoral response are called antibodies.

B cell epitopes are recognized by antibodies, and are in general exposed on the surface of infectious organisms. Antibodies contribute to the immunity against pathogens in three main ways. The first is by neutralization, a process in which antibodies bind functional sites of antigens, thereby hindering the binding of antigens to receptors on target cells. Neutralization is important for immunity against microorganisms, which infect host cells by adhesion, and subsequently, enter the target cell to multiply. Neutralization can additionally prevent bacterial toxins from entering cells. Opsonization is the second way of antibody-mediated immunity; in this process, antibodies cover the surface of pathogens and mediate the destruction of the pathogen by phagocytosis. Finally, binding of antibodies can activate the complement system, which leads to binding of complement proteins to the pathogen. Complement proteins facilitate opsonization of the pathogen, and can form pores in the membrane of pathogens, thus leading to lysis of the cells.

15.2 B Cell Epitopes: Classification and Structural Characteristics

Antibodies can bind a variety of different types of molecules, including proteins, peptides, haptens, and polysaccharides. Often, a B cell epitope is characterized by the type of antigen it is derived from, and this can be misleading. An antibody recognizes an entity composed of atoms with specific chemical features and spatial arrangement. The entity may be formed by antigens of different nature. For instance, a peptide may be able to cross-react with an antibody raised against a carbohydrate epitope.[2] Such peptides, which can mimic natural epitopes, are called mimotopes. However, because the main focus of this chapter is the identification of epitopes in protein antigens, the term epitope will refer to residues in proteinaceous B cell epitopes.

B cell epitopes are classified into two different groups.[3] The first group consists of linear or continuous epitopes. A continuous epitope comprises a single, consecutive stretch of amino acids in the protein sequence, which is specifically recognized by an antibody raised against the intact protein.

The second group is formed by conformational or discontinuous epitopes. These are epitopes composed of residues separated in the protein sequence, but in spatial proximity because of the protein fold.

There is often no clear distinction between the two groups of B cell epitopes. A discontinuous epitope may consist of several linear epitopes, which together form the antibody interaction site. In addition, continuous epitopes may contain residues that are not interacting with the antibody.[3] Since the majority of antibodies that are raised against complete proteins do not cross-react with peptide fragments, which are derived from the same protein, it is thought that the most epitopes are discontinuous. It has been estimated that approximately 90% of B cell epitopes in globular proteins are discontinuous in nature.[4,5]

Andersen *et al.* studied statistics of discontinuous epitopes in a data set of epitopes derived from the Protein Data Bank.[6] Figure 15.1 shows the distributions for the number of residues per epitope, the number of residues per sequential stretch in epitopes, and the longest sequential stretch per epitope. The total number of residues per epitope was ranging from 9–22 and more than 60% of the epitopes consisted of 14–19 residues [Fig. 15.1(a)]. Segments with a single epitope residue represented more than 45% of the segments [Fig. 15.1(b)]. The longest sequential stretch of identified residues per epitope was ranging between 3–12 residues, and more than 75% of epitopes comprised a sequential stretch of a maximum length of 4–7 residues [Fig. 15.1(c)]. This confirms that most epitopes in the data set are

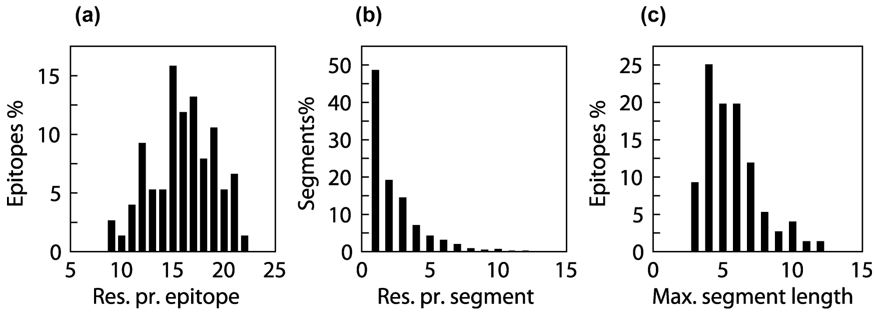


Fig. 15.1 Analysis of a dataset of discontinuous B cell epitopes. **(a)** Distribution of the number of residues per epitope. **(b)** Distribution of the number of residues per sequential stretch of epitopes. **(c)** Distribution of the maximum length of a sequential stretch per epitope. Adapted from Ref. 6, courtesy of Protein Science.

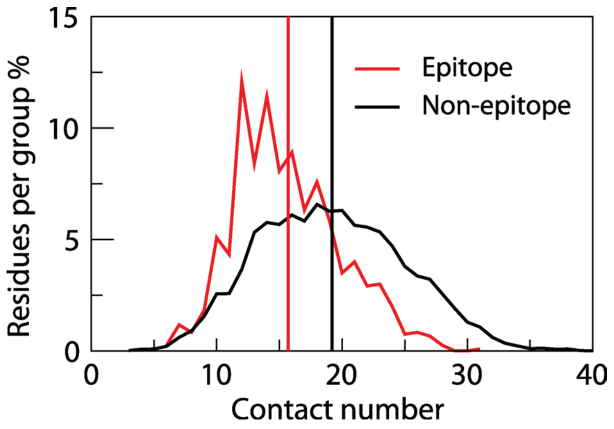


Fig. 15.2 Contact numbers of epitope residues in the dataset compared to non-epitope residues. The curves show the distribution of contact numbers for epitope residues (red curve) compared to non-epitope residues (black curve). The vertical lines represent the mean value of contact numbers for the epitope residues (red line) and for the non-epitope residues (black line) Adapted from Ref. 6, courtesy of Protein Science.

indeed discontinuous and composed by small parts of the antigen sequence forming a binding region for the antibody.

Figure 15.2 shows the surface exposure for epitopes and non-epitopes as measured by determining the number of intramolecular

C_α atom contacts for each residue. A low contact number correlates with localization close to the surface or in protruding regions of antigen structures. It can be seen that epitopes are in exposed or protruding regions, and this is in agreement with the previous analysis of B cell epitopes.[5,7]

15.3 B Cell Epitope Prediction Methods

Antibodies have been studied for many decades, and much effort has been put into the delineation of interactions between antibodies and epitopes. Even though "immunological bioinformatics" is a rather new term,[8] computational analysis and prediction of B cell epitopes have been major areas of research for more than 20 years.[4,5,9] As the entire field of biotechnology has expanded tremendously within the last two decades, developments of new methods have led to more insight into the antibody-antigen interactions as well as larger amounts of data. This, in turn, allows for the development of new bioinformatics tools and databases.[10]

In this section, we review B cell epitope predictions based on protein primary sequence and structural information.

15.3.1 *Sequence-based Prediction of B Cell Epitopes*

Sequence-based epitope prediction methods typically use propensity scales for the calculation of a prediction score. Propensity scales are composed by values that describe intrinsic features for each of the 20 amino acids. No single physico-chemical feature definitively distinguishes between epitopes and non-epitopes, but atoms that interact with the paratope have to be surface exposed. In B cell epitope prediction, some of the most successful features have been hydrophilicity, accessibility, flexibility, or loop/turn structures. In general, the predictions of these classical propensity scales correlate with features of surface exposure. Table 15.1 lists a number of the different propensity scales that have been used for epitope prediction.

In prediction methods for protein sequences, log-odds ratios are often used to reflect the probability that a given amino acid has the

Table 15.1 A Variety of Propensity Scales Used for B Cell Epitope Prediction

Feature	Year	Reference
Hydrophilicity	1981	9
Hydrophilicity	1986	11
Hydrophobicity	1982	12
Antigenicity	1985	13
Accessibility	1976	14
Surface probability	1978	15
Surface accessibility	1985	16
Backbone flexibility	1985	17
Secondary structure	1978	18
Secondary structure	1978	19
Turn prediction	1993	20

predicted property. For example, log-odds ratios can be calculated for each amino acid to reflect likelihood for it to be part of a B cell epitope. The probability of finding a given type of amino acid in an epitope of a data set can be described as:

$$p_{aa} = \frac{n_{aa}}{n} \tag{15.1}$$

where p_{aa} is the probability for an amino acid, n_{aa} is the number of times the amino acid is observed in epitopes of a data set, and n is the total number of amino acids in the data set. Because amino acids have different background frequencies, q_{aa}, the probabilities are often divided by the background frequencies observed in a large database such as SwissProt.[21] The log-odds ratio is calculated as:

$$L = \log_2\left(\frac{p_{aa}}{q_{aa}}\right) \tag{15.2}$$

A high log-odds ratio L indicates that the amino acid is more frequently observed in B cell epitopes than in the Swiss-Prot database.

The log-odds ratios may be calculated for both epitope residues and non-epitope residues and then subtracted to get the final epitope log-odds ratios:

$$L_{e-ne} = L_e - L_{ne} \tag{15.3}$$

where L_e is the log-odds ratio of a given amino acid type derived from epitope residues, L_{ne} is the log-odds ratio of a given amino acid type derived from non-epitope residues and L_{e-ne} is the final epitope log-odds ratio for the given amino acid type.

If a data set contains many similar examples for training, the log-odds ratios can easily be biased toward the redundant examples. To avoid this problem, several refinement techniques can be applied.[8] For instance, sequence weighting can be used; first the similar sequences are clustered, then weights are assigned for each sequence to down-regulate the influence of highly similar sequences.

Propensity scales are often used in combination with smoothing procedures. The simplest type of smoothing is based on the sliding of a window through the protein sequence and averaging the propensities of the residues within the window. The mean value of the window is then assigned to the residue in the middle of the window. This simple type of smoothing has been used frequently for B cell epitope prediction, but more sophisticated methods using a weighted average, or a Gaussian smoothing curve have also been applied.[22] The smoothing results in a scoring profile, where the high-scoring regions are predicted to be antigenic.

Several tools for antigenicity prediction using combinations of propensity scales have been developed.[23–27] However, the most extensive sequence-based study, involving 484 different propensity scale methods on a new data set of 50 proteins, concluded that most propensity scales perform close to random, and the use of more sophisticated machine-learning methods such as artificial neural networks (ANNs) was proposed.[28]

Recently, different approaches using advanced machine learning methods were published. Saha *et al.* proposed a prediction method based on recurrent ANNs trained on a data set of 700 continuous epitopes from the Bcipep database.[29,30] Larsen *et al.* presented the Bepipred method[31] based on predictions of a Hidden Markov Model (HMM) in combination with the Parker hydrophilicity scale.[11] The method was trained on continuous epitopes of 127 proteins from the AntiJen database.[32] Söllner *et al.* developed a classification algorithm based on the combinations of propensities with sequential residue neighborhood parameters.[33] The classification algorithm was based on decision trees and nearest neighbor approaches, and was trained on publicly available data sets from Bcipep and FIMM[29,34] and a large amount of proprietary data. In total, 1211 epitopes and non-epitope sequences were used for training, and the performance was shown to be greatly increased compared to single propensity scale methods.[33] Although the performance of these new methods is improved compared to propensity scale methods, accurate prediction of continuous epitopes remains a challenge in the field of immunological bioinformatics.

15.3.2 *Prediction of B Cell Epitopes Based on Protein Structure*

Protein 3D structures have also been used in the prediction of B cell epitopes. Particularly, the prediction of discontinuous epitopes is thought to require this information. The growing number of experimental protein structures in the PDB further expands possibilities to predict discontinuous epitopes for novel targets based on 3D homology models.[35]

The first prediction methods on the basis of protein structure were published in 1986, and were based on epitopes being on the surface of proteins. The method proposed by Novotny *et al.*[7] was based on the contacts of a large sphere (called a probe) on the Van der Waals surface of the protein. A similar method is used for calculating solvent accessible areas.[36] Novotny *et al.* found that the contacts of a probe with 10 Å radius correlate well with antigenic epitopes in hen egg white lysozyme, sperm whale myoglobin, cytochrome C, and myohemerythrin. It was additionally observed that the contacting residues of the large probe also had high solvent accessibility scores, as determined by using a probe size of 1.4 Å.

Thornton *et al.*[5] used ellipsoids with the same inertia moment as the protein structure in order to model the overall shape of the protein. The sizes of the ellipsoids were varied and chosen so that for protrusion index 9 (PI 9), 90% of the atoms in the structure were inside the ellipsoid. The rest of the atoms (10%) were protruding from (sticking out of) the structure. In general, it was found that antigenic peptides tend to protrude from structures of the proteins lysozyme, myoglobin, and myohemerythrin. In addition, it was found that protruding residues had a tendency to be more flexible and accessible.

After the publication of studies by Novotny *et al.* and Thornton *et al.*, little was reported about epitope prediction based on protein structure. However, the numbers of available structures of both antigens and antibody-antigen complexes have increased, and new approaches have recently been used for this type of epitope prediction.

The server for Conformational Epitope Prediction (CEP)[37] is based on the calculation of the relative surface accessibility (RSA). Sequence fragments of surface-exposed residues are identified and condensed with other proximal exposed fragments in the structure; this defines regions on the 3D surface that are exposed, and which possibly act as antigenic regions.

DiscoTope is another structure-based method that predicts residues which are part of epitopes.[6] DiscoTope combines surface localization and the spatial properties of a protein structure with a novel epitope propensity scale. The combination is defined in terms of a simple weighted sum of the contact number and a sum of sequentially averaged epitope log-odds ratios of spatially proximate residues.

To evaluate predictions of B cell epitope residues, which are mapped using other types of methods than X-ray crystallography, the predictions of DiscoTope were tested on the structure of the ectodomain from the *Plasmodium falciparum* apical membrane antigen-1 (AMA1).[38,39] No AMA1 epitopes were included in the training set for the method. However, two separate epitopes recognized by monoclonal antibodies Mab1F9 and Mab4G2 have been experimentally mapped on the AMA1 ectodomain. The Mab1F9 epitope was mapped using phage display of peptides and point mutations of E197.[40] The discontinuous Mab4G2 epitope was mapped in detail by point mutation of nine residues.[39] In addition, five residues (including E197 and other residues in the same region of the structure) were classified as highly polymorphic in AMA1 sequences.[38] It has been suggested that the polymorphism is caused by selection pressure on the antigen to avoid the host immune system. The DiscoTope threshold of –4.7, corresponding to a specificity of 90%, and 24% sensitivity was used in epitope prediction.[6] In AMA1, 43 of 311 residues were predicted as epitope residues. Most of the predicted epitope residues cluster in three separate regions of the AMA1 structure (Fig. 15.3). DiscoTope successfully identified two of the eight residues in the 1F9 epitope, which were mapped using phage display (D196 and E197). In the discontinuous 4G2 epitope, all nine residues except D348 were predicted to be part of epitopes. All of the five highly polymorphic residues were predicted to be located in epitopes. Thus, DiscoTope

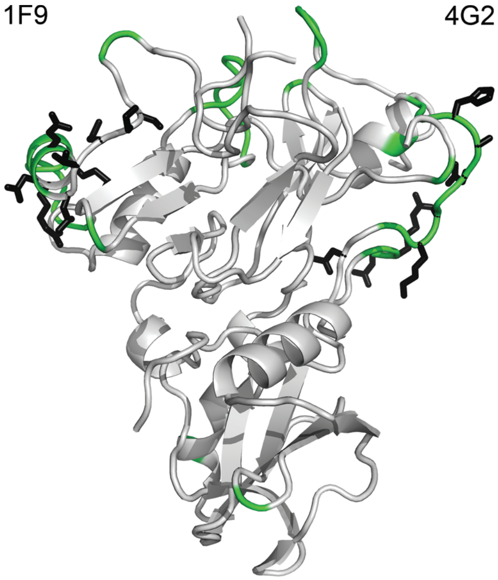


Fig. 15.3 Predicted epitope residues of the AMA1 ectodomain. Backbone atoms of residues predicted by DiscoTope as parts of epitopes are highlighted in green. Side-chains of the residues mapped to the monoclonal antibodies 1F9 and 4G2 are shown in black. Adapted from Ref. 6, courtesy of Protein Science.

successfully predicted epitope residues of AMA1 that had been mapped by using diverse methods.

The Epitope Mapping Tool (EMT)[41] is based on a sequence library of epitopes in different proteins identified by phage display. In the prediction method, the antigen structure is searched to find surface-exposed regions containing motifs of the library.

Rapberger *et al.* recently published a study of antigen-antibody interaction site prediction.[42] Their method uses surface accessibility measured with a probe radius of 3 Å. Additionally, the shape complementarity to paratopes and interaction energies are included to identify residues with high probability of being part of discontinuous epitopes. Similarly to the DiscoTope method, the work was based on discontinuous epitopes derived from PDB structures of antibody-antigen complexes. The method was tested using structures of free antigens in unbound conformations. It was shown to have a moderate performance and captured three of eight residues in the 1F9 epitope of the AMA1 protein, which were identified by using phage display.[40]

A number of tools have been developed to facilitate the mapping of epitopes on protein structures by using mimotopes. The methods analyze the protein 3D structure to find regions that can be mimicked by peptides.[43]

15.4 Future Outlook

One of the main goals of B cell epitopes research is to develop vaccines or diagnostic tools. Historically, vaccines have been based on responses to the entire pathogen. Killed or attenuated organisms have been used for the vaccines to raise the immunity while avoiding hazardous effects. These strategies have been effective in diminishing the occurrence of major diseases, such as smallpox and polio. However, there are several drawbacks of these types of vaccines. In general, the practical use of killed or attenuated pathogens can be affected by problems caused by producing pathogen in sufficient amounts, safety, and the genetic evolution of pathogens.[44]

Today, the field is moving more toward "rational vaccine design." The basic idea of this approach is to use knowledge about the pathogen, the immune responses against the pathogen, and general host-pathogen interactions in order to design more efficient vaccines. However, even the rational approach to vaccine design is still heavily dependent on experimental trials, since it is hard to predict the response of a new vaccine in complex systems such as the human body.[45] Some of the challenges in rational vaccine design are discussed below.

One general approach of modern vaccine design is the use of more simple vaccine formulations containing non-infecting subunits of the pathogen.[46] Subunit vaccines have shown to be useful for vaccination. For instance, virus-like particles (VLPs) are assembled of the human papilloma virus (HPV) proteins L1 and L2, and the VLPs are used for HPV vaccines preventing viral infection and lowering the risk of genital cancer.[47] One of the advantages of this approach is that these new vaccines should be safer and not lead to infections.

Another example of a recombinant subunit vaccine, which has been approved for human use, is the hepatitis B virus (HBV) DNA vaccine.[48] This vaccine is based on the HBV surface antigen (HBsAg) and produced in genetically modified yeast.

Recently, one of the major research topics in rational vaccine design has been human immunodeficiency virus (HIV) vaccines. A vaccine, which elicits neutralizing antibodies and effectively protects against HIV infection, has shown to be extremely difficult to develop, and multiple approaches using modern vaccine design technologies are still used in the pursuit of an HIV vaccine.[49] Other examples of major research projects are B cell-based vaccines against infection of malaria parasites. The RTS,S subunit vaccine has been an outcome of these malaria research projects. This vaccine is based on a fusion protein of a polypeptide from *Plasmodium falciparum* circumsporozoite protein and HbsAg,[50] and clinical trials of the vaccine have been promising.[51] However, even though the number of technologies for vaccine design is growing, developing a vaccine is a complex task that is still mostly based on labor-intensive experimental studies.

B cell epitope-based diagnostic tools also constitute a major research topic. In the diagnosis of infectious diseases, B cell epitope binding assays can be used to detect humoral responses against pathogen-specific epitopes.[52,53] In addition, B cell epitopes have a potential for the diagnosis of autoimmune diseases,[54,55] allergy,[56] and cancer.[57] The development of effective B cell epitope-based diagnostic tools can be challenging because non-specific antibody cross-reactivity and reactivity resulting from exposure, but not infection, can affect the rate of false positives.

15.4.1 *Vaccines Based on Linear Epitopes or Peptides*

Peptides containing linear epitopes are considered to have a high potential for vaccines. In addition to the advantages of the subunit vaccines mentioned above, peptides are easily synthesized, purified, stored, and handled. However, it has become clear that efficient peptide-based vaccines, in general, are complex to develop.

Peptides in vaccines must be immunogenic, have the ability to elicit antibodies that cross-react with the native protein, and that protect against infection or pathogenesis. Studies testing peptide epitopes are based on cross-reactivity: the ability for a peptide to be recognized by an antibody raised against a native protein. However, these cross-reactive peptides are usually not very immunogenic and there can be several reasons for this lack of immunogenicity: Most peptide-based vaccines would rely on $CD4^+$ T helper cell-initiated immune responses, and the B cell epitope itself may not contain an MHC class II epitope. In future vaccine design, this can be solved by fusing the peptide with residues containing an efficient MHC class II epitope.[58]

Another major problem is that the binding of antibodies to continuous epitopes is conformation-dependent. A peptide in solution may have a broad variety of structural conformations, and the conformations to which antibodies are developed may not be similar to the conformation in the native protein. To solve this problem, conformationally restricted peptides have been tested; for instance, disulphide bridges or other covalent links have been introduced in peptides to stabilize loop conformations.[59,60] Conformational restriction of peptides can also help to circumvent another problem of short peptides, which is fast degradation by peptidases in the human body.[61] We consider computational protein design to be useful in the future development of stable peptides that are more conformationally restricted, and therefore, able to present epitopes more efficiently.

Another problem in peptide-based vaccines is that the humoral immune response is more efficiently initiated when the epitopes are repeatedly presented on larger antigens. To circumvent this problem, a number of different adjuvant and carrier systems have been studied. For instance, VLPs are used in development of a foot-and-mouth disease virus vaccine to present continuous epitopes in a manner that facilitates an immune response.[62]

15.4.2 *Vaccines Based on Discontinuous Epitopes*

Discontinuous epitopes are in general more difficult to use than linear ones for vaccine design. Because the epitope is composed of different parts of the protein sequence, it is more complex to conserve the native conformation of the epitope in a recombinant protein or a peptide. As mentioned earlier in this chapter, the majority of natural epitopes are thought to be discontinuous. Therefore, much effort is put into the development of vaccines based on discontinuous epitopes. Mimotopes are considered to have a high potential in vaccines by mimicking discontinuous epitopes,[63,64] and future prediction algorithms for mimotopes on the basis of protein structure have a potential in vaccine design. Recombinant proteins are also considered useful for the presentation of discontinuous epitopes in vaccines. Koide *et al.* successfully applied a structure-based design for a Lyme disease vaccine candidate.[65] The authors removed approximately 45% of the residues from the OspA and stabilized the engineered protein by promoting hydrophobic interactions. The engineered protein presented discontinuous epitopes and had an affinity to monoclonal antibodies similar to the full-length OspA protein. A recent study investigated the binding of the neutralizing antibody b12 to stabilized protein constructs derived from the HIV protein gp120,[66] and suggested such constructs have a high potential for HIV vaccine design.[49]

However, the use of engineered proteins also has disadvantages. Recombinant proteins for vaccines must be stable and easily produced in sufficient amounts; for some proteins, this is a limiting factor. In addition, the engineered proteins may contain new epitopes on the surface that could be immunodominant and lead to unwanted immune responses.

Computational protein design has a great potential to influence future vaccine development. Using protein structure prediction, a suitable scaffold protein could be chosen to present an epitope. Scoring functions could help to predict mutations that stabilize new recombinant proteins used for vaccines. However, the predicted constructs would need to be tested intensively in experimental studies to identify which of them present relevant epitopes that elicit protecting antibodies, and at the same time do not lead to unwanted side-effects.

In general, the increased number of publicly available methods for B cell epitope prediction shows that even though the performances are still quite low, when compared to MHC class I epitope prediction, there is much optimism among researchers in the field. Many research groups are continuously working on improving methods, building databases, and evaluating different methods.[67] Thus, the field of B cell epitope prediction is expected to lead to further improved methods, which can aid in experimental epitope mapping and vaccine design.

Acknowledgments

The authors wish to thank Marc H. V. Van Regenmortel for fruitful discussions and comments on the manuscript.

References

1. Paul WE. (2003) *Fundamental Immunology*, 5th ed. Lippincott Williams and Wilkins, Philadelphia, USA
2. Lo Passo C, Romeo A, Pernice I, *et al.* (2007) Peptide mimics of the group B meningococcal capsule induce bactericidal and protective antibodies after immunization. *J Immunol* **178**: 4417–4423.

3. Van Regenmortel MHV. (1996) Mapping epitope structure and activity: from one-dimensional prediction to four-dimensional description of antigenic specificity. *Methods* **9**: 465–472.
4. Barlow DJ, Edwards MS, Thornton JM. (1986) Continuous and discontinuous protein antigenic determinants. *Nature* **322**: 747– 748.
5. Thornton JM, Edwards MS, Taylor WR, Barlow DJ. (1986) Location of "continuous" antigenic determinants in the protruding regions of proteins. *EMBO J* **5**: 409–413.
6. Andersen P, Nielsen M, Lund O. (2006) Prediction of residues in discontinuous B cell epitopes using protein 3D structures. *Protein Sci* **15**: 2558–2567.
7. Novotny J, Handschumacher M, Haber E, *et al.* (1986) Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc Natl Acad Sci USA* **83**: 226–230.
8. Lund O, Nielsen M, Lundegaard C, Kesmir C, Brunak S. (2005) *Immunological Bioinformatics (Computational Molecular Biology),* 1st ed. MIT Press, Cambridge, Massachusetts.
9. Hopp TP, Woods KR. (1983) A computer program for predicting protein antigenic determinants. *Mol Immunol* **20**: 483–489.
10. Korber B, LaBute M, Yusim K. (2006) Immunoinformatics comes of age. *PLoS Comput Biol* **2**: E71.
11. Parker JM, Guo D, Hodges RS. (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry* **25**(19): 5425–5432.
12. Kyte J, Doolittle RF. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**: 105–132.
13. Welling GW, Weijer WJ, van der Zee R, Welling WS. (1985) Prediction of sequential antigenic regions in proteins. *FEBS Lett* **188**: 215–218.
14. Chothia, C. (1976) The nature of the accessible and buried surfaces in proteins. *J Mol Biol* **105 (1)**: 1–12.
15. Janin J, Wodak S. (1978) Conformation of amino acid side-chains in proteins. *J Mol Biol* **125**: 357–386.
16. Emini EA, Hughes JV, Perlow DS, Boger J. (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol* **55**: 836–839.
17. Karplus PA, Schulz GE (1985) Prediction of chain flexibility in proteins: a tool for the selection of peptide antigens. *Naturwissenschaften* **72**: 212–213.
18. Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* **47**: 45–148.

19. Garnier J, Osguthorpe DJ, Robson B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* **120**: 97–120.
20. Pellequer JL, Westhof E (1993) PREDITOP: a program for antigenicity prediction. *J Mol Graph* **11**(3): 204–210.
21. Bairoch A, Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl Acids Res* **28**: 45–48.
22. Pellequer JL, Westhof E, Van Regenmortel MH. (1991) Predicting location of continuous epitopes in proteins from their primary structures. *Meth Enzymol* **203**: 176–201.
23. Alix AJ. (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine* **18**: 311–314.
24. Jameson BA, Wolf H. (1988) The antigenic index: a novel algorithm for predicting antigenic determinants. *Comput Appl Biosci* **4**: 181–186.
25. Kolaskar AS, Tongaonkar PC. (1990) A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett* **276**: 172–174.
26. Maksyutov AZ, Zagrebelnaya ES. (1993) ADEPT: a computer program for prediction of protein antigenic determinants. *Comput Appl Biosci* **9**: 291–297.
27. Pellequer JL, Westhof E. (1993) PREDITOP: a program for antigenicity prediction. *J Mol Graph* **11**: 204–210.
28. Blythe MJ, Flower DR (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci* **14**: 246–248.
29. Saha S, Bhasin M, Raghava GPS. (2005) Bcipep: a database of B cell epitopes. *BMC Genom* **6**: 79.
30. Saha S, Raghava GPS. (2006) Prediction of continuous B cell epitopes in an antigen using recurrent neural network. *Proteins* **65**: 40–48.
31. Larsen JEP, Lund O, Nielsen M. (2006) Improved method for predicting linear B cell epitopes. *Immunome Res* **2**: 2.
32. Toseland CP, Clayton DJ, McSparron H, *et al.* (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res* **1**: 4.
33. Söllner J, Mayer B. (2006) Machine learning approaches for prediction of linear B cell epitopes on proteins. *J Mol Recogn* **19**: 200–208.
34. Schonbach C, Koh JLY, Flower DR, Wong L, Brusic V. (2002) FIMM, a database of functional molecular immunology: update 2002. *Nucl Acids Res* **30**: 226–229.
35. Guex N, Peitsch, MC. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**: 2714–2723.
36. Lee BK, Richards FM. (1971) The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**: 379–400.
37. Kulkarni-Kale U, Bhosle S, Kolaskar AS. (2005) CEP: a conformational epitope prediction server. *Nucl Acids Res* **33**: 168–171.

38. Bai T, Becker M, Gupta A, *et al.* (2005) Structure of AMA1 from plasmodium falciparum reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. *Proc Natl Acad Sci USA* **102**: 12736–12741.
39. Pizarro JC, Vulliez B, Chesne SML, *et al.* (2005) Crystal structure of the malaria vaccine candidate apical membrane antigen 1. *Science* **308**: 408–411.
40. Coley AM, Parisi K, Masciantonio R, *et al.* (2006) The most polymorphic residue on plasmodium falciparum apical membrane antigen-1 determines binding of an invasion-inhibitory antibody. *Infect Immunol* **74**: 2628–2636.
41. Batori V, Friis EP, Nielsen H, Roggen EL. (2006) An *in silico* method using an epitope motif database for predicting the location of antigenic determinants on proteins in a structural context. *J Mol Recogn* **19**: 21–29.
42. Rapberger R, Lukas A, Mayer B. (2007) Identification of discontinuous antigenic determinants on proteins based on shape complementarities. *J Mol Recogn* **20**: 113–121.
43. Meloen RH, Puijk WC, Slootstra JW. (2000) Mimotopes: realization of an unlikely concept, *J Mol Recogn* **13**: 352–359
44. Arnon R, Ben-Yedidia T. (2003) Old and new vaccine approaches. *Int Immunopharmacol* **3**: 1195–1204.
45. Van Regenmortel MHV. (2007) The rational design of biological complexity: a deceptive metaphor. *Proteomics* 7: 965–975.
46. Ellis RW. (1999) New technologies for making vaccines. *Vaccine* 17: 1596–1604.
47. Ljubojevic S. (2006) The human papillomavirus vaccines. *Acta Dermatovenerol Croat* **14**: 208.
48. Keating GM, Noble S. (2003) Recombinant hepatitis B vaccine (Engerix B): a review of its immunogenicity and protective efficacy against hepatitis B. *Drugs* **63**: 1021–1051.
49. Douek DC, Kwong PD, Nabel GJ. (2006) The rational design of an AIDS vaccine. *Cell* **124**: 677–681.
50. Heppner DGJ, Kester KE, Ockenhouse CF, *et al.* (2005) Towards an RTS, S-based, multistage, multi-antigen vaccine against falciparum malaria: progress at the Walter Reed Army Institute of Research. *Vaccine* **23**: 2243–2250.
51. Matuschewski K. (2006) Vaccine development against malaria. *Curr Opin Immunol* **18**: 449–457.
52. Hamby CV, Llibre M, Utpat S, Wormser GP. (2005) Use of peptide library screening to detect a previously unknown linear diagnostic epitope: proof of principle by use of lyme disease sera. *Clin Diagn Lab Immunol* **12**: 801–807.
53. Hsueh, PR, Kao CL, Lee CN, *et al.* (2004) SARS antibody test for serosurveillance. *Emerg Infect Dis* **10**: 1558–1562.
54. Mahler M, Bluthner M, Pollard KM. (2003) Advances in B-cell epitope analysis of autoantigens in connective tissue diseases. *Clin Immunol* **107**: 65–79.

55. Selak S, Mahler M, Miyachi K, Fritzler ML, Fritzler MJ. (2003) Identification of the B-cell epitopes of the early endosome antigen 1 (EEA1). *Clin Immunol* **109**: 154–164.
56. Eigenmann PA. (2004) Do we have suitable *in vitro* diagnostic tests for the diagnosis of food allergy? *Curr Opin Allergy Clin Immunol* **4**: 211–213.
57. Valmori D, Souleimanian NE, Hesdorer CS, *et al.* (2005) Identification of B cell epitopes recognized by antibodies specific for the tumor antigen NY-ESO-1 in cancer patients with spontaneous immune responses. *Clin Immunol* **117**: 24–30.
58. Sabhnani L, Manocha M, Sridevi K, *et al.* (2003) Developing subunit immunogens using B and T cell epitopes and their constructs derived from the F1 antigen of *Yersinia pestis* using novel delivery vehicles. *FEMS Immunol Med Microbiol* **38**: 215–229.
59. Cabezas E, Wang M, Parren PW, Stanfield RL, Satterthwait AC. (2000) A structure-based approach to a synthetic vaccine for HIV-1. *Biochemistry* **39**: 14377–14391.
60. Sabo JK, Keizer DW, Feng ZP, *et al.* (2007) Mimotopes of apical membrane antigen-1: structures of phage-derived peptides recognized by the inhibitory monoclonal antibody 4G2dc1 and design of a more active analogue. *Infect Immun* **75**: 61–73.
61. Hans D, Young PR, Fairlie DP. (2006) Current status of short synthetic peptides as vaccines. *Med Chem* **2**: 627–646.
62. Zhang YL, Guo YL, Wang KY, *et al.* (2007) Enhanced immunogenicity of modified hepatitis B virus core particle fused with multiepitopes of foot-and-mouth disease virus. *Scand J Immunol* **65**: 320–328.
63. Saphire EO, Montero M, Menendez A, *et al.* (2007) Structure of a high-a.nity "mimotope" peptide bound to HIV-1-neutralizing antibody b12 explains its inability to elicit gp120 cross-reactive antibodies. *J Mol Biol* **369**: 696–709.
64. Untersmayr E, Szalai K, Riemer AB, *et al.* (2006) Mimotopes identify conformational epitopes on parvalbumin, the major.sh allergen. *Mol Immunol* **43**: 1454–1461.
65. Koide S, Yang X, Huang X, Dunn JJ, Luft BJ. (2005) Structure-based design of a second-generation Lyme disease vaccine based on a C-terminal fragment of *Borrelia burgdorferi* OspA. *J Mol Biol* **350**: 290–299.
66. Zhou T, Xu L, Dey B, *et al.* (2007) Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature* **445**: 732–737.
67. Greenbaum JA, Andersen PH, Blythe M, *et al.* (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J Mol Recogn* **20**: 75–82.

This page intentionally left blank

Chapter 16

Computational Antibody Engineering

T. K. Nevanen*,†, N. Munck† and U. Lamminmäki‡

16.1 Introduction

Antibodies (immunoglobulins) are binding proteins produced by B lymphocytes to defend an organism against invading pathogens and foreign macromolecules. They are widely used as a research tool, in diagnostics, and recently also as therapeutic agents. The unique features of antibodies are their enormous diversity and specificity of recognition, which are, together with their high affinity, the most essential attributes for their applications. The specificity and affinity of the binding originates from the unique structural elements of the antibody proteins, as reviewed in Section 16.1.1. The generation of antibodies *in vivo* is induced mainly by infection or by immunizing with an immunogen. The development of phage-displayed antibody gene libraries as *in vitro* sources of various binding specificities has enabled the selection of antibodies to almost any antigen, as recently reviewed by Hoogenboom.[1]

However, occasionally the properties of primary antibodies need to be optimized for their application. Antibody structures, when available, are very useful in antibody engineering. Sequence data and

*Corresponding author.
†VTT Technical Research Centre of Finland, P.O.Box 1000, FI-02044 VTT Espoo, Finland. Email: tarja.nevanen@vtt.fi.
‡Department of Biotechnology, University of Turku, 20520 Turku, Finland.

homology modeling provide an alternative way for knowledge-based design of engineering when the crystal structure is not available. Computational antibody engineering, which includes the use of both structure- and homology-based modeling, depending of the data available for the antibody of interest, has been utilized to obtain structural information on the antigen-combining site, antibody-antigen complex, and to design mutants for improved performance as reviewed in Section 16.4. Here, the focus is on homology modeling using existing databases of antibody structures as a starting point (Section 16.2).

16.1.1 *Structural Elements of Antibodies*

Four polypeptides, two identical heavy chains, and two identical light chains, form the basic structure of a Y-shaped antibody as drawn schematically in Fig. 16.1(a). Heavy and light chain polypeptides fold into four and two barrel-like globular domains ("immunoglobulin domains"), respectively having about 110 amino acids in each domain. Domains have two antiparallel β-sheets, which form the

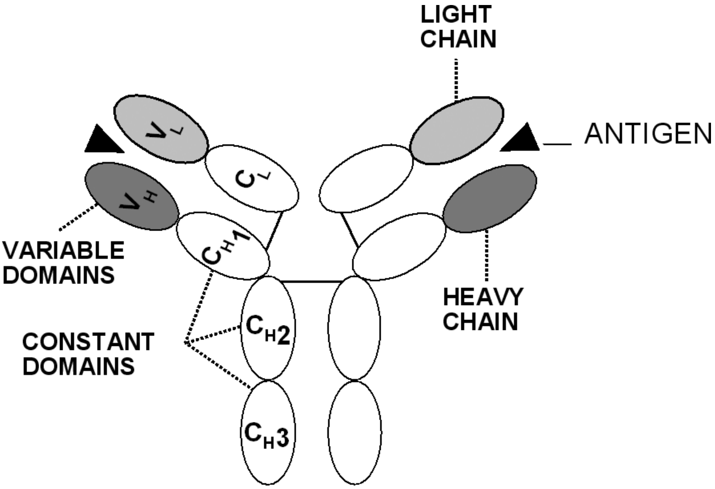


Fig. 16.1 **(a)** Domain structure of an IgG antibody. **(b)** The barrel-like framework structure of variable domains. Both domains have three hypervariable loops responsible for antigen binding.

framework structure [Fig. 16.1(b)] and are stabilized by a disulfide bond. Domains pack against each others by interactions between β-sheets, except the C_H2-C_H2 interaction, which is mediated by carbohydrate moieties. Constant domains (C_L or C_H1, C_H2, C_H3) have conserved sequences and structures among each sub-class (isotypes are κ or λ for the light chain and γ, μ, ε, δ or α for the heavy chain). The carboxyl-terminal domains of the heavy chain have various effector functions, which are important for the cellular immune response.

The amino terminal domains of both heavy and light chains, called variable domains (V_H and V_L), are responsible of the diversity of the specific antigen binding. Each variable domain has four conserved framework sequences forming the barrel-like structure of two antiparallel β-sheets.[2] The three hypervariable loops in each chain are H1, H2, and H3 for the heavy chain and L1, L2, and L3 for the light chains.[3] Hypervariable loops are located between the framework sequences in each chain [Fig. 16.1(b)]. An antigen binding site is a combination of these six loops (complementarity determining regions, CDRs) of the variable domains of heavy and light chains. The most variable residues at CDRs are mostly in contact with the antigen and they are called the specificity determining residues.[4] Framework regions have mainly a structural role in variable domains and the contribution of the framework residues to the affinity is usually indirect. They support the structure of the CDRs[5–7] or affect the correct folding.[8] However a framework residue may also be in direct contact with the antigen.[9,10] The hypervariable loops, especially the CDR3 of the heavy chain, vary in length and have high sequence diversity. The length of the loop and certain conserved key residues in CDRs and in the framework region determine the main-chain conformation of the hypervariable loop. The critical amino acids are important in packing and hydrogen bonding, or have suitable torsion angles. Thus the hypervariable loops have canonical structures,[11–16] which facilitate the computational modeling of the structure.

The conformations of uncomplexed and complexed antibodies vary in many cases in the crystal structures, suggesting a structural adaptation of the antibody conformation in the binding event (more detailed in Section 16.3). Changes in conformation upon binding bring challenges to the modeling.

16.2 Antibody Modeling

The computational modeling of antibodies, the history of which goes back over three decades,[17,18] plays a pivotal role in antibody research and engineering today. Antibodies are favorable targets to be studied by molecular modeling because of several reasons. First of all, despite their capacity to provide enormous functional diversity, the members of this protein class are highly conserved in large parts of their structure. Modeling of the strictly conserved constant regions, forming the majority of the antibody structure, is straightforward with standard homology modeling techniques, and major efforts can be focused to the prediction of the structure of a fairly limited region — the antigen combining site. In fact, the term antibody modeling is primarily understood to refer to the prediction of the structure of the variable domain, and in particular, the structure of the antigen-combining site within it. Secondly, the intensive studies on the structure/function of antibodies have yielded a block of useful data concerning the structure and properties of antibody variable domains. Yet another aspect facilitating the modeling is the large and constantly increasing number of experimentally determined structures of antibody molecules available in the Protein Data Bank (PDB)[19] to be used as templates for homology modeling. Generally, at least one high-resolution template structure with sequence identity ranging from 45% to over 90% can be identified for an antibody sequence of interest.

16.2.1 *Modeling of the Framework Regions*

The first step in building a model for an antibody variable region is the modeling of the framework. The PDB is searched for templates with high sequence identity. The templates of the light and heavy chain do not need to be derived from the same structure, and consequently, the searches are performed separately for the variable domains of the two chains. As most of the variable domain is encoded by the V-gene, it can be worth running the searches using only the V-gene encoded part of the light and heavy chain as a query in order to avoid biased ranking of the search results by highly variable CDR3 (especially in the case of heavy chain) region and J-gene encoded region. Most commonly, one light and one heavy chain variable domain structure showing the highest sequence identity with the target sequences are used as templates. However, if several template candidates with similar level of identity are available, it is useful to select a non-redundant (each template originates from different antibody) set of templates for modeling. The comparison of the multiple templates can allow the identification of regions where the structure is locally distorted, e.g. by a mutation or crystal packing effects. In the presence of multiple templates, the distorted regions as well as the segments having high B-factors can easily be grafted from one of the templates with a proper structure. High resolution structures should be favored over low resolution ones, and if several templates are available, the low resolution ones can be dropped. Morea *et al.* (2000)[20] suggested that, when using a single template, the structure having significantly higher resolution should be used as a template if the difference in sequence identity between the high and low resolution structures is 5% or less. An alternative way to generate a template for modeling is to calculate average coordinates of several overlaid antibody variable domains.[21] Generation of the sequence alignment for the modeling is facilitated by the facts that well-conserved templates are generally available, and that there are arrays of other antibody sequences that can be used for the construction of multiple sequence alignments. The structures of template molecules (light and heavy chain templates separately) are superimposed by using the Cα-atoms of the residues in the conserved B-sheets of the framework.[22] The coordinates for the framework residues can be assigned automatically by various homology modeling programs. A potential source of confusion in antibody modeling lies in the fact that there are several different numbering schemes for variable domains, the most well-known of which are the Kabat,[23] Chothia,[11] IMTG,[24] and AHo[22] numbering schemes. For example, the residue numbering in the known structures can be either based on some of these schemes or the standard PDB numbering. Helpful information for understanding the differences of the schemes can be found in the article by Honegger and Plückthun (2001)[22] as well as on well-maintained internet sites (http://www.bioinf.org.uk/abs/ and http://www.bioc.uzh.ch/antibody/).

16.2.2 *Modeling of the Hypervariable Regions*

In the next phase, the CDR loops are built on the framework model. The methods used for the modeling of CDRs involve both knowledge-based as well as *ab initio* modeling techniques. The most widely used approach takes advantage of the fact that most of the CDRs, especially CDR1 and CDR2 of the light and heavy chain as well as CDR3 of the light chain, tend to adapt a limited number of backbone conformations, which have been termed canonical structures.[11] Strict sequence based rules have been defined to allow the prediction of the canonical type of the CDRs on the basis of the loop length and on the presence of certain key-residues within the CDRs or at neighboring framework positions.[11–13,16] Helpful rules for the identification of CDRs in the sequence can be found on the internet page of Andrew Martin (Table 16.1). Those CDRs of the template that have the correct canonical structure can be retained in the model and the other canonical CDRs are grafted from other antibody structures. The grafted CDRs are fitted to the model by overlaying the main chain atoms of a few residues preceding and following the CDR region.[20]

Sometimes, part of the CDR1 and CDR2 loops of both chains or the CDR3 loop of the light chain fall out of the canonical classification. One approach to predict the structures of these loops is to produce a large number of different loop structures to saturate the conformational space, and then to use a scoring algorithm for ranking the loop candidates. The sampling of the conformational space can be obtained by using database searches and/or *ab initio* conformational sampling programs such as CONGEN.[25] These two techniques are combined in the program CAMAL Combined Antibody Modeling Algorithm,[13] which was designed for antibody loop

Table 16.1 A Collection of Internet Resources Providing Useful Information and Tools for Antibody Engineering and Modeling

Name of Resource	Short Description of Content	URL Address
WAM	Web antibody Modelling server by Anthony Rees's group for automated modelling of antibodies.	http://antibody.bath.ac.uk/
IMTG	A large database of nucleic acid sequence information on immunoglobulins and other molecules of immunological interest.	http://imgt.cines.fr/
V-base	A comprehensive directory of human germline variable region sequences.	http://vbase.mrc-cpe.cam.ac.uk/
V-base2	Database of human and mouse immunoglobulin sequences extracted from the EMBL-bank.	http://www.vbase2.org/
Kabat DB	A large database of sequences of immunoglobulins. Currently shut down! Content of the database available for compensation.	http://www.kabatdatabase.com/
AAAAA	Annemarie Honegger's antibody page with a plenty of useful information on antibody sequence and structure data as well as on antibody modelling.	http://www.bioc.uzh.ch/antibody/
Andrew Martin's homepage	Useful information and tools for antibody research.	http://www.bioinf.org.uk/abs/
Mike Clarck's homepage	Information on antibody structure, function and humanization.	http://www.path.cam.ac.uk/~mrc7/
Homepage of Juan Carlos Almagro's research group	Directories of immunoglobulin variable region sequences of mouse and large farm animals. Directory of antibody structures determined experimentally until year 2000.	http://www.ibt.unam.mx/vir/index.html

(*Continued*)

Table 16.1 (*Continued*)

Name of Resource	Short Description of Content	URL Address
Stefan Dubel's recombinant antibody page	Information about recombinant antibodies and lot's of links.	http://rzv054.rz.tu-bs.de/Biotech/SD/SDscFVSite.html
Humanization by Design	José Saldanha's page providing introduction into antibody humanization and a directory with descriptions of humanized antibodies.	http://rzv054.rz.tu-bs.de/Biotech/SD/SDscFVSite.html
The antibody society	Recently established forum for the field of recombinant antibodies, antibody engineering and related areas.	http://www.antibodysociety.org/

modeling applications. Alternatively, the loops having insertions or deletions are derived from a known loop with most similar sequence by adding or deleting, respectively, residues in the existing loop.[26] In this case, the torsion angles of the residues preceding and following the added or deleted residues are adjusted to adopt the insertion and deletion, and the structure of the loop is refined by molecular mechanics and dynamics tools such as simulated annealing and energy minimization.

Before the modeling of CDR-H3, the models of the heavy and light chain variable domains are combined in order to provide the more authentic structural environment for the modeling of CDR-H3, which generally is cradled by the interface of the domains. For combining the light and heavy chain models, they are superimposed on the structure of an existing Fv fragment by a least-square fit of the Cα-atoms of the conserved residues in the domain interface[27] or in the whole variable domain.[20] The reference structure is most commonly one of the templates used for the modeling of either the heavy or light chain variable domain, however, other structures showing high overall sequence similarity (over both V_L and V_H) can also be considered to orientate the domains. Alternatively, an average β-barrel generated from the variable domains of the known antibody structures can be used as a reference.[16] The relative orientation of V_H and V_L differ among the known antibodies, and significant changes in the orientation can sometimes occur upon antigen binding.[28] The orientation of the domain influences on the combining site structure by determining the relative positions of the light and heavy chain CDRs, and evaluation of several alternative orientations can be useful if high accuracy is required. Apart from plausible stereochemistry, there is, unfortunately, not an unambiguous way to evaluate the tenability of the orientation.

The major challenge in antibody modeling is the building of the CDR-H3 loop. This loop is characterized by extensive variability, and it has not been possible to define canonical classification for this loop. Modeling of CDR-H3 is, however, facilitated by a set of sequence-based rules described for the conformation of the stem of the loop, in particular its C-terminal part. Shirai *et al.*[29] first reported that the conformation of the C-terminal part of CDR-H3 can be divided in two subclasses, "kinked" and "extended," and the results of this study have later been verified and complemented in several other studies.[30–34] Whenever possible, the stem region of the CDR-H3 is grafted from the antibody structure having the stem of same subclass, and preferably, of similar sequence. The stem residues are fitted to the model in the same way as other grafted CDRs. The modeling of apex CDR-H3, or the whole loop if the stems do not fit within the known rules, is based on the same techniques as the modeling of any other non-canonical loop, i.e. on database searches, use of conformation search algorithms or combinations of these two. In practice, the modeling can be considered to be feasible in the case of short and mid-sized loops (up to about 13 residues). There are also data concerning preferential conformations of the apices among such loops.[30,34] The modeling of the apex of CDR-H3 becomes increasingly difficult and highly unreliable when the loop length further increases.

16.2.3 *Side Chain Modeling and Refining of the Model*

Modeling of side-chains can be either based on knowledge-based methods, on the use of computational algorithms, or on a combination of these two approaches. The knowledge-based approach is commonly used in the case of framework residues (side-chains built simultaneously with the main chain) as well as many side-chains in CDRs with known canonical structure. In the procedure described by Morea *et al.*,[20] the conformation of the side-chain in the model is retained if the same residue is present in the corresponding position of the template. If the residue type does not match in the model and the template, the side-chain is built in the conformation existing in the most similar antibody having the same residue in the corresponding position. The backbone conformation affects the rotamer preferences of side-chains, and consequently, the conformations of the CDR residues should be taken from CDRs with the same canonical structure. Side-chains of other residues can be built by selecting the sterically and/or energetically most plausible conformation from a rotamer library or by retaining the relative length of the side-chain as far as permitted. The recent data showing that the side-chain chi-1 angles of certain residues in canonical CDRs, as well as in the stem of CDR-H3 belonging to a kinked subclass, are conserved[34] can be utilized in modeling these residues. Computational search methods, based either on iterative searches, e.g. by CONGEN[25] or simultaneous global optimization[35] of side-chain conformations, can be used to build the side-chains of the residues in non-canonical CDR loops, and these methods can also be used to model side-chains of residues in canonical CDRs, especially in cases of unusual residues. If the main chain of the loop is generated by the conformational sampling algorithm, the side-chains are often produced simultaneously. Once all the side-chains are introduced, the model should be analyzed for stereochemistry, and severe sterical clashes, if any, should be removed by rebuilding the relevant side-chains. Alternatively, global optimization of all the side-chains can be performed after model building and initial energy refinement.[36]

Energy minimization can be used to refine the model by removing non-allowed torsion angels and sterical clashes between atoms that are introduced during the model building process. In order to avoid the introduction of additional distortions, it is useful to perform the minimization by using a stepwise procedure. After initial minimization of the hydrogen atoms, the splice sites between the segments originating from different structures are minimized (only the atoms of the two residues forming a splice site are allowed to move). In the next phase, only the side-chains are free to move, then also the backbone atoms of CDRs, and finally, all the atom in the model are released.

Considering the current level of interest in antibody research as well as the fact that there are a virtually innumerable number of potential target structures to be determined, it is not surprising that fully automated tools for building antibody modeling have been developed. These tools including program ABGEN[37] and WAM — Web antibody modeling server[16] provide rapid access to the 3D-structure of the antibody of interest. The benefit of the construction of the model "manually," in turn, is that good insight to the characteristics of the model is obtained in the course of the modeling process. Internet resources for antibody engineering and modeling are collected in Table 16.1.

16.3 Modeling of the Antibody-Antigen Complexes

16.3.1 *Classifying the Antibody Binding Sites*

Based on the contact analysis and structural studies of antibody binding sites,[38–40] it has been suggested that the shape of the antigen-antibody combining site is correlated with the nature of an antigen. By comparing the overall shape of the binding site of an antibody, it has been possible to categorize the antibodies into the classes, such as concave, ridged, and planar. Comparisons of the free and bound forms of antibody-antigen complexes determined experimentally[41] have revealed changes in the antibody binding sites upon ligand binding. The changes in structure may vary from small side-chain rearrangements[42] to substantial displacements, especially the conformation and localization of the CDR-H3.[43,44] Rearrangements of the hypervariable loop conformations contribute to complex stabilization. The complementarity may be further enhanced by additional structural water molecules at the interface of the complex as in antigen-bound Fab of HyHEL-63.[45]

16.3.2 *Construction of Models*

In the absence of the crystal structure of the antibody-antigen complex, modeling provides a rapid alternative to explore molecular interactions. The modeling of the structures of molecular complexes by using a homology model as a starting point is, however, severely complicated by the limited accuracy of homology models. One possible way to compensate for this is to generate several models, which differ in particular in the regions where structural inaccuracies are most likely to occur, i.e. the non-canonical CDRs and the CDR-H3 or in the relative orientation of the V_H and V_L domains.

By using a known antibody with high sequence identity as a template, it is possible to construct models of reasonable level of quality, following which complex modeling becomes meaningful. Soft docking algorithms/procedures[46–48] or long trajectory molecular dynamics simulations[49] can be used to map the conformational space of the ligand or both the ligand and the binding site of the antibody. The likelihood of the complex structure is then evaluated by different scoring functions, either physics-based potentials (force field type) or statistical potentials based on structural databases derived from protein structures in the PDB.[20] In general, the modeling of antibody-antigen interactions becomes increasingly difficult in the case of large and complex antigens such as proteins or in the case of very flexible molecules. In addition, potential conformational changes of the antibody occurring upon binding of the antigen make the modeling of antibody complexes very challenging. The prediction of binding or binding modes of antigen(s) by docking are often used for hapten or peptide antigen-antibody complexes, as, for instance, described by Stigler *et al.*[50] However, studies for larger protein antigens have also been published[51] where rigid body docking by ZDOCK[52,53] is first performed to create the starting structure, which in turn is refined by molecular dynamics simulations using GROMACS[54] in explicit solvent. Through ligand docking or using long trajectory molecular dynamics simulations, it has been possible to obtain specific knowledge about interacting residues of the protein ligand complex to engineer the specificity of an antibody[7,49,55] or interpret the experimental results.

16.4 Utilization of Antibody Modeling

The ideal starting point for antibody engineering is a crystal structure.[56] However, the amount of crystal structures of antibodies lags far behind the extensive sequence data that can be used to build homology models. Antibody modeling can provide useful information about the binding site and the interactions with the antigen[49,51,57–60] and also with related, possibly cross-reacting molecules.[61–62] Antibody models have been applied in adjusting the binding properties, affinity,[55,63–67] and specificity[61,63,64,68,69] of the recombinant antibody fragments, as well as in interpreting the effects of mutations obtained by random approaches.[70] Model-based design of mutations ranges from targeting single residues to the areas to be randomized.

Stability is an important requirement for all antibody fragments. It varies among different antibodies and the variable domains are mostly responsible for the stability differences between antibodies from the same subclass.[71,72] Targeted mutations, structure-based framework engineering, and CDR-grafting to more stable frameworks have been used to improve the stability of antibodies.[73–77] Sequence analyses and homology modeling have been applied to identify destabilizing residues[78] or to confirm how exposed they are in antibody structures.[79] Different aspects of stability engineering of single chain antibody fragments (scFv) have been reviewed by Wörn and Plückthun,[80] and approaches for stability engineering of variable domains have been collected by Monsellier and Bedoulle.[81]

Human antibodies have proven to have potential in therapy but are difficult to produce. On the other hand, the more easily available monoclonal antibodies of rodent origin have limited use in human therapy due to their immunogenicity and non-optimal effector functions. In addition, the binding properties, stability, internalization efficiency, solubility, or folding kinetics may need to be tailored. Model-based humanization of mouse antibodies has been used to overcome these problems.[82–84] In tandem with the modeling of the variable regions of the mouse antibody, the human frameworks are selected. The design of humanized variable regions includes the identification of those framework residues that may contribute to the binding properties and cause immunogenicity. For recent reviews of the design and engineering of therapeutic antibodies, see Presta[85] and Carter.[86]

In addition to the above-mentioned major efforts to improve antibodies, some of their other properties have been modified by the application of computational antibody engineering methods. Targeted and knowledge-based engineering have been applied to study and improve the crystallization properties of antibody fragments.[87,88] Structural information from a model has been used to design mutations to a catalytic antibody and improve the catalysis[89] as well as to identify important residues in metal-chelate recognition of antibodies for radiotherapy and imaging.[90]

Design of synthetic antibody libraries on single or several frameworks has been based on the crystal structure or homology models.[91–94] In the construction of fully synthetic human combinatorial antibody library (HuCAL), an example of unfocussed library, the structural alignment of approximately 100 antibodies was performed. Homology models were used to obtain insight to framework properties, packing, and the conformations of CDRs. Unfavorable torsion angles, unusually exposed hydrophobic regions, sets of hydrogen bonds, and canonical structures were checked as described in more detail by Knappik *et al.*[93] As an example of focused library, homology modeling was applied to design an antigen-specific library with biased specificity to haptens.[94] According to the model, the residues lining the binding cavity as well as the residues potentially binding the carrier protein were identified prior to targeted mutagenesis. Parallel screenings with a conventional antibody library showed that antibodies obtained from this cavity optimized library bound soluble hapten whereas the antibodies from the conventional antibody library showed carrier protein dependence.

16.5 Future Outlook

In the future, recombinant antibody libraries will take an increasingly prominent role as a source of new antibody molecules. As the recombinant antibodies provide direct access to primary structure, there will most likely be increasing interest to use molecular modeling tools to systematically characterize the antibodies at the level of their tertiary structure. The development of synthetic antibody libraries based on only one or a few defined framework genes will further facilitate the modeling; once the crystal structure of the framework(s) used in the library is(are) determined, a template with very high sequence identity will always be available.

The primary challenge in antibody modeling is the prediction of the conformations of the CDR loops, future improvements in the loop modeling techniques will therefore be of great value to antibody modeling. The constantly increasing number of experimentally determined antibody structures will most likely allow the identification of new, more rarely, occurring canonical structures for CDRs. With a large enough dataset, improvements in knowledge-based modeling of CDR-H3 can be expected, at least in the case of short and midsized loops. There will always be outliers among CDR structures that cannot be predicted by rules, and improvements in the conformational search methods are needed. Current, *ab initio* loop generation algorithms as well as advanced loop databases[95] seem to allow efficient sampling of the conformational space in most cases, but the bottleneck seems to be unreliable ranking of search results. Improved algorithms for the evaluation of the quality of loop candidates would be very valuable for antibody modeling.

References

1. Hoogenboom HR. (2005) Selecting and screening recombinant antibody libraries. *Nat Biotechnol* **23**: 1105–1116.

2. Poljak RJ, Amzel LM, Avey HP, *et al.* (1973) Three-dimensional structure of the Fab' fragment of a human immunoglobulin at 2.8-Å resolution. *Proc Natl Acad Sci* **70**: 3305–3310.
3. Kabat EA, Wu TT. (1971) Attempts to locate complementarity-determining residues in the variable positions of light and heavy chains. *Ann NY Acad Sci* **190**: 382–393.
4. Padlan EA, Abergel C, Tipper JP. (1995) Identification of specificity — determining residues in antibodies. *FASEB J* **9**: 133–139.
5. Lamminmäki U, Pauperio S, Westerlund-Karlsson A, *et al.* (1999) Expanding the conformational diversity by random insertions to CDRH2 results in improved anti-estradiol antibodies. *J Mol Biol* **291**: 589–602.
6. Daugherty PS, Chen G, Olsen MJ, Iverson BL, Georgiou G. (1998) Affinity maturation using bacterial surface display. *Protein Eng* **11**: 825–832.
7. Kusharyoto W, Pleiss J, Bachmann TT, Schmid RD. (2002) Mapping of a hapten binding site: molecular modeling and site-directed mutagenesis study of an anti-atrazine antibody. *Protein Eng* **15**: 233–241.
8. deHaard HJ, Kazemier B, van der Bnet A, *et al.* (1998) Absolute conservation of residue 6 of immunoglobulin heavy chain variable regions of class IIA is required for correct folding. *Protein Eng* **11**: 1267–1276.
9. Sheriff S, Silverton EW, Padlan EA, *et al.* (1987) Three-dimensional structure of an antibody-antigen complex. *Proc Natl Acad Sci* **84**: 8075–8079.
10. Wedemayer GJ, Wang LH, Patten PA, Schultz PG, Stevens RC. (1997) Crystal structures of the free and liganded form of an esterolytic catalytic antibody. *J Mol Biol* **268**: 390–400.
11. Chothia C, Lesk AM. (1987) Canonical structure for the hypervariable regions of immunoglobulins. *J Mol Biol* **196**: 901–917.
12. Chothia C, Lesk AM, Tramontano A, *et al.* (1989) Conformations of immunoglobulin hypervariable regions. *Nature* **342**: 877–883.
13. Martin ACR, Thornton JM. (1996) Structural families in loops of homologous proteins: automatic classification, modeling, and applications to antibodies. *J Mol Biol* **263**: 800–815.
14. Al-Lazikani B, Lesk AM, Chothia C. (1997) Standard conformations for the canonical stuctures of immunoglobulins. *J Mol Biol* **273**: 927–948.
15. Vargas-Madrazo E, Paz-Garcia E. (2002) Modifications to canonical structure sequence patterns analysis for L1 and L3. *Proteins: Struct Funct Genet* **47**: 250–254
16. Whitelegg NR, Rees AR. (2000) WAM: an improved algorithm for modeling antibodies on the WEB. *Protein Eng* **13**: 819–824.
17. Kabat EA, Wu TT. (1972) Construction of a three-dimensional model of the polypeptide backbone of the variable region of kappa immunoglobulin light chains. *Proc Natl Acad Sci USA* **69**: 960–964.

18. Padlan EA, Davies DR, Pecht I, Givol D, Wright C. (1976) Model-building studies of antigen-binding sites: the hapten-binding site of mopc-315. *Cold Spring Harb Symp Quant Biol* **41**: 627–637.
19. Berman HM, Westbrook J, Feng Z, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**: 235–242.
20. Morea V, Lesk AM, Tramontano A. (2000) Antibody modeling: implications for engineering and design. *Methods* *2*0: 267–279.
21. Eigenbrot C, Randal M, Presta L, Carter P, Kossiakoff AA, (1993) X-ray structures of the antigen-binding domains from three variants of humanized anti-p185^{HER2} antibody 4D5 and comparison with molecular modeling. *J Mol Biol* **229**: 969–995.
22. Honegger, A, Plückthun A. (2001) Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol* **309**: 657–670
23. Kabat EA, Wu TT, Perry HM, Gottesmann KS, Foeller C. (1991). Sequences of Proteins of Immunological Interest, 5th edn. NIH Publication No. 91-3242 US Department of Health and Human Services.
24. Lefranc MP, Giudicelli V, Ginestoux C, *et al.* (1999). IMGT, The International ImMunoGeneTics database. *Nucl Acids Res* **27**: 209–212.
25. Bruccoleri RE, Karplus M. (1987) Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **26**: 137–168.
26. Mandal C, Kingery BD, Anchin JM, Subramaniam S, Linthicum DS. (1996) ABGEN: a knowledge-based automated approach for antibody structure modeling. *Nat Biotechnol* **14:** 323–328.
27. Chothia C, Novotný J, Bruccoleri R, Karplus M. (1985) Domain association in immunoglobulin molecules. The packing of variable domains. *J Mol Biol* **186**: 651–663.
28. Wilson IA, Stanfield RL. (1994) Antibody-antigen interactions: new structures and new conformational changes. *Curr Opin Struct Biol* **4**: 857–867.
29. Shirai H, Kidera A, Nakamura H. (1996) Structural classification of CDR-H3 in antibodies. *FEBS Lett* **399**: 1–8.
30. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. (1997) Antibody structure, prediction and redesign. *Biophys Chem* **68**: 9–16.
31. Morea V, Tramontano A, Rustici M, Chothia C, Lesk AM. (1998) Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol* **275**: 269–294.
32. Oliva B, Bates PA, Querol E, Avilés FX, Sternberg MJ. (1998) Automated classification of antibody complementarity determining region 3 of the heavy chain (H3) loops into canonical forms and its application to protein structure prediction. *J Mol Biol* **279**: 1193–1210.

33. Shirai H, Kidera A, Nakamura H. (1999) H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Lett* **455**: 188–197.
34. Whitelegg N, Rees AR. (2004) Antibody variable regions: toward a unified modeling method. *Meth Mol Biol* **248**: 51–91.
35. Desmet J, Spriet J, Lasters I. (2002) Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **48**: 31–43.
36. Staelens S, Desmet J, Ngo TH, *et al.* (2006) Humanization by variable domain resurfacing and grafting on a human IgG4, using a new approach for determination of non-human like surface accessible framework residues based on homology modeling of variable domains. *Mol Immunol* **43**: 1243–1257.
37. Mandal C, Kingery BD, Anchin JM, Subramaniam S, Linthicum DS. (1996) ABGEN: a knowledge-based automated approach for antibody structure modeling. *Nat Biotechnol* **14**: 323–328.
38. MacCallum RM, Martin AC, Thornton JM (1996) Antibody-antigen interactions: contact analysis and binding site topography. *J Mol Biol* **262**: 732–745.
39. Lee M, Lloyd P, Zhang X, *et al.* (2006) The shapes of antibody binding sites: qualitative and quantitative analyses based on a geomorphic classification scheme. *J Org Chem* **71**: 5082–5092.
40. Collis AV, Brouwer AP, Martin AC. (2003) Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J Mol Biol* **325**: 337–354.
41. Wilson IA, Stanfield RL, Rini JM, *et al.* (1991) Structural aspects of antibodies and antibody-antigen complexes. *Ciba Found Symp* **159**: 28–39.
42. Parkkinen T, Nevanen TK, Koivula A, Rouvinen J. (2006) Crystal structures of an enantioselective Fab-fragment in free and complex forms. *J Mol Biol* **357**: 471–480.
43. Stanfield RL, Takimoto-Kamimura M, Rini JM, Profy AT, Wilson IA. (1993) Major antigen-induced domain rearrangements in an antibody. *Structure* **1**: 83–93.
44. Monaco-Malbet S, Berthet-Colominas C, Novelli A, *et al.* (2000) Mutual conformational adaptations in antigen and antibody upon complex formation between an Fab and HIV-1 capsid protein p24. *Structure* **8**: 1069–1077.
45. Li Y, Li H, Smith-Gill SJ, Mariuzza RA. (2000) Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63. *Biochemistry* **39**: 6296–6309.
46. Gray JJ, Moughon S, Wang C, *et al.* (2003) Protein-protein docking with simultaneous optimization of rigid body displacement and side-chain conformations. *J Mol Biol* **331**: 281–299.
47. Gabb HA, Jackson RM, Sternberg MJ. (1997) Modelling protein docking using shape complementarity, electrostatistics and chemical information. *J Mol Biol* **272**: 106–120.

48. Li CH, Ma XH, Chen WZ, Wang CX. (2003) A soft docking algorithm for predicting the structure of antibody-antigen complexes. *Proteins* **52**: 47–50.
49. Curcio R, Caflisch A, Paci E. (2007) Change of the unbinding mechanism upon a mutation: a molecular dynamics study of an antibody-hapten complex. *Protein Sci* **14**: 2499–2514.
50. Stigler RD, Hoffman B, Abagyan R, Schneider-Mergener J. (1999) Soft docking an L and D peptide to an anticholera toxin antibody using internal coordinate mechanics. *Structure* 7: 663–670.
51. Autore F, Melchiorre, Kleinjung J, Morgan WD, Fraternali F. (2007) Interactions of malaria parasite-inhibitory antibodies with the merozoite surface protein MSP1(19) by computational docking. *Proteins* **66**: 513–527.
52. Chen R, Weng Z. (2002) Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins* **47**: 281–294.
53. Chen R, Li L, Weng Z. (2003) ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**: 80–87.
54. Berendsen HJC, van der Spoel D, van Drunen R. (1995) GROMACS: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun* **91**: 43–56.
55. Nevanen TK, Hellman ML, Munck N, *et al.* (2003) Model-based mutagenesis to improve the enantioselective fractionation properties of an antibody. *Protein Eng* **16**: 1089–1097.
56. Lippow SM, Wittrup KD, Tidor B. (2007) Computational design of antibody-affinity improvement beyond *in vivo* maturation. *Nature Biotech* **25**: 1171–1176.
57. Kalsi JK, Martin AC, Hirabayashi Y, *et al.* (1996) Functional and modeling studies of the binding of human monoclonal anti-DNA antibodies to DNA. *Mol Immunol* **33**: 471–483.
58. Geva M, Eisenstein M, Addadi L. (2004) Antibody recognition of chiral surfaces. Structural models of antibody complexes with leucine-leucine-tyrosine crystal surfaces. *Proteins: Struct Funct Genet* **55**: 862–873.
59. Kearns-Jonker M, Barteneva N, Mencel R, *et al.* (2007) Use of molecular modeling and site-directed mutagenesis to define the structural basis for the immune response to carbohydrate xenoantigens. *BMC Immunol* **8**: 3.
60. Fenalti G, Hampe CS, O'Connor K, *et al.* (2007) Molecular characterization of a disease associated conformational epitope on GAD65 recognised by a human monoclonal antibody b96.11. *Mol Immunol* **44**: 1178–1189.
61. Korpimäki T, Rosenberg J, Virtanen P, *et al.* (2003) Further improvement of broad specificity hapten recognition with protein engineering. *Protein Eng* **16**: 37–46.
62. Paula S, Monson N, Ball WJ Jr. (2005) Molecular modeling of cardiac glycoside binding by the human sequence monoclonal antibody 1B3. *Proteins: Struct Funct Bioinform* **60**: 382–391.

63. Chames P, Coulon S, Baty D. (1998) Improving the affinity and the fine specificity of an anti-cortisol antibody by parsimonious mutagenesis and phage display. *J Immunol* **161**: 5421–5429.
64. Roberts S, Cheetham JC, Rees AR. (1987) Generation of an antibody with enhanced affinity and specificity for its antigen by protein engineering. *Nature* **328**: 731–734.
65. Ruff-Jamison S, Glenney JR (1993) Molecular modeling and site-directed mutagenesis of an anti-phosphotyrosine antibody predicts the combining site and allows the detection of higher affinity interactions. *Protein Eng* **6**: 661–668.
66. Hemminki A, Niemi S, Hoffren A, *et al.* (1998) Specificity improvement of a recombinant anti-testosterone Fab fragment by CDRIII mutagenesis and phage display selection. *Protein Eng* **11**: 311–319.
67. Casipit CL, Tal R, Wittman V, *et al.* (1998) Improving the binding affinity of an antibody using molecular modeling and site-directed mutagenesis. *Protein Sci* 7: 1671–1680.
68. Iba Y, Hayashi N, Sawada J, Titani K, Kurosawa Y. (1998) Changes in specificity of antibodies against steroid antigens by introduction of mutations into complementarity-determining regions of the V_H domain. *Protein Eng* **11**: 361–370.
69. Dubreul O, Bossus M, Graille M, *et al.* (2005) Fine-tuning of the specificity of an anti-progesterone antibody by first and second sphere residue engineering. *J Biol Chem* **280**: 24880–24887.
70. Luginbühl B, Kanyo Z, Jones RM, *et al.* (2006) Directed evolution of an anti-prion protein scFv fragment to an affinity of 1 pM and its structural interpretation. *J Mol Biol* **363**: 75–97.
71. Ewert S, Huber T, Honegger A, Plückthun A. (2003) Biophysical properties of human antibody variable domains. *J Mol Biol* **325**: 531–553.
72. Ewert S, Honegger A, Plückthun A. (2003) Structure-based improvement of the biophysical properties of immunoglobulin VH domains with a generalizable approach. *Biochemistry* **42**: 1517–1528.
73. Ewert S, Honegger A, Plückthun A. (2004) Stability improvement of antibodies for extracellular and intracellular applications: CDR crafting to stable frameworks and structure-based framework engineering. *Methods* **34**: 184–199.
74. Garber E, Demarest SJ. (2007) A broad range of Fab stabilities within a host of therapeutic IgGs. *Biochem Biophys Res Co* **355**: 751–757.
75. Donini M, Morea V, Desiderio A, *et al.* (2003) Engineering stable cytoplasmic intrabodies with designed specificity. *J Mol Biol* **330**: 323–332.
76. Willuda J, Honegger A, Waibel R, *et al.* (1999) High thermal stability is essential for tumor targeting of antibody fragments: engineering of humanized

anti-epithelial Glycoprotein-2 (epithelial adhesion molecule) single-chain Fv fragment. *Cancer Res* **59**: 5758–5767.

77. Teerinen T, Valjakka J, Rouvinen J, Takkinen K. (2006) Structure-based stability engineering of mouse IgG1 Fab fragment by modifying constant domains. *J Mol Biol* **361**: 687–697.
78. Arndt M, Krauss J, Schwarzenbacher R, *et al.* (2003) Generation of a highly stable, internalizing anti-CD22 single-chain Fv fragment for targeting non-Hodgkin's lymphoma. *Int J Cancer* **107**: 822–829.
79. Chowdhury PS, Vasmatzis G, Beers R, Lee B, Pastan I. (1998) Improved stability and yield of a Fv-toxin fusion protein by computer design and protein engineering of the Fv. *J Mol Biol* **281**: 917–928.
80. Wörn A, Plückthun A. (2001) Stability engineering of antibody single-chain Fv fragments. *J Mol Biol* **305**: 989–1010.
81. Monsellier E, Bedoulle H. (2006) Improving the stability of an antibody variable fragment by a combination of knowledge-based approaches: validation and mechanisms. *J Mol Biol* **362**: 580–593.
82. Li B, Wang H, Zhang D, *et al.* (2007) Construction and characterization of a high-affinity humanized SM5-1 monoclonal antibody. *Biochem Biophys Res Co* **357**: 951–956.
83. Tsurushita N, Hinton PR, Kumar S. (2005) Design of humanized antibodies: from anti-Tac to Zenapak. *Methods* **36**: 69–83.
84. Staelens S, Desmet J, Ngo TH, *et al.* (2006) Humanization by variable domain resurfing and grafting on a human IgG4, using a new approach for determination of non-human like surface accessible framework residues based on homology modeling of variable domains. *Mol Immunol* **43**: 1243–1257.
85. Presta LG. (2005) Selection, design and engineering of therapeutic antibodies. *J Allergy Clin Immunol* **116**: 731–736.
86. Carter PJ. (2006) Potent antibody therapeutics by design. *Nat Rev Immunol* **6**: 343–357.
87. Wingren C, Emundson AB, Borrebaeck CAK. (2003) Designing proteins to crystallize through β-strand pairing. *Protein Eng* **16**: 255–264.
88. Honegger A, Spinelli S, Cambillau C, Plückthun A. (2005) A mutation designed to alter crystal packing permits structural analysis of a tight-binding fluorescein-scFv complex. *Protein Sci* **14**: 2537–2549.
89. Zheng L, Manetsch R, Woggon W-D, Baumann U, Reymond J-L. (2005) Mechanistic study of proton transfer and hysteresis in catalytic antibody 16E7 by site-directed mutagenesis and homology modeling. *Biorg Med Chem* **13**: 1021–1029.
90. Delehanty JB, Jones RM, Bishop TC, Blake DA. (2003) Identification of important residues in metal-chelate recognition by monoclonal antibodies. *Biochemistry* **42**: 14173–14183.

91. Sidhu SS, Li B, Chen Y, *et al.* (2004) Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. *J Mol Biol* **338**: 299–310.
92. Hoet RM, Cohen EH, Kent RB, *et al.* (2005) Generation of high-affinity human antibodies by combining donor-derived and synthetic complementary-determining-region diversity. *Nat Biotechnol* **3**: 344–348.
93. Knappik A, Ge L, Honegger A, *et al.* (2000) Fully synthetic human combinatorial antibody librarries (HuCAL) based on molecular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol* **296**: 57–86.
94. Persson H, Lantto J, Ohlin M. (2006) A focused antibody library for improved hapten recognition. *J Mol Biol* **357**: 607–620.
96. Fernandez-Fuentes N, Oliva B, Fiser A. (2006) A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucl Acids Res* **34**: 2085–2097.

Section III

Drug Discovery and Pharmacology

This page intentionally left blank

Chapter 17

Small Molecule Docking

R. A. Friesner[*,†], M. Repasky[‡] and R. Farid[‡]

17.1 Introduction

In this chapter, we shall discuss computational methods for modeling the interaction of small molecule ligands with protein receptors in aqueous solution. Such interactions form the basis of the mechanism of the great majority of pharmaceutically active compounds. The ability to determine the structures and free energy of binding of protein-ligand complexes is, therefore, the key objective of computational structure-based drug design.

In principle, this problem can be solved by simulation (e.g. molecular dynamics) of the protein and ligand in solution. In practice, there are formidable problems associated with a straightforward, physical chemistry-based approach, including the large amount of CPU time required, and limitations in force field accuracy. These considerations have led to the development of approximate methods that, while still based on physical chemistry principles, endeavor to embody such principles in empirically optimized models ("scoring functions") and determine structures via specially designed conformational search algorithms ("docking algorithms"). The combination of a docking

[*] Corresponding author: Email: rich@chem.columbia.edu.

[†] Department of Chemistry, Columbia University, New York, NY 10025. Email: rich@chem.columbia.edu.

[‡] Schrödinger, LLC, 120 W. 45th Street, New York, NY 10036.

algorithm and a scoring function, implemented in a self-contained software system, constitutes a "docking program." A number of such programs are now widely used in both academia and industry.

The first docking program, DOCK, was developed in the laboratory of Tack Kuntz at the University of California, San Francisco, beginning in the early 1980s.[1] Over the next decade, DOCK incorporated the basic principles common to most reasonably efficient docking programs:

Ligand dihedral angles are varied, while keeping bond lengths and angles fixed. The protein is assumed to be rigid. The rigid protein representation greatly reduces the number of degrees of freedom sampled, enabling development of algorithms for rapid docking of large numbers of ligands. The problem with this approximation, which remains a central challenge of docking methods today, is that it ignores induced fit effects. Induced fit effects arise from changes in the protein structure induced by ligand binding and can be substantial, to the point where known active compounds may completely fail to dock in a rigid receptor of incompatible shape.

Keeping the protein rigid enables the protein structure and interactions to be represented on a grid, thus avoiding the need to explicitly calculate interactions between all protein atoms and ligand atoms at each step of the conformational search. The development of a suitably accurate, grid-based representation of the intermolecular interactions is a nontrivial task that has been addressed in DOCK and other programs, reducing CPU times required for docking by as much as one to two orders of magnitude.

DOCK begins with a set of ligand conformations and uses a variety of conformational sampling algorithms to determine reasonable initial locations for these conformations. Minimization of the ligand in the field of the protein is then employed to assess which starting conformations yield the lowest energy after optimization. Subsequent docking programs have used a wide range of sampling approaches and methods for "pose selection," that is, choosing the optimal structural prediction for the protein-ligand complex. Here, *pose* refers to the complete specification of the ligand, including position and orientation relative to the receptor of a single conformation of the ligand.

Given a predicted protein-ligand complex geometry (in principle, corresponding to the results of an X-ray crystallographic experiment), DOCK uses an empirically optimized scoring function to predict protein-ligand binding affinity. While the details of subsequent empirical scoring functions differ, all docking programs attempt to infer binding affinity from the geometrical relationship between the protein and ligand in the docked pose selected by the program (as described in the above paragraph), incorporating elements such as hydrogen bonding between protein and ligand, hydrophobic interactions, formation of salt bridges, and restriction of the dihedral angle freedom in the ligand by the protein. Parameters of the scoring function are typically fit to experimental structural and binding affinity data, the quantity of which has increased exponentially over the past 15 years.

The vast increase in the number of publicly available protein-ligand complexes in the Protein Data Bank (PDB)[2] has made it possible to extensively test and optimize docking programs in a fashion that was simply not possible 15 years ago, when there were only about 900 crystal structures in the PDB, many of which had no co-crystallized ligands. In contrast, there are, as of July 2007, over 44 000 crystal structures in the PDB of which about 36 000 are X-ray structures of proteins, and of those, nearly 28 000 contain a ligand. The development of curated datasets, including experimental binding affinities, such as PDBBind[3] and MOAD,[4] further simplifies the development of training and test sets.

Currently available scoring functions, docking algorithms, and their accuracy in lead discovery will be discussed in Section 17.2. Before doing so, however, a fundamental examination of the validity of assumptions underlying docking algorithms and scoring functions, summarized above, is useful to provide a context in which docking and scoring as a basic research enterprise can be understood. A brief outline of the key issues is presented in the paragraphs that follow.

The first and most obvious issue is the validity of the rigid receptor hypothesis. Empirical data, accumulated over the past decade of docking efforts with many programs in many laboratories, suggests that roughly 30–50% of a diverse set of known active compounds will fail to dock in anything resembling the correct binding mode, due to steric clashes, in a typical receptor. The percentage may be much lower if one insists on a truly accurate binding mode, in which all experimentally observed hydrogen bonds and hydrophobic interactions are formed as is likely necessary (but not sufficient) for reliably rank-ordering compounds by binding affinity. Hence, it is clear that overcoming limitations of the rigid receptor hypothesis is a crucial objective in turning docking methodology into a true platform for drug discovery.

Assuming a "good enough" receptor conformation into which one can dock, achievable either via induced fit algorithms or by using an ensemble of initial structures, the next question is what types of computational methods and models are necessary and sufficient to predict the binding mode of the ligand-receptor complex with precision and reliability. Studies of "self-docking," i.e. redocking a ligand into its native cognate conformation of the receptor, as determined by crystallography, provide an initial assessment of the effectiveness of various algorithms and model assumptions.

The problem of estimating the binding free energy of a protein-ligand complex is extraordinarily complex. Compared to a fully-rigorous, statistical mechanical treatment, the use of an empirical scoring function and a single conformation of the protein-ligand complex requires significant approximations. These approximations may place a fundamental limit on the accuracy of binding free energy predictions by empirical scoring functions. Theoretical arguments can be made for inferring the binding affinity of the complex from the single, dominant structure observed in a crystallographic experiment, which is presumably the target in a docking experiment. If a wide range of ligand conformations is sampled in the bound state, presumably the resulting crystallographic data would not be resolved to a suitable degree of precision. Generally, poorly resolved substructures of a ligand lie in solution, exhibiting little interaction with the receptor; these substructures are unlikely to make a significant contribution to the binding affinity. The crystallographic conformation is generally obtained at temperatures lower than room temperature, and indeed a criticism of empirical scoring functions could be based on the premise that they are in effect a low temperature theory. However, excitations of the complex from low temperature (at least as judged by molecular dynamics simulations) are generally small enough such that development of generic approximations for such excitations does not seem unreasonable.

However, even if this argument is correct, it remains the case that a scoring function that does not embody the physics of binding, regardless of the volume of data used in the parameterization, is unlikely to be successful when applied to ligands or targets that differ from those in the training set. The driving force for protein-ligand binding is dominated by displacement of water molecules in the protein cavity by a ligand that is complementary to the protein groups surrounding it. This is the case because water molecules in the active site typically have unfavorable free energies relative to bulk water molecules. Other important factors such as entropy loss of the protein and ligand upon binding and burial of protein or ligand charges must also be considered. For many years, relatively simplistic functional forms in the scoring function have implicitly represented the free energy contribution of displacement of water molecules, as well as the other factors enumerated above. A key question, which will be considered in greater detail below, is whether these simplistic functional forms embody realistic physics, and whether improved models can be designed that lead to better agreement with experimental data.

A final issue of central importance in empirical scoring function development and analysis is that of ligand and protein strain energy, also referred to as the energy required for conformer focusing.[5] The thermodynamic cycle in Fig. 17.1 illustrates the contributions of protein and ligand strain energies to the free energy of binding. One great advantage of an empirical scoring function is that rather than computing a binding affinity (typically a small number) by subtracting two large numbers, as in MM/MD-PBSA[6] or MM/MD-GBSA[7] calculations, the binding affinity (i.e. the small number) is calculated directly. This advantage can compensate for many of the approximations inherent in the methodology; the noise in brute force binding calculations of protein-ligand complexes can be quite large, even if the basic description of the interactions is reasonable. This noise leads to the non-intuitive finding that MM-GBSA/PBSA methods that utilize a single, relaxed complex structure tend to perform better than MD-GBSA/PBSA methods that utilize a more rigorous ensemble of complex structures.[8] However, it is important to realize that such a scoring

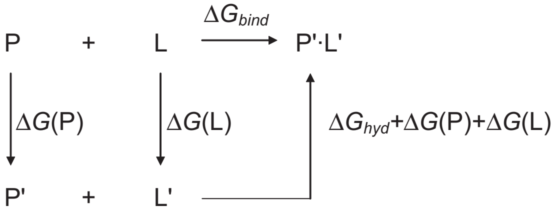


Fig. 17.1 Thermodynamic cycle illustrating the effect of ligand and protein strain energy on the free energy of binding. Here, ΔG(P) and ΔG(L) are the energies required to take the protein and ligand, respectively, from their solution conformation distributions (P and L) into a distribution of generally more strained conformations (P′ and L′) suitable for binding.

function only evaluates the interaction energy and does *not* include the conformational changes in ligand and receptor required to form the complex. Unfortunately, scoring functions typically only account for the interaction energy, ignoring the protein/ligand strain energies.

In principle, both protein and ligand strain energy can be calculated using physical chemistry-based approaches such as thermodynamic integration/free energy perturbation theory, continuum solvent modeling, etc. In practice, there is virtually no data in the literature calibrating the accuracy of such first principles calculations for proteins. Ligand strain energies from force field-based calculations have been shown to manifest significant errors in many cases.[5] An alternative is to rely on experimental data and on analogies between conformational changes of related receptors. Even a crude approximation is better than none at all from the standpoint of putting experimental binding affinities for different ligands and different receptors on an equal footing. Both approaches will be discussed in greater detail below.

17.2 Rigid Receptor Docking

17.2.1 *Docking Programs*

The DOCK program was the first of its kind, and continues to be used in both academia and industry. However, over the past decade, a number of new programs have been developed, incorporating efforts to improve sampling algorithms, pose selection functions, and the scoring function for binding affinity. By our estimation, the FlexX,[9,10] GOLD,[11,12] and Glide[13–15] programs account for greater than 90% of the docking carried out in ongoing drug discovery projects in pharmaceutical and biotechnology companies, and hence, represent a good snapshot of current state of the art practice. Quite a few other programs are in use, including Surflex,[16,17] MolDock,[18] LigandFit,[19] eHiTS,[20] ICM,[21] AutoDock,[22] and FRED (Open Eye Scientific Software; Santa Fe, NM); however, due to space limitations, we will not discuss these programs. Benchmarking of new methodologies is essential to continued progress and many such studies have been performed.[23] While each method has its strengths and weaknesses, to our knowledge, none of the alternatives listed above has demonstrated significant superiority across a wide range of test cases in pose prediction or binding affinity prediction to the current versions of GOLD, FlexX, and Glide. We focus in the present section on the docking algorithms and their prediction of protein-ligand complex geometry. A discussion of binding affinity prediction is presented in Section 17.4.

The FlexX program, developed in the early to middle 1990s by Lenguaer, Klebe, and coworkers, performs flexible docking via an incremental construction algorithm. The program first searches for locations where small molecular fragments can be favorably positioned. Starting from these "base" fragments, the remainder of the molecule is grown incrementally, taking torsional flexibility into account. As the molecule is grown, only the highest scoring structures are retained, and then clustering is used to eliminate redundant structures. An empirically derived pose selection function is then used to rank order the surviving solutions generated from the ensemble of base fragments via the growing algorithm.

The GOLD program was developed in a similar time frame through collaboration between the University of Sheffield, GlaxoSmithKline, and the Cambridge Crystallographic Data Center. In contrast to FlexX, GOLD uses a genetic algorithm to locate docking solutions, propagating multiple copies of a flexible model of the ligand in the active site of the receptor, and recombining segments of these copies randomly until a converged ensemble of structures is generated. Pose selection is accomplished using a molecular mechanics-based function, incorporating hydrogen bonding energy, steric interactions between protein and ligand via a modified 4–8 van der Waals term, which is significantly softened compared to the usual 6–12 Lennard-Jones model, and ligand internal energy is modeled using a molecular mechanics potential function.

The Glide program, developed in collaboration between the Friesner laboratory and Schrödinger, Inc., was created in a somewhat later time period (beginning in the late 1990s). Glide SP (standard precision) employs yet a third strategy for sampling the ligand position and conformation, namely a hierarchical search methodology that in principle is exhaustive. This algorithm is probably the closest to the original strategy employed in DOCK, although it is very different in its details. An ensemble of ligand conformations is pregenerated using a filter designed to favor the more open structures typically found when ligands are bound to protein cavities. These conformations are then screened by a series of increasingly demanding filters to locate initial guesses for ligand poses. Finally, the initial poses are refined via minimization and torsional sampling to yield a final ensemble of poses. A high-accuracy version of the algorithm, Glide XP (extra precision), has been developed that contains extensive modifications to SP in both the sampling algorithm and scoring function. The XP sampling algorithm begins with SP docking, but then refines the poses by extracting base fragments and executing a high resolution growing algorithm, similar in some ways to FlexX. As with GOLD and FlexX, a specialized pose selection function has been developed for Glide SP and XP, in this case, principally based on the OPLS-AA[24] molecular mechanics force field combined with relatively small admixtures of other terms. There is a mechanism for softening the van der Waals interaction by scaling the parameters, so as to accommodate minor steric clashes due to inaccuracies in the crystal structures and small induced fit effects.

17.2.2 *Docking Accuracy: Self-Docking*

The accuracy of a docking program can be assessed by comparing the predicted structures of protein-ligand complexes with crystallographic data. For rigid docking, the simplest comparisons are made by docking each ligand into its own cognate receptor conformation; this is referred to as *self-docking*.

Average ligand RMSDs to the crystal structure of the best scoring prediction from self-docking experiments with FlexX, GOLD, and Glide are summarized in Ref. 13. The FlexX scoring function is relatively "soft," and hence, has some difficulty in recovering the native structure even in self-docking. The developers of FlexX instead emphasize examining a larger set of docked poses and employing such things as protein-ligand constraints and visual inspection to select the correct pose. This sort of human intervention, while on occasion essential, is quite labor intensive and would be difficult to apply to a large number of molecules. Both GOLD and Glide perform reasonably well on the indicated test sets.

It should be noted that these results are dependent on a few factors, including the experience of the individual performing the experiments, preparation of the protein and ligand structures prior to docking, and the fact that the average RMSD, used as a measure of docking accuracy, can be adversely affected by a few ligands with very poor placements. Perhaps more interesting than the RMSD comparisons is the question of why docking programs have consistently displayed a residual fraction of complexes, typically 20–30% of the data set, for which self-docking RMSDs are greater than 2.0 Å. We have been investigating such cases in detail over the past several years and have identified a few factors that appear to be significant:

In roughly 10–15% of the cases, the use of polarized charges on the ligand, derived from mixed quantum mechanics/molecular mechanics (QM/MM) calculations in which the ligand charges respond to the protein environment, have been shown to yield a selection of low RMSD pose in preference to an incorrect pose with the wrong hydrogen-bonding pattern. Specific examples can be found in Ref. 25.

In a smaller fraction of cases (~3–5%), lack of flexibility in saturated ring structures has been identified as a barrier to achieving accurate structural prediction. In these cases, the saturated ring typically has large, bulky groups attached to a number of points on the ring. Docking programs in many cases sample grossly different ring conformations (chair, boat), but not perturbations of the angles in the ring. Such perturbations, which have relatively low energetic costs, are necessary in a small fraction of cases to properly position large attached groups in available pockets. Similar phenomena can also be observed in bond angles outside rings, which can occasionally display significant lever arm effects due to long/bulky attachments. The introduction of a limited degree of flexibility in angle bending is therefore able to drastically improve RMSDs in a small, but noticeable, number of self-docking cases, particularly when "unbiased" input ligand geometries are used that differ in detail from the native co-crystallized geometry.

Metal-containing systems can pose particular challenges to docking algorithms and pose selection functions. The description of ligand-metal interactions via classical force fields is of questionable accuracy, and errors in such descriptions can lead to structural ambiguity. This issue requires significant further exploration.

When the ligand has a large number of rotatable bonds, the sampling problem becomes more difficult, and sampling errors, as opposed to energetic errors in the pose selection function, are observed more frequently. This issue can in principle be addressed by applying additional computational effort when the ligand size exceeds a given threshold, which seems to occur at about ~15–20 rotatable bonds.

Some cases of "misdocking" are undoubtedly due to problems external to the docking methodology, for example, errors in the protein structure or misassignment of ionization/tautomeric states. Others are attributable to ligand groups that extend into solution in the complex and do not make significant contact with the receptor. For such groups, B-factors in the X-ray structure may be high or the energetic effects may be too subtle for current empirical energy models to discriminate.

Our most recent results, using development versions of Glide with polarized QM/MM charges on the ligand, produce poses with RMSD < 2.0 Å in about 90–95% of self-docking experiments cases by addressing the above factors with a reasonable level of effort. At this point, we believe issues due to cross-docking errors (presumably arising from induced fit effects) and errors in binding free energy predictions are greater contributors to docking/scoring failures in practical application than self-docking errors.

17.2.3 *Docking accuracy: Cross-docking*

A more complex and demanding test than self-docking involves docking a ligand into a conformation of the receptor that is *not* its cognate, but rather either the apo structure of the receptor or one from co-crystallization with a different ligand. In many such cases, only approaches utilizing receptor flexibility to treat induced fit, discussed below in Section 17.3, can lead to high-quality pose prediction. When two protein conformations are sufficiently similar, cross docking can generate reasonable ligand poses, albeit, usually not quite as accurate in details as obtained from self-docking. Cross-docking can also introduce noise into the calculations with the effect of creating smaller energy gaps between experimentally observed and incorrect docking solutions, introducing an energetic preference for an incorrect solution, and/or making barriers in sampling more difficult to overcome, typically because the ligand is a "tighter fit" into the cross-docked structure than into the self-docked structure.

It is relatively straightforward to set up cross-docking tests for rigid receptor docking. Given a set of superimposed co-crystallized complexes of a given receptor, one simply takes all of the ligands and attempts to dock them into all or some fraction of the receptor conformations, and records the ligand RMSDs in a cross-docking "matrix." In practice, it is more common to dock a database of ligands into multiple receptor conformations, a closely related methodology referred to as ensemble docking. A number of publications performing ensemble or cross-docking studies are available in the literature.[26–31]

There is some difference in cross-docking performance among programs, although this difference is typically smaller than what is seen in self-docking. Such a result is not surprising; ligands that fit poorly into a given receptor conformation (i.e. exhibit significant steric clashes) cannot be docked by any rigid docking program, no matter how sophisticated the docking algorithm. The degree of "hardness" of the pose selection function is also a highly relevant factor in this type of calculation; softer potentials may enable a higher fraction of cross-docking calculations to succeed, although they may also prove detrimental in enrichment studies by enabling false positives to achieve good scores.

However, the principal conclusion to be drawn from the cross-docking results is that docking failures due to cross-docking effects are a substantial cause of poor performance of rigid docking programs in enrichment, rank ordering, and other key tasks. If one believes accurate structures are required for high-quality binding free energy prediction, and if 30–50% of actives are scored incorrectly due to poor binding mode prediction, the impact on any reasonable performance metric is going to be severe. Hence, if docking is to become a true platform for driving structure-based drug design, it is imperative that receptor flexibility be addressed in an accurate, yet cost effective fashion.

The simplest approach to receptor flexibility is ensemble docking, where ligands are docked into multiple receptor conformations. In the naïve implementation of this strategy, the computational cost of docking into *N* conformations is simply *N*-times the cost of docking into one conformation. More sophisticated approaches can reduce the CPU time per conformation.[32–36] However, the most critical question is how much improvement such an approach can provide in terms of docking accuracy and corresponding binding free energy prediction. This question is just beginning to be explored in the literature. There are some major technical issues that must be addressed before performance can be rigorously evaluated. One of the most important is how one selects the correct binding mode, and binding affinity prediction, from among the ensemble. For example, a simple approach would be to choose the binding mode with the best predicted binding affinity. However, such an approach will often favor more "open" forms of the receptor, which can result in significant over-prediction if the ligand in fact should fit into a tighter pocket. Opening the receptor generally costs reorganization energy; if the ligand is large, and otherwise could not fit, this reorganization energy is a requisite penalty for achieving any sort of reasonable docked pose. Unless an estimate of this reorganization energy is included in the pose selection and scoring functions, however, an over-prediction of the binding affinity will result. We refer readers to Refs. 37 and 38 for other perspectives on ensemble docking.

An issue with ensemble docking is that one cannot proliferate receptor conformations indefinitely. If the conformations are obtained from crystallography, there will typically be a limited number of alternatives available, particularly in the early stages of a discovery project when accurate docking and scoring results would be most valuable. Generation of conformations from high-resolution protein structure prediction methods (such as Schrödinger's Prime program) is possible in principle, but studies have not yet been done to validate the conformations that would be so obtained. If a relatively small number of conformations[3–5] are used for ensemble docking, a significant number of misdockings are still observed, and the problem of obtaining accurate scores from the ensemble remains. Hence, while ensemble docking will almost certainly be a part of any long-term solution, particularly when large loop motions of the receptor are involved, for example, in studying ligands that bind to the DFG-in and DFG-out forms of p38 MAP kinase,[39] it is unlikely to solve the cross-docking problem completely. This is particularly true if the goal is to rank order compounds as opposed to approximately separate active and inactive compounds. To achieve greater robustness and higher accuracy, introduction of protein motion explicitly into the calculation is required. Such methods, which we shall refer to as induced fit approaches, are the subject of the next section.

17.3 Flexible Receptor Docking: Treatment of Induced Fit Effects

In principle, flexible treatment of both ligand and receptor to produce an accurate prediction of the structure of a ligand-receptor complex is a straightforward problem in biomolecular simulation. One could, for example, use molecular dynamics (MD) methods with explicit solvation, starting with the ligand in solution, and under appropriate conditions with regard to effective ligand concentration, a sufficiently long MD simulation should converge to the bound state as thermodynamic equilibrium is reached. In practice, such an approach would require far too much computation time to be practical.

A more efficient approach is to use rigid receptor docking methods to generate many initial poses, for which at least one is close to the correct structure, and then apply simulation methods that account for both ligand and receptor flexibility to generate a final ensemble of structures, selecting the correct answer via a scoring function, typically based on the molecular mechanics energy of the complex. A number of methodologies of this type have been reported in the literature.[32,40] A faster, but arguably less powerful alternative is to enable a small number of protein groups, typically protein side-chains, to move within the docking procedure itself.[41–44] The effectiveness of any of these approaches depends critically upon details of the sampling technology and energy model. Below, we briefly outline the key issues and summarize the current state-of-the-art of our induced fit methodology,[40] which combines the Glide and Prime programs.

The induced fit docking approach involves docking followed by a limited protein conformation search. A key requirement in this approach is the generation of an initial pose close to the experimental pose by the docking algorithm. If structural incompatibility between ligand and receptor conformation is relatively minor, standard docking protocols may be sufficient to generate a ligand pose within 2–3 Å RMSD, although it may not be the top ranked pose in the initial docking run. As the incompatibility increases, larger steric clashes would be required to position the ligand in the receptor consistent with crystal structures, and softening of the potential in some fashion is required to generate a reasonable initial guess. The simplest approach is to scale the protein-ligand van der Waals potential so that it is more forgiving of overlaps. Various techniques for doing this are available.[45,46] A second approach is to remove side-chain atoms of residues responsible for a major blockage of the ligand, for instance, through mutation to an alanine residue. Both techniques are most effective if focused on a small subset of residues; too much softening leads to promiscuous ligand binding in a wide range of positions, orientations, and conformations. Mobile residues can be identified via several approaches, including X-ray B-factors, family-based structural analysis, examination of multiple crystal structures of the receptor complexed with different ligands, and conformational energy analysis. Automated construction of a suitably softened model is an area of current research. Currently, human intervention based on target expertise, enabling integration of all the above factors, is often helpful in obtaining accurate results.

It is assumed that the ensemble of poses from the initial docking will include at least one pose that is "close enough" to the experimental structure. Of course, the question of what is "close enough" and whether it is possible to recover at least one suitable pose in all cases are basic research topics requiring further study. The next step is to refine the ensemble of initial poses and use an energy and/or scoring model to select the pose closest to that observed experimentally. A variety or even combination of Monte Carlo, molecular dynamics, and conformational search algorithms can be used for this refinement; the pose selection function in general contains some molecular mechanics component, possibly an implicit solvation model, as well as empirical terms. Only a small number of attempts have been made to evaluate a complete protocol (sampling plus scoring) for a significant number of challenging test cases.[47] Our work, described in Ref. 40, provides grounds for encouragement. Twenty-one systems were examined, of which 17 failed (often rather drastically) in rigid-receptor cross-docking, presumably due to induced fit effects. Substantial improvement was obtained for all of these cases via the induced fit protocol, and an average RMSD of 1.3 Å, as compared to 5.5 Å for standard cross-docking using Glide, was reported. Reference 47 also reported some successes, although data sets were much smaller than those examined in Ref. 40.

Induced fit methods have already had a significant impact on a number of drug discovery projects, despite the relatively steep computational requirements (2 CPU hours/ligand or 15 minutes/ligand distributed over 20 CPUs). An obvious application is in the discovery of the binding mode of a novel lead compound prior to obtaining the crystal structure. A significant number of users of the Glide/Prime induced fit methodology of Ref. 40 have reported successful efforts along these lines. However, it is not yet clear whether the present methods are sufficiently robust to enable deployment in larger scale virtual screening exercises, or to improve accurate prediction of relative binding affinities in lead optimization. To answer these questions, a substantially larger data set should be examined and the speed of the algorithms should be enhanced for higher throughput application.

17.4 Binding Affinity Prediction

17.4.1 *Standard Empirical Scoring Functions*

A wide range of computational methods are available for predicting the binding affinity of protein-ligand complexes. These include highly computationally-intensive simulation-based approaches, such as thermodynamic integration/free energy perturbation theory (Chapter 19), more approximate, faster methods based on end-point simulations (linear interaction energy approaches,[48] MD/MM-PBSA, MD/MM-GBSA), and empirical or knowledge-based scoring functions that attempt to predict binding affinity given a protein-ligand complex geometry. There are also approaches based on fitting to structure-activity data of closely related ligands binding to the same receptor; such QSAR-based methods are often applied in the absence of any structural information. In the present section, we shall focus primarily on empirical scoring functions, as these are what are ordinarily utilized in currently available docking programs.

It is widely believed that the principal contributor to protein-ligand binding affinity is the free energy gained by displacing water molecules from the receptor active site. The largest free energy gains arise from displacing waters in hydrophobic regions of the receptor. In most empirical scoring functions, this free energy contribution is approximated as being proportional to either the hydrophobic surface area of the ligand in contact with hydrophobic groups of the protein, or as an atom-atom pair term summed over lipophilic ligand and protein atoms. The van der Waals interaction between ligand and protein can also be used.

Hydrogen bonds between the ligand and protein can also be a source of free energy gain upon ligand binding. The gain again comes from displacement of waters bound to the protein hydrogen-bonding partner by the ligand. Since no net hydrogen bonds are generally made or broken, the gain is smaller than the typical gas phase hydrogen bonding strength (~3–5 kcal/mol for two neutral groups, 10–15 kcal/mol for two charged groups), but can be as large as ~1–2 kcal/mol, principally arising from improved entropy of the water molecules upon transfer from their hydrogen bonded locations in the protein active site to bulk. Some scoring functions treat all hydrogen bonds identically, while others differentiate depending upon whether each partner atom is charged or neutral. The score is also typically modulated by a geometrical factor, reducing the predicted free energy gain as the geometry deviates from ideal.

A third commonly used term is aimed at representing the loss of ligand entropy upon binding, due to restriction of ligand torsional flexibility. This term is typically small, and not treated particularly well, given the approximations made and the neglect of other, similar terms such as ligand vibrational, rotational, and translational entropy. Nevertheless, various functional forms have been tried, and parameters optimized along with the remainder of the scoring function.

Finally, specialized terms are often introduced to represent metal-ligand binding.[13] Because there can be a covalent component to this binding, its contribution to overall binding affinity is particularly difficult to model. The contribution is difficult to obtain accurately as it requires estimation of the contribution to binding affinity (a small energy) from the difference in interaction energies of the metal ion with solvent and the ligand (both large energies).

Given a functional form implicitly or explicitly including the four terms described above, and automated or manual fitting using experimental binding affinity data to determine optimal parameter values, a standard empirical scoring function is defined. There are many such scoring functions in the literature,[22,49–56] one widely used version is the

ChemScore function developed by Eldridge and coworkers.[57] Other scoring functions, such as PLP,[56] use a significantly larger number of parameters by defining more atom-types and hence more atom-atom pair interaction terms, although the functional form is not very different from ChemScore. However, it is unclear whether the increased complexity of using more adjustable parameters yields higher accuracy when applied to systems beyond the training set.

It is, of course, possible to define *local* empirical scoring functions. These functions are trained on a single receptor or a small set of related receptors. A general-purpose docking program, however, employs a global scoring function that is intended to apply to all protein-ligand complexes with equal accuracy. The standard procedure is to fit the parameters of the scoring function to a "diverse" set of complexes, taken from the Protein Data Bank. Despite many attempts, it has proven difficult to achieve accuracy and robustness with the functional form and fitting protocol described above. When large and diverse data sets are examined, average errors of ~3 kcal/mol in binding affinity prediction are observed, and exceptionally large outliers with errors of 5–10 kcal/mol are not uncommon. An example of such an outlier is binding of biotin to streptavidin (PDB complex 1STP). Despite the small size of biotin, which has only 16 non-hydrogen atoms, it is the most tightly bound complex found in the PDB with a binding affinity of −18.3 kcal/mol. Most empirical scoring functions with the form outlined above yield results for this complex in error by ~5–10 kcal/mol, an extremely large error in both absolute and percentage terms.[58] Furthermore, until quite recently, no physical explanation for this discrepancy was offered.

17.4.2 *Improved Representation of the Hydrophobic Effect*

Description of the hydrophobic effect based on hydrophobic surface area contacts as in standard scoring functions can be shown to work reasonably well in describing small hydrophobic solutes in water. However, a body of evidence from the physical chemistry literature suggests that as the system becomes larger and more complex, significantly different behaviors can arise as the geometry of the hydrophobic moiety is altered. In particular, a region displaying hydrophobic enclosure of water molecules, i.e. enclosed on two sides at a 180-degree angle by lipophilic protein atoms, can create a situation where a surface area model underestimates the gain in free energy from displacing these waters. Clear examples of such enclosure include water in a carbon nanotube[59] and water between two hydrophobic plates.[60,61] If the nanotube diameter, or distance between plates, is comparable in size to the diameter of a water molecule, the water molecules in these regions can lose hydrogen bonds. In an extreme situation, this can cause dewetting of the cavity. This situation can be contrasted with simulations of water at a hydrophobic wall, as studied by Rossky and McCammon more than 20 years ago.[62] Water molecules at a hydrophobic interface do not on average lose hydrogen bonds — rather, they exhibit loss of entropy due to an inability to hydrogen-bond with the hydrophobic surface. Thus, one would expect that the free energy gain for displacing enclosed water molecules is significantly larger than that available from displacing water molecules in contact with a single hydrophobic surface. Yet, standard scoring functions treat both situations with the same functional form and parameterization.

Algorithms can be developed to recognize regions of hydrophobic enclosure in protein active sites, as is done in Glide v4.5 SP and XP. When groups of lipophilic ligand atoms occupy such sites, the predicted free energy is adjusted to reflect the additional free energy gained beyond the standard scoring function representation of the hydrophobic effect. Hydrophobically enclosed sites are common in pharmaceutical targets and represent regions targeted by medicinal chemistry, as these are regions where a small number of atoms can yield large gains in potency. The hydrophobic enclosure model in Glide provides, for the first time, a rapid, reasonably accurate way to recognize such sites and evaluate how well various ligands capture the free energy gains available due to the restrictive water environment. Figure 17.2 shows the napthyl moiety of the ligand in 1kv2, a p38 MAP kinase inhibitor that binds to the DFG-out mode of p38. The enclosure by hydrophobic protein residues is indicated by the green spheres

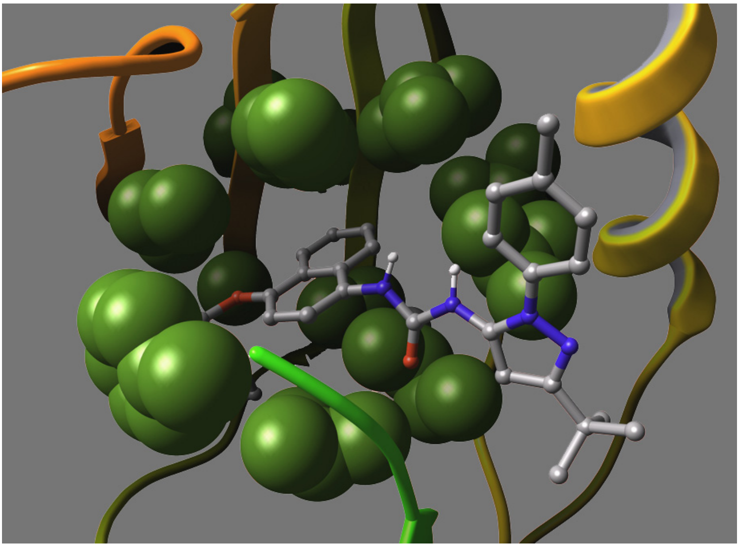


Fig. 17.2 Doramapimod (BIRB-796 from Boehringer Ingelheim) bound to human p38 map kinase (PDB entry 1kv2). Protein residues forming a region of hydrophobic enclosure about the naphthyl group are represented as green spheres.

surrounding the group. SAR data indicates that this group contributes a substantial amount of binding affinity.

A particularly important type of hydrophobically enclosed site is one in which the ligand also makes hydrogen bonds to the protein. The combination of hydrophobic enclosure and hydrogen bonding imposes severe constraints on water molecules occupying such a site (in the absence of ligand), leading to extremely unfavorable entropies of the water molecules. A characteristic motif is the presence of multiple hydrogen bonds in such sites (that we refer to as correlated hydrogen bonds), which, for example, are made to backbone NH and CO groups of the protein. The hinge region in kinases is an important example of this type of binding; Fig. 17.3 shows staurosporine bound to the kinase CDK2, in which a double correlated hydrogen bond is formed in a region with strong hydrophobic enclosure. A final example of this motif is the binding of biotin to streptavidin (Fig. 17.4), as discussed above, where a triply-correlated hydrogen bond is found in a region of extensive hydrophobic enclosure. Recent work employing

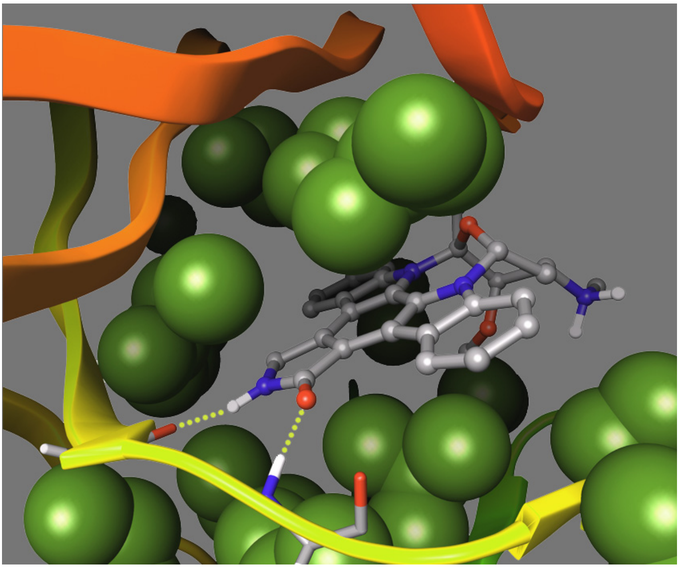


Fig. 17.3 Staurosporine bound to human cyclin dependent kinase. Key hydrogen bonds are shown with residues forming a region of hydrophobic enclosure represented as green spheres.

molecular dynamics simulations[63] confirms the unusual behavior of water molecules in such environments and is consistent with the enhanced free energy gains implemented in the Glide XP scoring function when such waters are displaced by suitable ligands.

17.4.3 *Enrichment Studies*

Over the past five years a considerable number of enrichment studies have been performed using a variety of scoring functions and rigid receptor docking method.[23] Enrichment studies utilizing flexible receptor docking have yet to be performed to any serious extent. In enrichment studies, ligands with sub-micromolar binding affinity (actives) to the receptor are seeded into a database of random, drug-like molecules (decoys) that are assumed to have larger-than-micromolar binding affinities. Not surprisingly, the character of the decoys can affect results significantly.[64] Various metrics for measuring enrichment have

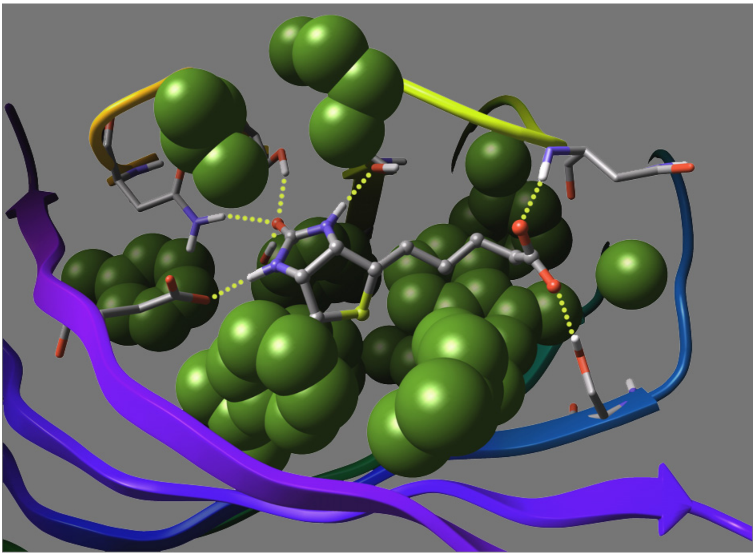


Fig. 17.4 Biotin bound to streptavidin. The identification of a triplet of correlated hydrogen bonds in the ring in a hydrophobically enclosed region, and the three hydrogen bonds to the ligand carbonyl within that ring results in very strong binding for this relatively small ligand.

been proposed, where the basic idea, given the ranked list of binding affinities predicted by the scoring function, is to examine the fraction of actives recovered as the percentage of the decoys recovered increases.[65]

Typical enrichment studies with state-of-the-art methods display a burst of "early" enrichment (in the top 1–2 % of the database, followed by a steady increase that often levels off, recovering 60–90% of the known actives. From our discussion above, we can interpret these results as follows. Early enrichment is due to compounds that fit particularly well into the receptor configuration. Weakness in recovering the bottom 30–40% of actives is most likely attributable to misdocking due to steric clashes. As discussed above, misdocking is then not really reflective of problems with the scoring function, but rather of the failure to treat induced fit effects. Until the induced fit effect is addressed more effectively, significant fractions of actives will continue to be misdocked, and hence, will be incorrectly scored.

This analysis suggests that a useful approach to calibrating the performance of scoring functions in enrichment is to work with data sets of actives that fit reasonably well into the target receptor conformation. A recent study indicates that Glide XP substantially outperforms alternatives when only well-docked actives are considered.[66] This is an encouraging result, as it shows that improvement of the physics of the scoring function leads to corresponding improvement in enrichment. However, until more progress is made in handling induced fit effects, misdocking will set fundamental limitations on the performance of Glide XP or any other scoring function in virtual screening.

In contrast to enrichment, empirical scoring functions have generally had difficulty in rank ordering compounds by binding affinities, an essential function if lead optimization problems are to be fully addressed by docking methods. Reasonable results are sometimes obtained, but for the most part this remains a very challenging problem, and a major focus of current research.

17.4.4 *Applications*

Initial application of docking and scoring in the early 1990s yielded anecdotal successes in facilitating the identification of lead compounds, particularly in the case of HIV protease inhibitors.[67,68] Over the next decade, enhancements in the cost/performance of computational platforms (culminating in the use of commodity personal computers, either via Linux clusters or grid computing) combined with improvements in docking software, algorithms, and scoring models, enabled virtual screening for lead discovery to be profitably performed on a wide range of targets in both academia and industry.[67–70] While complete separation of active from inactive compounds is not yet possible (for the reasons outlined above), even a relatively modest enrichment factor of 3–5 can enhance the efficiency of the lead discovery process. Higher enrichments, in the 10–30 range, can enable a small number of compounds to be evaluated with the expectation of obtaining some low micromolar hits with modest financial expenditure. Furthermore, the hits obtained from virtual screening are often complementary to those found in experimental HTS, facilitating the discovery of novel classes of compounds that might otherwise have been missed entirely. An interesting recent result is the use of virtual screening methods to identify an inhibitor of a protein-protein interaction, a particularly difficult class of target to access.[71]

The application of docking and scoring methods to lead optimization requires greater accuracy and reliability to make a significant impact. Nevertheless, over the last several years, use of these methods in practical drug discovery projects has expanded with a number of successful efforts reported in the literature.[72–77] Docking algorithms can be used to generate structures of new compounds in a lead series, providing powerful insight into the origin of structure-activity relationships and suggesting further directions for compound refinement. A particularly interesting study is that of Ref. 76 where the induced fit methodology of Ref. 39 was used to elaborate structures for a medicinal chemistry series. Interesting and important differences were found in the induced fit structures of closely related compounds, illustrating the need to consider protein flexibility even when investigating congeneric series.

In summary, computational approaches with docking and scoring as their core technologies are increasingly being integrated into the drug discovery processes at large pharmaceutical companies, biotechnology companies, and academic laboratories. For each type of organization, there are specific issues that will dictate the optimal deployment of computational tools. For example, large pharmaceutical companies have made a major investment in HTS facilities; they will benefit most from the complementary integration of virtual and HTS screening methods in lead discovery. At the opposite end of the spectrum, academic groups rarely have access to the million-compound HTS experiments that are routinely run in a large pharmaceutical company environment. For them, virtual screening, if it works, is highly attractive as the primary means of discovering lead compounds due to the low capital requirements and minimal ongoing expenses. Biotechnology companies fall somewhere in the middle of this spectrum; they can enhance HTS efforts with a much larger virtual compound collection by computationally screening a library of purchasable compounds. In the lead optimization phase, there is not yet sufficient data to confidently assess the potential impact of computation on improving the efficiency of a project, but this is an exciting direction that can be pursued by all the entities discussed above.

17.5 Future Outlook

The application of docking methods to lead discovery requires screening large numbers of compounds and the ability to handle active compounds complementary with many different receptor conformations. To achieve the large enrichment factors that would make virtual screening an essential lead discovery technology, receptor flexibility will have to be combined with a reasonably accurate scoring function at an acceptable computational cost for each ligand screened. We believe the current direction of scoring function improvements, as described above, will be sufficient to yield the requisite accuracy within a five-year period. The necessary speed can be achieved through a combination of the use of ensemble docking, acceleration of induced fit methods, and expanded computing capacity including grid computing. Because docking approaches are embarrassingly parallelizable, once accuracy of the methodology has been established, investments will be made in deploying thousands or tens of thousands of processors for virtual screening calculations. In five years, Moore's law predicts that the cost/performance per processor will improve from current benchmarks by roughly a factor of 10. Assuming the use of 5000 processors, if one wanted a virtual screen of one million compounds to complete in one day, one could afford to spend approximately seven minutes per ligand, which translates into 70 minutes per ligand using existing technology.

The more computationally expensive docking calculations currently require approximately five minutes per ligand with current processors. In contrast, current successful induced fit methods require hours of CPU time on a single processor. However, this typically involves rebuilding a large number of residues around the ligand for a set of top ligand poses. By intelligently reducing the number of side-chains that are treated flexibly (valid for the great majority of induced fit cases) and the number of initial poses to be retained (fewer poses will be required if very large perturbations of the active site are not being considered), major reductions in the CPU time required for induced-fit calculations can be achieved. These estimates suggest that computational power is not the real issue in deploying virtual screening on a large-scale basis, at least in large pharmaceutical companies. The key is to achieve a high level of accuracy for a wide range of targets — *robustness is at least as important as accuracy*. Generating poses with small RMSDs relative to experimental structures and high-quality predicted binding affinities by treating induced fit effects are some of the additional challenges that will need to be met as the technology matures.

Application of docking methods in lead optimization requires the ability to accurately treat induced fit effects and to rank order compounds by predicted binding affinities. The precision required is higher than for lead discovery; on the other hand, it is acceptable to require more CPU time per compound, since fewer compounds are typically considered.

The question of whether a high-accuracy, global empirical scoring function can be developed is an interesting one. There may be fundamental limits to such an approach. However, as more experimental data is obtained for a project, it should be possible to build and improve local empirical scoring functions, using what could be thought of as a structure-based QSAR approach. A few attempts in this direction have been published,[78] and this is an area we expect will develop rapidly in the next five years. Development is expected to be based on the use of improved functional forms and the widespread public availability of large quantities of experimental structural and binding data. An alternative is to follow docking calculations with other computational approaches, such as MD/MM-GBSA or MD/MM-PBSA, linear response calculations, or thermodynamic integration/free energy perturbation (FEP) calculations. All of these methods are undergoing continual improvement, and increases in computational power will enable more widespread application of simulation-based methods such as FEP, which has for many years been considered too computationally expensive for production use in a pharmaceutical environment. It is difficult to predict which approach or combination of approaches will be the most successful. What does seem clear is that, as docking methods are able to produce increasingly accurate initial poses, other downstream calculations will benefit as well from having an improved starting structure. Improvements in force fields, including the use of quantum chemical calculations directly in docking and simulations, will also be of increasing importance in ensuring robust treatment of a wide variety of chemical moieties.

In a wide range of science and engineering disciplines, including, for example, aerospace, petroleum exploration, and semiconductor chip design, computational modeling has become the central means by which new products are designed or discovered. Drug discovery requires atomistic modeling in contrast to the larger length scales relevant to the fields cited above, and must confront the fact that binding affinity is but a single property among many, such as pharmacokinetics, metabolism, and toxicity, to be optimized in a successful drug. Nevertheless, as computational power increases, and models and algorithms improve, we believe a technology transition in which computation forms a true platform for drug discovery projects will occur. The prospects for this transition taking place over the next five to 10 years are, in our view, substantial, and there is some possibility that it will occur even more rapidly. The computing power is there; the key is achieving accuracy and robustness in pose prediction and binding free energy prediction. While an optimized methodology along these lines will not directly address all of the properties cited above (although structure-based ADMET calculations are increasingly becoming feasible), the ability to explore huge chemical spaces rapidly, locate the tiny fraction of ligands with suitable binding affinities reliably, and remain in the tight binding region of chemical space as optimization of the other properties is carried out during the late stages of lead optimization (thus avoiding synthesis of dead compounds), can in principle provide a compelling advantage as compared to alternatives.

References

1. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. (1982) A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **161**: 269–288.
2. Berman HM, Westbrook J, Feng Z, *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**: 235–242. Also see www.rcsb.org.
3. Wang R, Fang X, Lu Y, Wang S. (2004) The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* **47**: 2977–2980. Also see www.pdbbind.org.
4. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. (2005) Binding MOAD (mother of all databases). *Proteins Struct Funct Bioinform* **60**: 333–340. Also see www.BindingMOAD.org.
5. Tirado-Rives J, Jorgensen WL. (2006) Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J Med Chem* **49**: 5880–5884.
6. Kuhn B, Kollman PA. (2000) Binding of a diverse set of ligands to avidin and streptavidin; an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J Med Chem* **43**: 3786–3791.
7. Nu H, Kalyanaraman C, Irwin JJ, Jacobson MP. (2006) Physics-based scoring of protein-ligand complexes: Enrichment of known inhibitors in large-scale virtual screening. *J Chem Inform Model* **46**: 243–253.
8. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M. (2005) Validation and use of the MM-PBSA approach for drug discovery. *J Med Chem* **48**: 4040–4048.
9. Rarey M, Kramer B, Lengauer T, Klebe GA. (1996) A fast flexible docking method using an incremental construction algorithm. *Chem Biol* **261**: 470–489.
10. Kramer B, Rarey M, Lengauer T. (1999) Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins Struct Funct Genet* **37**: 228–241.
11. Jones G, Wilett P, Glen RC, Leach AR, Taylor R. (1997) Development and validation of a generic algorithm and an empirical binding free energy function. *J Mol Biol* **267**: 727–748.
12. Nissink JW, Murray C, Hartshorn M, *et al.* (2002) A new test set for validating predictions of protein-ligand interaction. *Proteins* **49**: 457–471.
13. Friesner RA, Banks JL, Murphy RB, *et al.* (2004) Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **47**: 1739–1749.
14. Halgren TA, Murphy RB, Friesner RA, *et al.* (2004) Glide: A new approach for rapid, accurate docking and scoring 2. Enrichment factors in database screening. *J Med Chem* **47**: 1750–1759.
15. Friesner RA, Murphy RB, Repasky MP, *et al.* (2006) Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **49**: 6177–6196.

16. Jain AN. (2003) Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem* **46**: 499–511.
17. Jain AN. (2007) Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J Comput Aided Mol Des* **21**: 281–306.
18. Thomsen R, Christensen MH. (2006) MolDock: A new technique for high-accuracy molecular docking. *J Med Chem* **49**: 3315–3321.
19. Venkatachalam CM, Jiang X, Oldfield T, Waldman M. (2003) LigandFit: A novel method for the shape-directed rapid docking of ligands to protein active sites. *J Mol Graph Model* **21**: 289–307.
20. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP. (2006) eHiTS: A new fast, exhaustive flexible ligand docking system. *J Mol Graph Model* 7: 421–435.
21. Totrov M, Abagyan R. (1997) Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins Struct Funct Genet* **(1)**: 215–220.
22. Morris GM, Goodsell DS, Halliday RS, *et al.* (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**: 1639–1662.
23. Warren GL, Andrews CW, Capelli AM, *et al.* (2006) A critical assessment of docking programs and scoring functions. *J Med Chem* **49**: 5912–5931.
24. Jorgensen WL, Maxwell D, Tirado-Rives J. (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **118**: 11225–11236.
25. Cho AE, Guallar V, Berne BJ, Friesner R. (2005) Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J Comput Chem* **26**: 915–931.
26. Kua J, Zhang Y, McCammon JA. (2002) Studying enzyme binding specificity in acetylcholinesterase using a combined molecular dynamics and multiple docking approach. *J Am Chem Soc* **124**: 8260–8267.
27. Kua J, Zhang YK, Eslami AC, Butler JR, McCammon JA. (2003) Studying the roles of W86, E202, and Y337 in binding of acetylcholine to acetylcholinesterase using a combined molecular dynamics and multiple docking approach. *Protein Sci* **12**: 2675–2684.
28. Frimurer TM, Peters GH, Iversen LF, *et al.* (2003) Ligand-induced conformational changes: Improved predictions of ligand binding conformations and affinities. *Biophys J* **84**: 2273–2281.
29. Cavasotto CN, Abagyan RA. (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. *J Mol Biol* **337**: 209–225.
30. Ota N, Agard DA. (2001) Binding mode prediction for a flexible ligand in a flexible pocket using multi-conformation simulated annealing pseudo crystallographic refinement. *J Mol Biol* **314**: 607–617.
31. Ferrari AM, Wei BQ, Costantino L, Shoichet, BK. (2004) Soft docking and multiple receptor conformations in virtual screening. *J Med Chem* **47**: 5076–5084.

32. Carlson HA. (2002) Protein flexibility and drug design: How to hit a moving target. *Curr Opin Chem Biol* **6**: 447–452.
33. Knegtel RMA, Kuntz ID, Oshiro CM. (1997) Molecular docking to ensembles of protein structures. *J Mol Biol* **266**: 424–440.
34. Broughton HB. (2000) A method for including protein flexibility in protein-ligand docking: Improving tools for database mining and virtual screening. *J Mol Graph Model* **18**: 247.
35. Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS. (2002) Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins Struct Funct Genet* **46**: 34–40.
36. Claussen H, Buning C, Rarey M, Lengauer T. (2001) FlexE: Efficient molecular docking considering protein structure variations. *J Mol Biol* **308**: 377–395.
37. Huang SY, Zou X. (2007) Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins* **66**: 399–421.
38. Polgar, T, Keseru, GM. (2006) Ensemble docking into flexible active sites. Critical evaluation of FlexE against JNK-3 and beta-secretase. *J Chem Inform Model* **46**: 1795–805.
39. Sherman W, Beard HS, Farid R. (2006) Use of an induced fit receptor structure in virtual screening. *Chem Biol Drug Des* **67**: 83–84.
40. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R. (2006) Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem* **49**: 534–553.
41. Leach AR. (1994) Ligand docking to protein with discrete side-chain flexibility. *J Mol Biol* **235**: 345–356.
42. Leach A, Lemon A. (1998) Exploring the conformational space of protein side-chains using dead-end elimination and the A* algorithm. *Proteins Struct Funct Genet* **33**: 227–239.
43. Frimurer TM, Peters GG, Iversen LF, *et al.* (2003) Ligand-induced conformational changes: Improved predictions of ligand binding conformations and affinities. *Biophys J* **84**: 2273–2281.
44. Anderson AC, O'Neil RH, Surti TS, Stroud RM. (2001) Approaches to solving the rigid receptor problem by identifying a minimal set of flexible residues during ligand docking. *Chem Biol* **8**: 445–457.
45. Carlson HA, McCammon JA. (2000) Acommodating protein flexibility in computational drug design. *Mol Pharmacol* **57**: 213–218.
46. Jiang F, Kim SH. (1991) Soft docking — Matching of molecular-surface cubes. *J Mol Biol* **219**: 79–102.
47. Moitessier N, Therrien E, Hanessian S. (2006) A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic B-secretase (BACE 1) inhibitors. *J Med Chem* **49**: 5885–5894.

48. Aquvist J, Medina C, Samuelsson JE. (1994) A new method for predicting affinity in computer-aided drug design. *Proteins* **34**: 395–402.
49. Bohm JJ. (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* **8**: 243–256.
50. Bohm JJ. (1998) Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from *de novo* design or 3D database search programs. *J Comput Aided Mol Des* **12**: 309–323.
51. Rarey M, Dramer B, Lengauer T, Klebe GA. (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**: 470–489.
52. Jones G, Willet P, Glen RC, Leach AR, Taylor R. (1997) Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**: 727–748.
53. Ewing TJ, Makino S, Skillman AG, Kuntz ID. (2001) DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **15**: 411–428.
54. Wang R, Lai L, Wang S. (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des* **16**: 11–26.
55. Wang R, Lu Y, Wang S. (2003) Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem* **46**: 2287–2303.
56. Gehlhhar DK, Verkhivker GM, Rejto PA, *et al.* (1995) Molecular recognition of the inhibitor AG-1343 by HIV-1 protease – Conformationally flexible docking by evolutionary programming. *Chem Biol* **2**: 317–324.
57. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* **11**: 425–445.
58. Krammer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M. (2005) LigScore: A novel scoring function for predicting binding affinities. *J Mol Graph Model* **23**: 395–407.
59. Hummer G, Rasaiah JC, Noworyta JP. (2001) Water conduction through the hydrophobic channel of a carbon nanotube. *Nature* **414**: 188–190.
60. Zangi R, Hagen M, Berne BJ. (2007) Effect of ions on the hydrophobic interaction between two plates. *J Am Chem Soc* **129**: 4678–4686.
61. Huang X, Zhou R, Berne BJ. (2005) Drying and hydrophobic collapse of paraffin plates. *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys* **109**: 3546–3552.
62. Lee CY, McCammon JA, Rossky PJ. (1984) The structure of liquid water at an extended hydrophobic surface. *J Chem Phys* **80**: 4448–4455.
63. Young T, Abel R, Kim B, Berne BJ, Friesner RA. (2007) Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc Natl Acad Sci USA* **104**: 808–813.

64. Huang N, Shoichet BK, Irwin JJ. (2006) Benchmarking sets for molecular docking. *J Med Chem* **49**: 6789–6801.
65. Truchon J-F, Bayly CI. (2007) Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J Chem Inform Model* **47**: 488–508.
66. Zhou Z, Felts AK, Friesner RA, Levy RM. (2007) Comparative performance of several flexible docking programs and scoring functions: Enrichment studies for a diverse set of pharmaceutically relevant targets. *J Chem Inform Model* **ASAP Article**: 10.1021/ci7000346.
67. Kitchen DB, Decornez H, Furr JR, Bajorath J. (2004) Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat Rev Drug Discov* **3**: 935–949.
68. Kubinyi H. (2006) Success stories in computer-aided design. In: B Wang (eds.), *Computer Series in Drug Discovery and Development*, pp. 377–424. Wiley Interscience.
69. Kenyon V, Chorny I, Carvajal WJ, Holman TR, Jacobson MP. (2006) Novel human lipoxygenase inhibitors discovered using virtual screening with homology models. *J Med Chem* **49**: 1356–1363.
70. Klebe G. (2006) Virtual ligand screening: Strategies, perspectives, and limitations. *Drug Discov Today* **11**: 580–594.
71. Siddiquee K, Zhang S, Guida WC, *et al.* (2007) Selective chemical probe inhibitor of Stat3, identified through structure-based virtual screening, induces antitumor activity. *Proc Natl Acad Sci USA* **104**: 7391–7396.
72. Sauerberg P, Mogensen JP, Jeppesen L, *et al.* (2007) Design of potent PPAR alpha agonists. *Bioorg Med Chem Lett* **17**: 3198–3202.
73. Joseph-McCarthy D, Baber JC, Feyfant E, Thompson DC, Humblet C. (2007) Lead optimization via high-throughput molecular docking. *Curr Opin Drug Discov Devel* **10**: 264–274.
74. Tripathy R, Ghose A, Singh J, *et al.* (2007) 1,2,3-Thiadiazole substituted pyrazolones as potent KDR/VEGFR-2 kinase inhibitors. *Bioorg Med Chem Lett* **17**: 1793–1798.
75. Lyne PD, Lamb ML, Saeh JC. (2006) Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J Med Chem* **49**: 4805–4808.
76. Pomel V, Klicic J, Covini D, *et al.* (2006) Furan-2-ylmethylene thiazolidinediones as novel, potent, and selective inhibitors of phosphoinositide 3-kinase gamma. *J Med Chem* **49**: 3857–3871.
77. Maeda K, Das D, Ogata-Aoki H, *et al.* (2006) Structural and molecular interactions of CCR5 inhibitors with CCR5. *J Biol Chem* **281**: 12688–12698.
78. Datar PA, Khedkar SA, Malde AK, Coutinho EC. (2006) Comparative residue interaction analysis (CoRIA): A 3D-QSAR approach to explore the binding contributions of active site residues with ligands. *J Comput Aided Mol Des* **20**: 343–360.

Chapter 18

Structure-based Pharmacophores and Screening

R. Lewis*,† and R. G. Karki†

18.1 Introduction

The concept of a pharmacophore is one of the oldest and most robust concepts in medicinal chemistry, and has been used to underpin many successful hit-finding campaigns before and after the advent of routine access to protein structural information. The combination of pharmacophores with knowledge derived from the structures of protein-ligand complexes has opened up new possibilities: the use of steric constraints and the more precise specification of the geometry of a pharmacophore are just two examples. In this chapter, the theory behind pharmacophore modeling will be briefly reviewed, with an emphasis on the fields of current investigation, such as scoring and feature definition. Then the influence of structural knowledge on pharmacophore modeling will be discussed, especially the derivation or refinement of pharmacophores from the binding sites. Some examples of structure-based pharmacophoric screening will be presented. Finally, a future vision for the field will be put forward.

*Corresponding author.

†Novartis Institutes of BioMedical Research. Postfach, 4002 Basel, Switzerland. Email: richard.lewis@novartis.com.

18.2 Overview

Obtaining experimental structures of some of the therapeutically interesting biological targets, especially the membrane bound proteins such as the GPCR's have posed great challenges in structural biology. In such cases, the structure of a known ligand for the target guides the drug design process. Determining the fundamental characteristics of the ligand required for biological activity, in terms of the nature and disposition of chemical groups is the basis of pharmacophore modeling.

A pharmacophore was first defined by Paul Ehrlich in 1909 as "a molecular framework that carries (phoros) the essential features responsible for a drug's (= pharmacon's) biological activity."[1] In 1977, this definition was updated by Peter Gund to "a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity."[2] The International Union of Pure and Applied Chemistry (IUPAC's) definition of a pharmacophore is "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response." Notice that a presence of a pharmacophore is not sufficient to endow a structure with the desired activity.

In drug design, the term pharmacophore refers to a set of features that is common to a series of active molecules. Hydrogen bond donors, acceptors, hydrophobes, positively and negatively charged groups are the typical features. A 3D pharmacophore specifies the spatial relationships between the groups. These relationships are often expressed as distances or distance ranges but may also include geometric measures such as angles and planes. Perceiving a pharmacophore is one of the critical step towards understanding the interaction between a ligand and its biological target. Pharmacophore models maybe used as virtual screening tools wherein the pharmacophoric queries are used to search three-dimensional (3D) databases of small molecules to find new leads. Alternatively, they maybe used to obtain feature-based alignments for other inhibitors of the target. These alignments can be used to explain structure activity relationship

(SAR) of the compounds. Since 3D searches of databases using pharmacophoric queries can be run faster than docking, pharmacophore models can be used to filter compounds from a large database, followed by docking only the hits into the biological target of interest. This can speed up virtual screening of large databases.

18.3 Ligand-based Pharmacophore Model Generation

Pharmacophores can be identified and models generated from a series of molecules that are active against the biological target of interest. This is also referred to as a ligand-based approach and the major requirement is a set of molecules spanning a broad range of activity against the biological target of interest (1000-fold is ideal, 50-fold is minimal). In this approach, one makes the assumption that all the molecules bind to the same active site region of the receptor in a similar way and have a common mechanism of action. This assumption can often be incorrect, but it is the most parsimonious. There are several steps that are key for generating a good and reliable pharmacophore model via the ligand-based approach. Since this has been covered in detail elsewhere,[3] it will only be briefly discussed here. The important steps in a ligand-based pharmacophore model generation are:

1. Dataset selection and preparation
2. Conformational analysis
3. Pharmacophore enumeration
4. Ranking and selection of the representative models
5. Validation.

The quality of the pharmacophore model is dependent on the dataset that is used for the model building, so one has to be very careful in selecting it and representing it correctly. The molecules used for the pharmacophore model building are referred to as the "training set." The rule of thumb is to start with a set of compounds that have been tested in the same bioassay procedure, preferably in the same laboratory. Structural diversity is the second consideration. In real life, one should use a chemically diverse set of molecules spanning a broad activity range, although one can also get a pharmacophore model with good statistical parameters with a set of compounds belonging to the same chemical series. This could be attributed to differences in binding modes across different chemical series which will be hard to detect in the absence of any experimental structural data. The data preparation step involves checking for correctness in the chemical representation of the molecule, especially atom types, bond orders, stereochemistry, tautomers and charged state.

The next step is exploring the conformational space of each molecule. It is a known fact that the bioactive conformation may not necessarily be the lowest energy conformer, so it becomes necessary to do an exhaustive conformational search so that the bioactive conformation is also enumerated. Based on an analysis of a number of different X-ray co-complexes, Nicklaus *et al.*[4] have found a range of values (0.0 to 18 kcal/mol) between the protein bound conformation and the calculated global energy minima. Bostrom *et al.*[5] determined that 70% of ligands bind with strain energy of less than or equal to 3 kcal/mol (this should be interpreted in the context that 1.4 kcal/mol corresponds to a 10-fold change in affinity). Perola and Charifson[6] re-analyzed X-ray co-complexes and obtained results similar to that reported by Bostrom *et al.* The starting geometry of the molecule and the energy threshold used for the conformational analysis both influence the quality of the conformations generated: this in turn affects the quality of the pharmacophore model. In a recent study on a D2 antagonist dataset,[7] we generated conformations at four different energy threshold values 2, 5, 10 and 20 kcal/mol, starting from a local minima conformation, and the correct pharmacophore emerged at an energy threshold of 4 kcal/mol.[8] From our experience, we feel that one should study the quality of the conformations for a subset of highly flexible molecules by exploring the conformational space at different energy threshold values and finally use the threshold value that gives good coverage for generating conformations for the molecules in the entire dataset.

Once the dataset is prepared, the next step is to enumerate the different possible pharmacophore models. The same dataset may be compatible with multiple pharmacophores, so identifying them and ranking them are the subsequent steps. Some pharmacophores are degenerate or underdetermined, having the same features but different geometries. This may reflect that the training set does not contain molecules with the necessary conformational restraints. Pharmacophores with different features may also be equally correct, if there are multiple binding modes, or reflect the inherent limitations of the approach. Ranking is normally based on fitness and/or mapping of the training set molecules on the pharmacophore model. HypoGen[9] ranks the hypotheses on their cost value; this consists of three components, namely, the weight cost, error cost and configuration cost. The weight component increases in a Gaussian form as the feature weight deviates from the idealized value of 2.0. The error cost increases as the RMS distance between the estimated and the measured activities for the training set increases. The configuration cost represents the complexity or the entropy of the hypothesis space being optimized and is constant for a given data set.

Catalyst first calculates the costs of two theoretical hypotheses, namely, the ideal hypothesis (fixed cost) and the null hypothesis. The ideal hypothesis has a minimal error cost and the slope of the activity correlation is one. The null hypothesis has a maximal error cost and the slope of the activity correlation is zero. Together they represent the upper and lower bounds on cost for the hypotheses that are generated. The greater the difference between them, the greater is the likelihood that a meaningful hypothesis can be found. The closer the cost of the generated hypothesis is to that of the ideal hypothesis, the higher the probability that the generated hypothesis represents a true correlation in the data. The hypothesis with the least cost ideally would map to all the features of the most active compounds in the training set. The cost is reported in bits and a difference of about 50–60 bits between the generated hypotheses and the null hypothesis suggests that the correlation may be significant, which in turn requires a difference of about 60–70 bits between the costs of the ideal and null hypotheses.

Some of the commercially available programs for automatic pharmacophore model generation are: Catalyst,[9] DISCO (DIStance COmparisons),[10] DISCOtech™ (A faster version of DISCO),[11] GASP (Genetic Algorithm Superposition program),[12] PHASE,[13] etc. Although each of the programs is meant for ligand-based pharmacophore model generation, apart from differences in the algorithm used for the pharmacophore model generation, there are also differences in the definition of the pharmacophoric features and scoring of the models. The common pharmacophoric features are hydrophobes, hydrogen bond donor and acceptor, and positive and negative ionizable group. Features are defined by substructural fragments. This approach to feature definition is both a strength and a weakness. The unified definition of a feature can encompass bioisosteres; however, the simple classification into features does not reflect the continuum of feature strength, e.g. the oxygen of a carbonyl has very different h-bonding acceptor characteristics compared to that of a furyl ether. It is very difficult to break the definitions into finer divisions without encountering issues of granularity and training set coverage. Attempts to invent categories based on charge[14] or h-bonding strength (e.g. Abraham descriptors[15]) have not led to crisper models. The default is to stick with the broad categories, making small, tailored adjustments based on the particular needs of the training set. Some programs have separate feature definition for aromatic rings and so they are not considered as hydrophobes. In spite of the fact that the concept of pharmacophore and programs for generating them have been around for a long time, we still do not have a separate definition for metal chelating groups. Most programs have these groups defined as acceptors. Several times the same chemical group could be mapped under different feature definition. For example: a carboxylic acid group (COOH). It would fit the definition of negative ionizable feature, an acceptor feature for the C=O group and a donor feature for the OH group. However, most ligand-based pharmacophore generation program will not assign both the donor and acceptor feature at the same time because of the distance constraints imposed between two feature mapping in a pharmacophore model. This limit is an adjustable parameter. A good rule-of-thumb is to consider if a carboxamide should contain one or two features. Naturally, having a short inter-feature distance will favor, even overweight, feature-rich groups such as carboxylate, whose function in the molecule may be more about solubilization than binding affinity. Also, it is difficult sometimes to guess the physiological state of the ligand in the active site as this would depend upon the nature of the amino acids in the ligand-binding site. All indirect experimental information, e.g. from NMR, pKa measurements should be considered at this stage. In such cases, a structure-based approach to generating pharmacophore models maybe more beneficial.[16]

18.4 Structure-based Pharmacophore Perception

Receptor-based pharmacophore models can be generated only if the structure of the active site of the receptor is known or if a ligand-receptor complexed structure is available. The receptor structure could be from an experimental structure, either X-ray or NMR, or in the absence of such a structure, a homology model may also be used as a starting point. If a ligand-receptor complexed structure is available, the pharmacophore model can be generated via translation of the ligand-receptor interactions into feature definitions either from a single structure or multiple complexed structures. Docking poses of a subset of ligands in a protein target of interest may alternatively be used as a starting point to derive structure-based pharmacophore models. Inherent protein flexibility can be incorporated in the process of pharmacophore model generation by using conformational ensembles of the protein that is either sampled from a molecular dynamics simulation[17] or obtained from multiple experimental structures of the protein. In either case pharmacophore models can be identified from conserved regions after overlaying the active sites from each protein conformation.

As in the ligand-based approach, there are some points that need to be considered for pharmacophore model generation using the structure-based approach.

18.4.1 *Data Preparation*

Check for missing residues and add them using a modeling program if they are part of the active site. The next step is the protonation of the amino acids. This will have to be done after consideration of the charged state of the amino acids. Since in most cases the assays are designed to mimic the physiological state of the protein in its natural environment, in our opinion protonating the amino acids considering the pH of the bioassay used for testing the compounds would be a good approach. When docking poses are used as starting points for the structure-based identification of pharmacophore models, it is important that even the ligands should be protonated after considering the ionization state. In experimental structures in protein databank (PDB) format, the bond orders for the ligands and the atom types are not defined. So one has to assign bond types and hybridization states from the geometrical information. Minimization of the ligand-receptor complex or just the receptor would be helpful to relieve the steric clashes but this is not essential.

18.4.2 *Identifying and Ranking the Pharmacophore Models*

From a ligand-receptor complexed structure a pharmacophore model can be perceived based on the interactions made by the ligand with the receptor. These interactions can be translated into features such as hydrogen bond donor, hydrogen bond acceptor, hydrophobic groups, and positive and negative ionizable features, similar to the ligand-based approach. Regions in the receptor that are not accessible by the ligand are defined as excluded volumes. One of the ways of scoring the pharmacophore models is by correlating the score with the dissociation constant for the ligand-receptor complex as implemented in structure-based Focusing in Cerius2.[18,19] Commercially available programs for structure-based pharmacophore model generation are Structure-based Focusing (SBF)[18] (Accelrys Inc); Sprout (SimBioSys Inc.); LigandScout (Inte:Ligand), and GRID. Each program follows a separate set of rules to generate a set of interaction sites for each atom or functional group of the protein that is capable of participating in a non-bonded contact. The rules are largely based on statistical analysis of experimental structures from the protein data bank and take into account the chemical nature of the atoms as well as energetically favorable orientations of chemical features such as hydrogen bond donors/acceptors and hydrophobic groups. For example, the hydrogen bond interaction distance in LigandScout has been extended from 2.5 (H_BOND_MIN_DISTANCE) to 3.8 (H_BOND_MAX_DISTANCE) Å in order to include all plausible interactions in low resolution PDB structures with additional geometric constraints for sp2 and sp3 donor atoms.[20] Although SPROUT is a *de novo* ligand design system, one of the modules called HIPPO (Hydrogen Bonding Interaction Site Prediction as Positions with Orientations) is meant to select interaction sites within a receptor site. These interaction sites are known as target sites in SPROUT and they are used as starting points for structure generation, while in LigandScout and SBF, the interaction sites are used to generate pharmacophore queries to search 3D databases. HIPPO can identify metal ions and residue motifs that tend to form covalent bonds to ligands (e.g. Ser-His-Asp triad) and generate the appropriate target sites for them. It can identify metal ions (Zn, Mg, Cu, Ca, Co, Fe, Ni, Mn) in the receptor PDB file, calculate the most likely direction of the free valency according to the existing connections (to protein or solvent atoms) and generate the appropriate target site. One may consider using the HIPPO identified target sites to define pharmacophoric queries outside of SPROUT to either align ligands and explain the SAR for other inhibitors of the target, or for searching 3D databases to find new leads.

Questions around the quality of our methods for generating pharmacophores are being raised, especially as there have been no major advances since GASP and DISCOtech. Researchers are actively revisiting some of the fundamental issues surrounding sampling of conformational space, the generation of ensembles of solutions, and the scoring of those solutions. The Sheffield group[21] have developed a multi-objective genetic algorithm (MOGA), based on their experiences with GASP. The conflicting objectives are conformational energy and the degree of overlap/similarity of the structures when overlaid according to the pharmacophore hypothesis. One inherent difficulty is that the “correct” answer is often not known. Most methods can produce several plausible solutions but this may reflect the difficulty the programs have in sampling the search space. The MOGA does find a wider range of solutions than other stochastic approaches. Another advantage is that conformational space is sampled on the fly, rather than relying on a pre-computed set of conformers, which will bias the search space. The disadvantage is that the MOGA does not allow for partial matches, so the pharmacophore needs to be built from compounds that all have (similar) high affinity.

18.5 Future Outlook

To enunciate a future vision, we must first understand the fundamental questions: typically, multiple pharmacophore possibilities are consistent with the SAR. So the problem lies in sifting through them to find the “real” one, and realizing that you have found it. The use of external indirect data would be especially valuable, if it could be incorporated into the methodology. How could one, for example, add in distance constraints derived from NMR either intraligand, or between residues in the protein and the ligand? A pharmacophore may not exist for an SAR. This may be because different ligands may have multiple binding modes. Can we derive statistical approaches that can work with multiple binding modes, and then discard those that do not explain sufficient information in the dataset, much as one discards variables.[2] However, using powerful statistical analysis tools, it is easy to find patterns in the data that are not physically real, and therefore have zero prospective utility (“Data when tortured long enough will confess to anything” – P. Hein). How does one estimate the robustness of the pharmacophore model? Possibly this could be done through a sensitivity analysis, and looking for strong changes for a few critical parameters, for example, energy cutoff. HTS techniques offer the possibility of much larger datasets, but with smaller dynamic range. Can this data be used in any way, even only to pull out the crudest of models? Despite the inherent simplicity of the pharmacophore concept, it is clear there is much enjoyable research still to be done.

References

1. Ehrlich P. (1909) Present status of chemotherapy. *Chem Ber* **42**: 17–42.
2. Gund P. (1977) Three-dimensional pharmacophoric pattern searching. *Prog Mol Subcell Biol* **5**: 117–143.
3. Kristam R, Gillet VJ, Lewis RA, Thorner D. (2005) Comparison of conformational analysis techniques to generate pharmacophore hypotheses using catalyst. *J Chem Inf Model* **45**: 461–476.
4. Nicklaus MC, Wang SM, Driscoll JS, Milne GWA. (1995) Conformational changes of small molecules binding to proteins. *Bio Med Chem* **3**: 411–428.
5. Bostrom J, Norrby PO, Liljefors T. (1998) Conformational energy penalties of protein-bound ligands. *J Comp Aid Mol Des* **12**: 383–396.
6. Perola E, Charifson PS. (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem* **47**: 2499–2510.
7. VanDrie JH. (1997) Strategies for the determination of pharmacophoric 3D database queries. *J Comp Aid Mol Des* **11**: 39–52.
8. van Drie JH. (2006) Identifying the optimal energy window in pharmacophore discovery. *Abs ACS* **232**.
9. Accelrys Software Inc. (2005) Catalyst, Release 4.11, Accelrys Software Inc., San Diego.
10. Martin YC, Bures MG, Danaher EA, *et al.* (1993) A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J Comp Aid Mol Des* 7: 83–102.
11. (2006) DISCOTECH, Sybyl 7.3, Tripos International, 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.
12. Jones G, Willett P, Glen RC. (1995) A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J Comp Aid Mol Des* **9**: 532–549.
13. Phase. (2007) Phase, version 2.5, Schrödinger, LLC, New York, NY.
14. Chau P, Dean P. (1994) Electrostatic complementarity between proteins and ligands. 2. Ligand moieties. *J Comp Aid Mol Des* **8**: 527–544.
15. Abraham MH, Ibrahim A, Zissimos AM, *et al.* (2002) Application of hydrogen bonding calculations in property based drug design. *Drug Disc Today* **7**: 1056–1063.
16. Mason JS, Morize I, Menard PR, *et al.* (2007) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of

combinatorial libraries containing privileged substructures. *J Med Chem* **42**: 3251–3264.

17. Carlson HA, Masukawa KM, Rubins K, *et al.* (2000) Developing a dynamic pharmacophore model for HIV-1 integrase. *J Med Chem* **43**: 2100–2114.
18. (2005) Accelrys, Inc., Cerius2 Modeling Environment, Release 4.10L, San Diego.
19. Bohm HJ. (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known 3-dimensional structure. *J Comp Aid Mol Des* **8**: 243–256.
20. Wolber G, Langer T. (2005) LigandScout: 3-d pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J Chem Inform Model* **45**: 160–169.
21. Cottrell SJ, Gillet VJ, Taylor R. (2006) Incorporating partial matches within multiobjective pharmacophore identification. *J Comp Aid Mol Des* **20**: 735–749.

Chapter 19

Molecular Dynamics-based Free Energy Simulations

M. A. Cuendet[†], V. Zoete[†] and O. Michielin[*,†,‡]

19.1 Introduction

Free energy represents the most important quantity to describe the behavior of a molecular system. The probabilities of the different states of a system are indeed directly related to the value of their free energy. In the case of proteins, for example, the conformational change between two states, the folding process, the association between two monomers, or the affinity of a small molecule for its receptor are all described by the free energy. For this reason, much effort has been devoted to the development of computational methods that allow reliable estimates of this quantity for a given molecular system and a given process under investigation.

The theoretical foundations for free energy simulations can, to a large extend, be attributed to Kirkwood for his 1930s pioneering developments[1] on the computation of free energy differences using thermodynamic integration (TI) and to Zwanzig for the free energy perturbation (FEP) method.[2] The first applications of this formalism to biological problems came in the early 1980s with the work of

[*] Corresponding author: E-Mail: olivier.michielin@unil.ch

[†] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland.

[‡] Ludwig Institute for Cancer Research and Multidisciplinary Cancer Center, Ch. des Boveresses 155, 1066 Epalinges, Switzerland.

Tembe and McCammon on protein-ligand binding,[3] followed by Peter Kollman and coworkers with the first alchemical simulations (see Section 19.2) to estimate the binding free energy difference between a wild type and a point-mutated protein.[4] Since these early days, free energy simulation techniques have been the subject of intense research efforts. Only recently have these methods become reliable, due on the one hand to the better sampling provided by the more powerful computers available today, but more importantly, to improved theoretical approaches with better convergence properties. In this chapter, we will review the basic methods used in the field as well as the most recent theoretical developments.

The statistical mechanics definition of the free energy of a system in a given state A is

$$G_A = -k_\mathrm{B} T \ln Z_A,$$

where k_B is the Boltzmann constant, T is the temperature, and Z_A is the partition function. In complex systems, such absolute free energies are intrinsically impossible to compute, because the partition function is essentially a measure of the full configuration space accessible to the system. In experiments as well as in simulation, free energies are always computed relatively to a reference state. Such a free energy difference between two states A and B is given by a ratio of partition functions,

$$\Delta G_{AB} = -k_\mathrm{B} T \ln \frac{Z_B}{Z_A}. \tag{19.1}$$

The main idea behind methods presented below is to avoid direct computation of the individual partition functions Z_A and Z_B by using the fact that the variations between states A and B of interest are often localized in relevant regions of the configuration space. Elsewhere, the corresponding partition functions Z_A and Z_B have a high degree of similarity. Most approaches correspond, therefore, to reformulating Equation (19.1) such that common parts of Z_A and Z_B not directly relevant to the process under investigation cancel out. A fundamental aspect of these approaches is that they express, as we will see, the free energy difference in terms of an ensemble average, which can be directly measured or calculated in a simulation, unlike an absolute free energy.

In the first section, we present methods derived from first principles that are exact at the statistical mechanics level. For these methods, the quality of the results (for a given model or force field) depends mainly on the quality of the sampling and on convergence properties. In the second section, approximate methods will be described. These methods are not exact at the statistical mechanics level but do show interesting convergence properties that make them very useful in some applications. The last section is a summary and outlook on the present and future potential of free energy methods.

19.2 Exact Methods

When addressing specific biological questions, various free energy differences can be of interest. For example, in the case of ligand binding to a receptor, one might want to compute:

(i) the relative binding free energy between different ligands
(ii) the absolute binding free energy of a ligand
(iii) the full binding free energy profile or potential of mean force (PMF).

The transformation under study can be of two types:

(1) The first may involve a change in the nature of the system. It can be a modification of certain interactions or the exchange of an entire group of atoms with another. Indeed, unlike in experiments, any parameter in the potential energy function describing the system can be varied in a simulation. In this case, the reaction coordinate is an external parameter, which connects the two physical states of interest with unphysical hybrid intermediate states. We call this an alchemical transformation.
(2) Alternatively, the transformation may involve a conformational change in the system, such as the binding of the ligand, originally

in solution, to a receptor. In this case, the reaction coordinate λ is a function of the atomic coordinates of the system. Then, all intermediate states are physical, and the full PMF along λ is of interest.

In the following, we first present the two most used exact statistical mechanical approaches to calculate free energy differences. Both approaches can be applied to either alchemical or conformational transformations. We then review alchemical methods for the calculation of relative free energy differences using thermodynamic cycles and of absolute binding free energy differences. We finally cover the most used methods to calculate PMFs. Excellent review articles[5–11] as well as a fine book[12] on free energy calculation are available. Other studies provide efficiency comparisons of various methods.[13–16]

19.2.1 *Exact Statistical Mechanics Methods for Free Energy Differences*

Here, we briefly derive FEP and TI expressions, which can be applied in molecular dynamics (MD) as well as Monte-Carlo simulation to calculate free energy differences.[17] At this level, both alchemical and conformational types of reaction coordinates can be treated in the same general way under the common notation λ. We however keep in mind that in the alchemical case λ is an external parameter changing the functional form of the system's Hamiltonian. In the case of conformational changes, $\lambda = \lambda(r)$ is a function of the coordinates r and the Hamiltonian remains unchanged, except for the addition of a biasing term.

19.2.1.1 *Free energy perturbation*

Consider a well-defined state A described by the Hamiltonian

$$H_A(r,p) = \sum_i \frac{p_i^2}{m_i} + U_A(r),$$

with p_i the momentum of particle i, and $U_A(r)$ the potential energy function. For a given number N of particles at constant volume and temperature, state A is described by the partition function

$$Z_A = \frac{1}{h^{3N} N!} \int e^{-\beta H_A(r,p)} drdp,$$

where $\beta = k_B T$. The normalization constant contains Plank's constant h, which is a measure of the elementary volume in phase space, and the factor $N!$, which should be present only when the particles are undistinguishable. Similarly, let state B be described by H_B and characterized by Z_B. By definition, the free energy difference between A and B is

$$\Delta G_{AB} = -k_B T \ln \frac{Z_B}{Z_A}. \tag{19.2}$$

By inserting a unity factor in the form $e^{+\beta H_A(r,p)} e^{-\beta H_A(r,p)}$ into the numerator, we get

$$\Delta G_{AB} = -k_B T \ln \frac{\int e^{-\beta H_B(r,p)} e^{+\beta H_A(r,p)} e^{-\beta H_A(r,p)} drdp}{Z_A}.$$

This can be seen as a phase space average of the quantity $e^{-\beta[H_B - H_A]}$ in state A,

$$\Delta G_{AB} = -k_B T \ln \left\langle e^{-\beta[H_B - H_A]} \right\rangle_A. \tag{19.3}$$

This approach is generally attributed to Zwanzig.[2] In practice, a single simulation in the reference state A is performed, during which the above phase space average is converged. The accuracy of the free energy evaluation can be improved if one can perform a simulation in state B as well. In such a case, FEP from A to B and from B to A can be optimally combined in a single expression using the so-called Bennett acceptance ratio,[18]

$$\Delta G_{AB} = -k_{\mathrm{B}}T \ln \frac{\left\langle \min\left(1, e^{-\beta[H_B - H_A]}\right)\right\rangle_A}{\left\langle \min\left(1, e^{-\beta[H_A - H_B]}\right)\right\rangle_B} .$$

The FEP method can give meaningful results only if the two states *A* and *B* overlap in phase space, meaning that configurations are sampled in which the difference $H_B - H_A$ is smaller than $k_{\mathrm{B}}T$. Often, for transformations of practical interest, this is not the case. The solution is to introduce *n* intermediate states between *A* and *B*, such that the overlap between successive states is good. The Hamiltonian $H(r,p,\lambda)$ is made a function of a parameter λ, which characterizes the intermediate states, such that $H(r,p,\lambda_A) = H_A(r,p)$ and $H(r,p,\lambda_B) = H_B(r,p)$. One is free to introduce as many intermediate λ-steps as necessary, since their free energy differences simply cumulate to give

$$\Delta G_{AB} = \Delta G_{A1} + \Delta G_{12} + \cdots + \Delta G_{nB} \cdot$$

The total free energy difference is then recovered by applying the FEP method between each successive intermediate states and summing all contributions,

$$\Delta G_{AB} = -k_{\mathrm{B}}T \sum_{i=0}^{n} \ln \left\langle e^{-\beta[H(\lambda_{i+1}) - H(\lambda_i)]}\right\rangle .$$

Note that the intermediate states, in other words, the unphysical path linking states *A* and *B*, are completely arbitrary since ΔG_{AB} is a thermodynamic state function. Thus, the intermediate states can be chosen such as to optimize the simulation convergence. For example, the functional form of the λ-dependence in different terms of $H(r,p,\lambda)$ can be adapted,[19] or smaller λ intervals can be chosen in regions where $dH/d\lambda$ is large. In particular, special care has to be taken to avoid numerical singularities when making Lennard-Jones particles appear, for example, by using the soft core scaling method.[20]

19.2.1.2 *Thermodynamic integration*

Assuming that the two states *A* and *B* are linked by a coupling parameter λ as defined above, and that the free energy *G* is a continuous function of λ, we have the identity

$$\Delta G_{AB} = G_B - G_A = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{dG}{d\lambda} \right\rangle_{\lambda} d\lambda .$$

Using the definition of the free energy, Equation (19.2), we have

$$\frac{dG}{d\lambda} = -\frac{k_B T}{Z_\lambda} \int \frac{d}{d\lambda} e^{-\beta H(\lambda)} drdp = \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} .$$

This leads to

$$\Delta G_{AB} = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda , \tag{19.4}$$

which is the TI formula.[1,21] Note that the FEP formula, Equation (19.3) can be recovered from Equation (19.4) by considering a first-order numerical approximation of the Hamiltonian derivative.[10] In practice, simulations are performed at a number of fixed λ-values between and including λ_A and λ_B, during which the analytical derivative of $H(\lambda)$ is calculated and the phase space average in Equation (19.4) is estimated. In the end, the integration over λ is performed numerically. Note that the same care has to be taken as for FEP in choosing the λ-dependence of $H(\lambda)$, in order to avoid numerical singularities and optimize convergence.

The recent adaptive integration method[22] seeks to estimate the same integral as TI, Equation (19.4). In addition to fixed-λ sampling, it uses a Metropolis Monte Carlo procedure to generate moves that change the value of λ during the simulation. This method seems to be one of the most efficient up-to-date.[14]

19.2.2 *Relative Free Energy Differences from Thermodynamic Cycles*

A common application of MD free energy calculation is to compute the relative binding free energy of two ligands L_1 and L_2 to a receptor R. In this case, one can avoid the computationally difficult task of computing directly the binding free energy of each ligand, ΔG_1 and ΔG_2, by using the thermodynamic cycle depicted in Fig. 19.1. Since free energy is a state function, the difference of the horizontal legs is equal to the difference of the vertical legs

$$\Delta\Delta G_{12} = \Delta G_2 - \Delta G_1 = \Delta G_{bind} - \Delta G_{solv}\,.$$

Therefore, $\Delta\Delta G_{12}$ can be obtained by calculating the solvation free energy difference ΔG_{solv} and the receptor interaction free energy difference (in solution) ΔG_{bind} between L_1 and L_2. In both cases, this is done by mutating one ligand into the other, and using one of FEP, TI, or even a nonequilibrium method (see below) to determine ΔG_{solv} and ΔG_{bind}. The method was devised in 1984[23] and first applied to a protein-ligand system in 1987.[4] The same approach can be used for

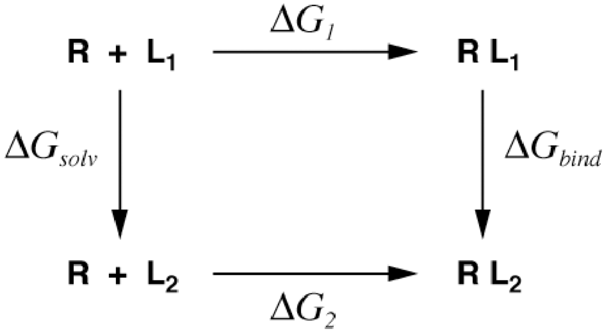


Fig. 19.1 Thermodynamic cycle for the binding of two ligands L_1 and L_2 to a receptor R.

various applications, such as relative solvation free energies or sequence dependence of protein-protein interactions.[24] Note that thermodynamic cycles can be extended to multiple ligands. A related approach based on FEP is the single-step perturbation method,[25] in which relative free energies for not too different compounds are estimated by perturbation from a single simulation of an unphysical reference state that encompasses the characteristic molecular features of the compounds.

19.2.3 *Absolute Binding Free Energy Differences Using the Double Decoupling Method*

Full description of the mechanism of a ligand L binding to a receptor R requires knowing the absolute binding free energy of the process

$$(R)_{sol} + (L)_{sol} \xrightarrow{\Delta G^0_{bind}} (RL)_{sol}, \tag{19.5}$$

instead of the free energy (double) difference between two ligands L_1 and L_2. Comparison with experimental results requires calculating ΔG^0_{bind} with respect to a given standard condition. The direct calculation of ΔG^0_{bind} according to Equation (19.5) would require a simulation that starts with P and L bound and then follows the (un)binding process to the completely separated ligands. This amounts to calculating a full PMF, a computationally intensive process described in Section 19.2.4. The Double Decoupling Method[26,27] overcomes this problem using the following thermodynamic cycle:

$$(L)_{sol} \xrightarrow{\Delta G_I} (L)_{gas}$$

$$(RL)_{sol} \xrightarrow{\Delta G_{II}} (R)_{sol} \cdots (L)_{gas} \xrightarrow{\Delta G^0_r} (R)_{sol} + (L)_{gas}.$$

In the first line, the ligand is transferred from the solution to the gas phase by decoupling its interactions with the solvent. The corresponding ΔG_I can be computed using TI or FEP methods. Similarly, in the second line, the ligand is first decoupled from both the receptor and the solvent. The intermediate state $(R)_{sol} \ldots (L)_{gas}$ is a hypothetical state in which all interactions of L with R as well as with the solvent have been turned off, but L is maintained by an artificial restraint in a position and orientation close to the bound state. The full binding free energy is recovered through

$$\Delta G^0_{bind} = \Delta G_I - \Delta G_{II} - \Delta G^0_r \cdot$$

Including the $(R)_{sol}\ldots(L)_{gas}$ state in the calculation has two benefits (not present in the prior Double Annihilation Method[28]). First, the restraint simplifies the calculation of ΔG_{II} using TI or FEP, because it alleviates the requirement of sampling all positions and orientations of L with respect to R. Second, it appears that all the standard state dependence of ΔG^0_{bind} is included in the term ΔG^0_r. Essentially, ΔG^0_r expresses the ratio of the phase space volume available to the ligand in the restrained and free states, and depends on the standard concentration C^0. Analytical expressions of ΔG^0_r were given for positional restraints with respect to a fixed point in space,[29] or for general positional and angular restraints with respect to the receptor.[27] Recently, a similar method based on RMSD restraints of a flexible ligand was proposed[30] and successfully applied together with FEP for the determination of protein-ligand binding free energies.[31]

19.2.4 *Potentials of Mean Force from Configurational Transformations*

To this point, we have considered transformations in which the reaction coordinate λ was an external parameter changing the nature of atoms or the strength of interactions between them. Albeit very useful in simulations, these transformations are unphysical, and only endpoint free energy differences are meaningful. Conversely, in PMF calculations, the reaction coordinate is associated with the position of atoms and describes the binding pathway of a ligand to a receptor or a conformational change in a protein. Thus, all intermediate states are physically relevant. In this case, $\lambda = \lambda(r)$, with $r = (r_1,r_2,\ldots,r_N)$

representing the coordinates of the N particles in the system. The PMF is then defined as

$$\Delta G(\lambda) = -k_{\mathrm{B}}T \ln \rho(\lambda) + C, \tag{19.6}$$

where C is a constant and $\rho(\lambda)$ is the probability of finding the system at λ on the reaction path,

$$\rho(\lambda) = \frac{1}{Z}\int e^{-\beta H(r,p)}\delta(\lambda - \lambda(r))\,drdp,$$

with $\delta(.)$ representing the Dirac function.

If the reaction coordinate is nonlinear in the Cartesian coordinates r, an additional term appears in the equations above, corresponding to the determinant of the Jacobian of the coordinate transformation.[13] For example, if $\lambda(r)$ is the Euclidean distance between two particles, the Jacobian corresponding to the polar coordinates (λ,θ,ϕ) of the second particle with respect to the first particle is $\lambda^2\sin\theta$. This leads to

$$\Delta G(\lambda) = -k_{\mathrm{B}}T \ln \rho(\lambda) + 2k_{\mathrm{B}}T \ln \lambda + C.$$

The additional term accounts for the increasing phase space volume corresponding to a given λ for increasing λ. For conciseness, we will leave this term out of the following developments, but we insist on its importance.

As the system evolves, $\lambda(r)$ changes spontaneously, responding to the forces at play. After enough time, the system would sample the whole available reaction path (unlike in the alchemical transformations discussed above). In this case, the most immediate way to get the PMF is to build a histogram of occurrences of system configurations around λ, which gives directly $\rho(\lambda)$. In most useful applications however, the PMF has barriers much higher than $k_{\mathrm{B}}T$, and the corresponding regions will be poorly sampled in the limited duration of a simulation. In the following, we describe two kinds of methods to improve the sampling of specific regions of the λ coordinate[13]:

(i) The system can be restrained in the vicinity of a given reaction coordinate λ_0 by adding to the Hamiltonian a potential energy term $u(\lambda(r),\lambda_0)$, which is usually chosen to be harmonic. This method is called umbrella sampling.[32]

(ii) The system can be constrained to move on a hypersurface $\lambda(r) = \lambda_0$. This effectively reduces the dimensionality of the phase space, and additional terms arise in the PMF, due to the fact that the momentum conjugate to the λ coordinate is zero.

19.2.4.1 *The umbrella sampling method*

Since its first application in MD in 1982,[33] the umbrella sampling method is probably the most popular approach to calculate PMFs. The method involves several simulations restrained around separate λ values noted $\{\lambda_i\}$, using bias potentials $u_i(\lambda) = u(\lambda,\lambda_i)$. In windows around each λ_i, a regional biased PMF is computed by constructing the biased λ-coordinate probability $\tilde{\rho}_i(\lambda)$. Local PMFs are finally unbiased and recombined as follows: the unbiased probability $\rho_i(\lambda)$can be expressed[10,34] in terms of the biased probability as

$$\rho_i(\lambda) = e^{+\beta u_i(\lambda)}\tilde{\rho}_i(\lambda)e^{-\beta G_i} ,$$

where the undetermined constant G_i defined from $e^{-\beta Gi} = \langle e^{-\beta ui}\rangle$ represents the free energy associated with the introduction of the bias potential u_i. The unbiased PMF around λ_i is then

$$\Delta G_i(\lambda) = -k_{\mathrm{B}}T \ln \tilde{\rho}_i(\lambda) - u_i(\lambda) + G_i + C \cdot \qquad (19.7)$$

In early applications, the constants G_i were obtained by manually adjusting the various PMFs of adjacent windows such that they match in the regions in which they overlap. There is, however, an efficient method for unbiasing, optimally determining the G_i, and combining each local PMF into a smooth free energy profile in one go. The weighted histogram analysis method (WHAM) was originally derived for Monte-Carlo data,[35] and was later applied to umbrella sampling.[34,36,37] The idea is that, at a given λ, the total probability $\rho(\lambda)$ is an average of the $\rho_i(\lambda)$, weighted according to the Boltzmann factor of u_i at λ and to the number n_i of data points collected in window i. The resulting expression to determine $\rho(\lambda)$ from the biased $\tilde{\rho}_i(\lambda)$is

$$\rho(\lambda) = \frac{\sum_i n_i \tilde{\rho}_i(\lambda)}{\sum_j n_j e^{-\beta\left[u_j(\lambda) - G_i\right]}} \cdot$$

The constants G_i are determined using the optimal estimate for the distribution function $\rho(\lambda)$,

$$e^{-\beta G_i} = \int d\lambda e^{-\beta u_i(\lambda)} \rho(\lambda) \cdot$$

Because $\rho(\lambda)$ itself depends on the constants $\{G_i\}$, the WHAM equation must be solved self-consistently. In practice, this is achieved through an iteration procedure starting with an initial guess for the $\{G_i\}$ and repeated until both equations are satisfied up to a fixed tolerance.

The umbrella sampling method together with the corresponding WHAM analysis can be straightforwardly extended to multidimensional reaction coordinates.[36,38] For an application to the ion conduction through the potassium channel, see Ref. 39. As another extension of the method, the use of adaptive bias potentials has been proposed.[40] These bias potentials are taken as the negative of the local PMF, and are iteratively adapted as the PMF estimate is refined.

The usual umbrella sampling method is based on the determination of the free energy by counting occurrences, via Equation (19.7). Simulations with restraining potentials can as well be combined with the two other free energy methods described previously: FEP and TI. The central idea of the FEP-based umbrella sampling method[41] is to calculate the free energy difference $\Delta G_{i,i+1} = G_{i+1} - G_i$ between windows u_i and u_{i+1} from an average performed in window i,

$$\Delta G_{i,i+1} = -k_{\mathrm{B}} T \ln \left\langle e^{-\beta[u_{i+1}(\lambda) - u_i(\lambda)]} \right\rangle_i .$$

If the windows are sufficiently close, the PMF can simply be approximated by the points $G_i = G_1 + \Delta G_{1,2} + \cdots + \Delta G_{i-1,i}$, which carry most of the PMF features. The more elaborate methods[41,42] use the G_i as offsets for local detailed PMFs, which are then combined to find the complete PMF.

The long unexplored combination of TI with umbrella sampling provides a substantial benefit over the regular umbrella sampling: no offsets between windows need to be estimated.[43] This comes from the fact that the derivative of the free energy $dG/d\lambda$ is extracted in each window, and not its absolute value. One way to estimate $dG/d\lambda$ from a biased simulation is based on the fact that the average of the restraining force equals the opposite of the average of the physical (unbiased) force along the reaction coordinate.[13] Thus, the contribution of window i for a given λ is

$$\frac{dG}{d\lambda} = \left\langle \frac{\partial u_i}{\partial \lambda} \right\rangle_i .$$

Alternatively,[43] $dG/d\lambda$ can be estimated using the biased probability of occurrence $\tilde{\rho}_i(\lambda)$ estimated from window i,

$$\frac{dG}{d\lambda} = -k_B T \frac{\partial \ln \tilde{\rho}_i}{\partial \lambda} - \frac{\partial u_i}{\partial \lambda} .$$

In order to avoid numerical noise in taking the derivative of $\tilde{\rho}_i(\lambda)$, it is approximated by a simple Gaussian. The contributions of all windows are averaged (without having to determine unknown offsets), and the resulting $dG/d\lambda$ profile is numerically integrated to find the PMF $\Delta G(\lambda)$.

19.2.4.2 *Constraint-based methods*

Instead of adding restraint potentials to the Hamiltonian in order to enhance the sampling of certain regions, constraints can be used. A constrained system is forced to evolve on a hypersurface of fixed $\lambda(r) = \lambda_0$, which raises two difficulties. First, the momentum conjugate to the generalized coordinate λ is zero. Second, if $\lambda(r)$ is an internal degree of freedom, different regions of configuration space may get different weight factors depending on the Jacobian, or more precisely, the mass-metric tensor of the transformation from Cartesian to internal coordinates. It took many years for accurate formulations of this problem to be found, one of the first being the Blue Moon Method.[44] Formulas are available for general reaction coordinates,[45,46] but as they become mathematically rather involved, we restrict ourselves here to constraints of the type

$$\lambda(r) = |r_1 - r_2| = \lambda_0.$$

Essentially, the terms that emerge in the expression of $\Delta G(\lambda)$ contain the determinant of the $(N_c \times N_c)$-matrix **g** containing the derivatives of the N_c constraints Q_c with respect to the Cartesian coordinates r,

$$\mathbf{g} = \left(\frac{\partial Q_c}{\partial r}\right)^T M^{-1}\left(\frac{\partial Q_c}{\partial r}\right),$$

with M the $(3N \times 3N)$ diagonal matrix containing the particle masses. With only one distance constraint, the determinant becomes $|\mathbf{g}| = 1/m_1 + 1/m_2$. Since it does not depend on λ, this factor is absorbed in the constant C.

A first possible expression[13] for the PMF around λ_0 has the form of FEP, Equation 19.3,

$$\Delta G(\lambda) = -k_B T \ln\left\langle e^{-\beta[V(\lambda)-V(\lambda_0)]}\frac{\lambda^2}{\lambda_0^2}\right\rangle_{\lambda_0} + C\cdot$$

Here, the Boltzmann factor includes the potential energy $V(\lambda)$ in internal coordinates. The constrained ensemble average is weighted by the factor λ^2/λ_0^2, which is in fact the ratio of the determinants of the Jacobians at λ and λ_0, characterizing the transformation from Cartesian to internal coordinates. The results from several simulations constrained at successive points $\{\lambda_i\}$ along the reaction coordinate can be combined with the WHAM to get the full PMF.

A second approach to obtain a PMF from a series of constrained simulations is based on TI. The average of the constraint force $f^c(\lambda)$ equals the opposite of the average (physical) force along the λ coordinate. Thus, $f^c(\lambda)$ can be used in the TI formula Equation 19.4,

$$\Delta G(\lambda) = +\int_{\lambda_0}^{\lambda} \left\langle f^c(\lambda')\right\rangle_{\lambda'} d\lambda' + 2k_{\mathrm{B}}T \ln\frac{\lambda}{\lambda_0} + C\cdot$$

The second term is due to the non-Cartesian nature of the reaction coordinate $\lambda(r) = |r_1 - r_2|$. If the constraint is enforced by using a Lagrange multiplier technique such as the SHAKE procedure,[47] the constraint force $f^c(\lambda)$ is obtained from the value of the Lagrange multipliers. Alternatively, the projection along λ of the unconstrained forces due to $V(r)$ can be used in the above formula.

19.2.4.3 *Advanced methods*

The two main limitations to free energy calculations are the accuracy of the model (force field or level of theory in quantum calculations), and conformational space sampling. Many methods have been devised to address the sampling problem,[48] and we restrict ourselves here to a few examples of advanced sampling methods directly related to the calculation of free energy differences.

Force bias methods[49–51] rely on replacing the force acting along λ with a force of zero mean or with an adaptive bias force. This leads to nearly uniform sampling of λ. The PMF can rigorously be recovered from the force along λ, in a similar way as in constrained simulations.

In the adiabatic free energy dynamics method,[52] an artificial adiabatic separation is created between the reaction coordinate and the rest of the system. Fictitious masses ensure that λ evolves slowly, while maintained at a temperature high enough to freely overcome free energy barriers. In a related and more elaborate method called metadynamics,[53] a set of collective variables is selected as a multi-dimensional reaction coordinate λ. The collective variables have their own dynamics, which drives the physical system while being adiabatically decoupled from it. Based on the Local elevation method,[54] small Gaussian potentials are accumulated in regions visited by the collective variable, which ensures a complete sampling of the reaction coordinate. The PMF is recovered from the sum of all small Gaussian potentials.

19.2.5 *Nonequilibrium Methods*

Nonequilibrium statistical mechanics have long remained an abstract theoretical field, even after Evans derived the fundamental fluctuation theorem in 1993.[55,56] The field has gained considerable interest[57] since 1997, when a nonequilibrium relation with direct practical perspectives, the Jarzynski identity (JI), was published.[58] The second law of thermodynamics states that the average work of a process cannot be smaller than the difference of free energies between the initial and the final states, $\langle W \rangle \geq \Delta G$. Conversely, the JI is a relation between the same quantities that holds regardless of the speed of the process,

$$e^{-\beta \Delta G_{AB}} = \left\langle e^{-\beta W_{AB}} \right\rangle_A . \tag{19.8}$$

Here, the nonequilibrium work W_{AB} is path-dependent, and the average $\langle \cdot \rangle_A$ is over different trajectories with independent canonically distributed initial conditions in state A. Substantial theoretical work has been devoted to the JI, which was proved to apply to a variety of dynamics,[59–63] including the specific thermostated or barostated equations of motion used in MD.[64,65] The JI opens the possibility of calculating equilibrium free energy differences from nonequilibrium processes, as verified experimentally.[66] A strong requirement, however, is to have a sufficiently large collection of trajectories for an accurate estimation of the exponential average in Ref. 8, which is a major concern for practical applications.

The JI is extremely relevant to molecular simulation, where the system can be perturbed at will. In steered molecular dynamics,[37,67] a time dependent external steering potential of the form

$$u(r,t)=\frac{1}{2}k(d(r)-\lambda(t))$$

is added to the Hamiltonian, in order to actuate a conformational change of the system. In the case of a ligand dissociation from a receptor, $d(r)$ is the instantaneous distance between the ligand and the receptor. $\lambda(t)$ is the reference reaction coordinate, which is monotonously increased during the simulation (k is an harmonic constant). During this operation, the work $W(t)$ as a function of time can be integrated. If this protocol is repeated with canonically distributed initial conditions (in practice, these are frames taken from a long equilibrium simulation in the bound state), the JI can be applied to reconstruct the unbinding PMF. This requires special care,[63,68] as three operations need to be performed:

(i) correction of the bias introduced by the steering potential
(ii) transformation of the fluctuating $W(t)$ into a smooth function of λ
(iii) evaluation of the exponential average.

The first two operations are addressed by the stiff spring approximation,[67] or an adapted WHAM analysis.[69] In far-from-equilibrium cases, where the dissipative part of the work is large, a strong bias can appear in the estimation of the exponential average in Equation 19.8. Several methods have been proposed to overcome this problem, such as the cumulant expansion method,[58,67] block averaging,[70] weighted sampling of the work values,[71] or a combination of the Jarzynski identity with transition path sampling.[70] However, the convergence of Equation 19.8 inherently relies on rare events,[72] which hampers the practical efficiency of the method, as shown in several studies.[15,73,74]

In 1999, Crooks[60] derived a result slightly more general than the JI for cases where work measurements are available in both directions, $A \to B$ and $B \to A$. The Crooks theorem[60] (CT) holds for Langevin or Hamiltonian dynamics, as well as for thermo-barostated MD,[75,76]

$$e^{-\beta\Delta G_{AB}} = e^{-\beta W}\,\frac{P_{AB}(W)}{P_{BA}(-W)}\,.$$

Here, $P_{AB}(W)$ is the probability of observing a given work value W in the forward process and $P_{BA}(W)$ in the reverse process. Table 19.1 places the CT in a general perspective among other free energy relations. The CT can be applied directly by identifying ΔG_{AB} with the work value where the forward and reverse work distributions intersect, $P_{AB}(W) = P_{BA}(-W)$. This was done in an experimental corroboration[77] of the CT.

A maximum likelihood approach based on a slightly more general expression of the CT[78] leads to the Bennett acceptance ratio method.[15,79] It provides an optimal estimate of ΔG_{AB} given a set of N_f nonequilibrium work values in the forward direction, and N_r in the reverse direction, as the solution of

$$\sum_{N_f}\left(1+e^{\beta[\eta+W_{AB}-\Delta G_{AB}]}\right)^{-1} = \sum_{N_r}\left(1+e^{\beta[-\eta+W_{BA}+\Delta G_{AB}]}\right)^{-1},$$

where $\eta = k_B T \ln(N_f/N_r)$.

Table 19.1 Synthetic View of Free Energy Relations

	Equilibrium	Non equilibrium
One way	$e^{-\beta\Delta G_{AB}} = \left\langle e^{-\beta[H_B-H_A]}\right\rangle_A$ Zwanzig[2] (1954)	$e^{-\beta\Delta G_{AB}} = \left\langle e^{-\beta W_{AB}}\right\rangle_A$ Jarzynski[58] (1997)
Two ways	$e^{-\beta\Delta G_{AB}} = \dfrac{\left\langle \min\left(1, e^{-\beta[H_B-H_A]}\right)\right\rangle_A}{\left\langle \min\left(1, e^{-\beta[H_A-H_B]}\right)\right\rangle_B}$ Bennett[18] (1976)	$e^{-\beta\Delta G_{AB}} = e^{-\beta W}\,\dfrac{P_{AB}(W)}{P_{BA}(-W)}$ Crooks[60] (1999)

19.3 End-Point Methods

End-point methods, which sample only the free and bound states and compute ΔG_{bind} by taking a difference, have been widely used recently to study macromolecular structural stability or association, as well as protein-ligand binding in relation with drug design (DD) applications. These methods are attractive because of their simplicity, their low computational cost compared to more exact methods such as FEP or TI, and the fact that they can be applied to structurally diverse compounds, since they do not need the simulation of an unphysical transformation between molecules. However, their theoretical foundation still needs to be strengthened, although efforts are being made in this direction.[80] Here, we will review two examples, the Linear Interaction Energy[81] (LIE) and the molecular mechanics Poisson-Boltzmann surface area[82,83] (MM-PBSA) models.

19.3.1 *LIE*

In the LIE[81,84,85] approach, two MD simulations are performed: one for the ligand alone in solution, and the other for the solvated complex. The MD simulations are generally performed in explicit solvent using ligand-centered stochastic boundary conditions. Frames are then extracted from the MD simulations and are used to compute the averaged van der Waals and electrostatic interaction energies between the ligand and its environment in the bound ($V_{vdW,bound}$ and $V_{elec,bound}$) and free states ($V_{vdW,free}$ and $V_{elec,free}$). The binding free energy, ΔG_{bind}, is then estimated using

$$\begin{aligned}\Delta G_{bind} = {} & \alpha\left(\left\langle V_{vdW,bound}\right\rangle - \left\langle V_{vdW,free}\right\rangle\right)\\ & +\beta\left(\left\langle V_{elec,bound}\right\rangle - \left\langle V_{elec,free}\right\rangle\right) + \gamma\ .\end{aligned}$$

The first term in the above equation holds for the nonpolar contributions to ΔG_{bind}. Its linear relationship with the surrounding van der Waals energies is based on the observation that solvation energies of nonpolar compounds are linearly correlated with the surrounding van der Waals energies.[84,86] The second term describes the electrostatic contribution to ΔG_{bind} according to the linear response approximation (LRA) theory.[81,87,88] γ is a constant that can be added to get the correct absolute binding free energies. In the initial implementation, β was fixed to ½ following the LRA approximation, while α was fitted empirically to a value of 0.16 to reproduce the experimental activity of four structurally related endothiapepsin inhibitors.[81] γ was kept to 0 to limit the over-parameterization. Although these parameters gave satisfying results for protein-ligand systems, it was found using FEP calculations that β could be considered has a function of the ligand nature. Values of 0.5, 0.43, 0.37, and 0.33 were suggested for ionic molecules, and neutral compounds with one, two, or more hydroxyl groups, respectively.[84] A 0.18 value was found to be optimal for α. Non-zero values of γ can be necessary to reproduce ΔG_{bind} for some systems.[89] More recently, it was suggested that γ could be expressed as a function of the buried solvent accessible surface area (SASA) of the ligand that is buried upon complexation,[90] leading to the modified equation

$$\Delta G_{bind} = \alpha\left(\left\langle V_{vdW,bound}\right\rangle - \left\langle V_{vdW,free}\right\rangle\right) + \beta\left(\left\langle V_{elec,bound}\right\rangle - \left\langle V_{elec,free}\right\rangle\right) + \lambda\left(\left\langle SASA_{bound}\right\rangle - \left\langle SASA_{free}\right\rangle\right).$$

However, this has been questioned since the buried SASA is correlated to the change in V_{vdW}, making the new term equivalent to adding a constant.[85,91] The current general view is that α, β and γ depend on the system that is studied and should be fitted to reproduce a set of experimental activities of known ligands.[92] The fitted parameters can in turn be used to estimate the activities of new or virtual compounds. However, this parameters variability has been questioned by Åqvist and coworkers, who found that α and β are both force field and system independent,[93] while γ remains the only free parameter. The latter depends on the hydrophobicity of the binding site and is a function of the fraction of hydrophobic surface area.[93]

Several important contributions to molecular recognition are neglected in LIE, such as the conformational rearrangement upon complexation of the ligand and the receptor, the receptor desolvation energy, and the entropies. However, it has been argued that these terms are implicitly taken into account by the LRA approximation and the adjustable parameters of the model.[90,91] This method is generally applied to structurally related molecules, and the cancellation of errors expected from that limitation contributes to explain the success of the method in estimating ΔG_{bind}.

Recently, efforts were made to replace the explicit solvent model in LIE by an implicit solvent model: the generalized Born model (GB)[90] or Poisson-Boltzmann (PB).[94] In this implementation, the electrostatic term is replaced by a function of the coulomb interaction energy between the ligand and the protein, and of the solvation reaction field energy. This variant is attractive due to the reduced CPU cost. The MD simulation can then be performed using explicit[94] or implicit solvent models.[90]

LIE has been intensively studied in the context of DD applications. A recent and detailed review can be found elsewhere.[90] The average root mean squared errors between experimental and LIE-determined energy is typically around 0.5 to 1.5 kcal/mol, thus similar to FEP or TI but at a much cheaper computational cost. The method has also been used successfully to study the effect of mutations on protein-protein association.[95,96]

19.3.2 *MM-PBSA*

In MM-PBSA, ΔG_{bind} is written as the sum of the gas phase contribution, ΔH_{bind}^{gas}, the energy difference due to translational and rotational degrees of freedom, $\Delta H_{trans/rot}$, the desolvation free energy of the system upon binding, ΔG_{desolv}, and an entropic contribution, $-T\Delta S$[82,83]:

$$\Delta G_{bind} = \Delta H_{bind}^{gas} + \Delta H_{trans/rot} + \Delta G_{desolv} - T\Delta S.$$

The term ΔH_{bind}^{gas} contains the van der Waals and electrostatic interaction energies between the two partners in the complex, and the internal energy variation (including bond, angle, and torsional angle energies) between the complex and the isolated molecules, ΔH_{intra}.

In the classical limit, $\Delta H_{trans/rot}$ is equal to *3RT*. This constant term is generally omitted in MM-PBSA calculations. ΔG_{deslov} is the difference between the solvation free energy, ΔG_{solv}, of the complex and that of the isolated parts. ΔG_{solv} is divided into the electrostatic, $\Delta G_{elec,solv}$, and the nonpolar, $\Delta G_{np,solv}$, contributions, such that

$$\Delta G_{solv} = \Delta G_{elec,solv} + \Delta G_{np,solv}.$$

In MM-PBSA, $\Delta G_{elec,solv}$ is calculated by solving the Poisson or the Poisson-Boltzmann equation,[97,98] depending whether the salt concentration is zero or nonzero. Recently, an approach related to MM-PBSA, where $\Delta G_{elec,solv}$ is determined using a GB[99] model, has been introduced under the name molecular mechanics-generalized Born surface area[82,100] (MM-GBSA). Despite its approximations, the GB model makes this variant attractive because it is much faster than PB. Recent advances of GB models[101,102] in reproducing the PB solvation energies of macromolecules as well as desolvation energies upon binding further support the use of GB in this context.[103] The term $\Delta G_{np,solv}$, which can be considered as the sum of a cavity term and a solute–solvent van der Waals term, is assumed to be proportional to the SASA,

$$\Delta G_{np,solv} = \gamma SASA + b.$$

This well-known and often used approximation comes from the fact that the ΔG_{solv} of saturated nonpolar hydrocarbons is linearly related to the SASA.[104,105] Several linear models exist. The surface tension γ and the constant b can be set to 0.00542 kcal mol^{-1} Å^{-2} and 0.92 kcal mol^{-1}, respectively, if $\Delta G_{elec,solv}$ is calculated from PB.[106] Values of 0.0072 kcal mol^{-1} Å^{-2} and 0 kcal mol^{-1},[107] or 0.005 kcal mol^{-1} Å^{-2} and 0 kcal mol^{-1} can be used together with GB models.[108] Recently, an alternative model for $\Delta G_{np,solv}$ using a cavity solvation free energy term plus an explicit solute–solvent van der Waals interactions energy term has been tested.[108] This model led to better results in estimating ΔG_{bind} for the Ras-Raf association, although the transferability of the results was questioned.[108]

The entropy term, due to the loss of degrees of freedom upon association, is decomposed into translational, S_{trans}, rotational, S_{rot}, and vibrational, S_{vib}, contributions. These terms are calculated using standard equations of statistical mechanics.[109,110] S_{rot} is a function of the moments of inertia of the molecule, whereas S_{trans} is a function of the mass and the solute concentration. S_{trans} is the only term in the free energy of an ideal solution that depends on solute concentration, leading to the concentration-dependence of the binding reactions. The vibrational entropy term is calculated with the quantum formula from a normal mode analysis (NMA).[110] A quasiharmonic analysis of the MD simulations is also possible. However, it has been found that it does not always yield convergent values, even using very long MD simulation trajectories, and also led to large deviations from the results obtained with NMA, giving an overall unreasonable entropic contribution.[108]

In the standard MM-PB(GB)SA protocol, the energy terms are averaged over 200 to 500 frames extracted from MD simulation trajectories, typically performed in explicit solvent. Both periodic and stochastic boundary conditions have been used. Explicit water molecules are removed prior to energy calculations, since the solvent effect is described according to a PBSA or GBSA implicit solvent model. More recently, some studies also performed the MD simulations using implicit solvent models.[111,112] The normal modes are usually calculated on a smaller number of frames, due to the CPU requirement of such calculations. Short 0.5 to 1 ns trajectories are generally performed and yield to converged energy terms. Longer simulations have been tested, up to 10 ns in length.[108] However, they were not found to provide better results, most probably because long simulations emphasize force field errors and limitations. Indeed, it has been found that MM-PBSA yields better results with MD simulations restrained around the X-ray structure, compared to unrestrained simulations.[113] Two possibilities are arising concerning the number of MD simulations to perform. In principle, one should make three trajectories, one for the complex and each of the isolated partners, and calculate the energy terms using the adequate simulation. However, a popular alternative consists in performing only one MD simulation for the complex. In this variant, the terms relative to one isolated partner are calculated after removing the atoms of the other partner in the frames extracted from the MD simulation of the complex. As a consequence, the reorganization energy of the molecules upon association is neglected ($\Delta H_{intra} = 0$). However, this variant is less CPU demanding and leads to increased convergence due to cancellation of errors, reduction of noise arising from flexible remote regions relative to the binding site, and conformational restraints imposed by the complex geometry. Thus, this one-simulation variant is attracting when ΔH_{intra} may be reasonably neglected. Comparisons between one- and three-trajectories results can be found in the literature.[51,103,113]

MM-PB(GB)SA is expected to estimate absolute ΔG_{bind} without adjustable parameters. Although several studies were able to reproduce experimental ΔG_{bind} for protein-protein association with an error lower than 2 kcal mol^{-1},[103,108] these results are open to discussion. Indeed, the approach contains several "hidden" parameters, like the force field used, the choice of PB or GB and that of the nonpolar solvation model, the use of one or three trajectories, and the different terms that can be included or neglected. As a consequence, it is sometimes possible to find a combination of such hidden parameters apparently allowing a fine estimation of ΔG_{bind} for a given system. However, the transferability of such results to other systems is questionable. Nevertheless, MM-PB(GB)SA has proven to be useful for several applications less sensitive to the choice of hidden parameters, such as the comparison of relative stabilities of macromolecular conformations, determination of relative affinities for different small ligands in DD applications, and estimation of the effect of mutations on association processes and fold stability.

Although some studies aimed at determining absolute ΔG_{bind} for ligand-protein association, MM-PB(GB)SA is usually used to estimate relative affinities for different ligands targeting the same protein. This allows additional approximations, like the neglect of the entropy terms for ligands of similar masses binding to the same site. Also, despite the fact that this approach is expected to tackle chemically diverse ligands, it is often applied to a series of chemically related ligands. This obviously simplifies the problem thanks to additional cancellation of errors, but it also reflects the usual DD processes that generally focus on families of similar ligands. A recent and detailed review of the numerous studies using MM-PB(GB)SA in the context of DD can be found elsewhere.[92] MM-PB(GB)SA has given variable results, ranging from poor correlations between experimental and calculated ΔG_{bind}, to very good ones, with correlation coefficients up to 0.96.[92] The performance seems to be a function of the nature of the targeted protein and of the range of activities encompassed by the ligands. Not surprisingly, the ranking is better for a broader range of affinities.[114]

MM-PB(GB)SA has been found to perform well at determining the effect of mutations on association processes, and identifying the hot-spots of protein-protein complexes.[100,103,108,115–118] Two main approaches exist. First, it is possible to perform a so-called computational alanine scanning[115,116] (CAS) in which the absolute ΔG_{bind} is calculated for the wild type system, as well as for several mutants in which one residue has been replaced by an alanine. The alanine mutation is introduced by modifying the frames extracted from the MD simulation of the wild-type system. The difference in ΔG_{bind} between the wild type system and the mutants may be compared directly to the results of an experimental alanine scanning[115,116] (AS). The second possibility is to perform a binding free energy decomposition[100] (BFED) for the wild type system. This process aims at calculating the contributions to ΔG_{bind} arising from each atom or groups of atoms (typically side-chains). Like CAS, the BFED also identifies the nature of the energy change in terms of interaction and solvation energies, or entropic contributions. The detailed description of the BFED process can be found elsewhere.[100,117] The MM-GBSA variant is attractive for BFED, not only because it is much faster than MM-PBSA, but also because the pair-wise nature of the GB equation allows the decomposition of $\Delta G_{elc,solv}$ into atomic contributions in a straightforward manner.[100] It is however interesting to note that the decomposition of a PB calculated $\Delta G_{elec,solv}$ can also be performed,[118] though it is more computationally demanding. Although its results cannot be compared directly to an experimental AS, the BFED offers a faster alternative to the CAS, since it only requires one binding free energy calculation.

Also, it allows studying the contributions from non-mutable groups of atoms, such as backbone atoms. In addition, contrary to CAS, the BFED is a non-perturbing approach that does not require introducing a mutation in the system. A comparison between CAS and BFED results can be found in Ref. 117. Obviously, these methods cannot be expected to provide results exactly comparable to values obtained from an experimental AS, since they both neglect the effect of the mutations on the protein conformation. However, fair agreements between the experimental and theoretical results have been found in several studies and open the way to rational protein engineering.[100,116,118,119] It has been found that the side-chain contributions to S_{vib} play an important role, and increase the quality of the correlation between experimental and calculated energy changes.[100] A theoretically exact way to calculate the contribution of a given group of atoms to S_{vib} is to zero their mass and recalculate the normal modes and the corresponding total entropy.[100] The difference between the wild type system S_{vib} and that of the system with some zeroed masses gives the contribution of the corresponding atoms. This approach is very time consuming since it requires a NMA for each group of atoms. Consequently, the entropic contribution is often neglected in such studies or calculated for the most important residues. However, a new vibrational entropy decomposition scheme has been introduced recently to circumvent this problem: the linear decomposition of the vibrational entropy[117] (LDVE) approach, which necessitates only one NMA for the wild type system, and is thus much faster. It is based on the idea that the most important contributions to S_{vib} originate from side-chains that contribute most to the vibrational amplitude.

Recently, the CMEPS[119] approach (computational mutations to estimate protein stability) has been introduced. It uses MM-GBSA calculations to study the impact of mutations on protein structural stability and determine the most important residues for the protein fold. It is based on the notion that the ΔG_{bind} corresponding to the alchemical complexation of a given side-chain (considered as a "pseudo-ligand") into the rest of the protein (considered as a "pseudo-receptor") reflects the importance of this side-chain to the thermodynamic stability of the protein. This method has been applied successfully to the study of insulin,[119] p53,[120] and PPAR[121] structural stability.

19.4 Future Outlook

Free energy simulations have gained a lot of maturity and robustness over the last one or two decades. They now provide an invaluable set of tools to assess the effect of a mutation in a protein receptor, a modification of a lead compound, the relative probability of two conformations in a protein, the folded state of a peptide, and many important biological questions.

Some of the new methods mentioned in this chapter are just starting to be used in biological applications. Seeing the impact that the standard approaches have had over the last years in various fields like protein design or drug design, it is very likely that these new developments are going to change the way we use molecular dynamics to understand molecular behavior. The continuous increase in computer power, though not sufficient *per se* to lead this evolution, will be useful in allowing the more rigorous methods to be tested in conditions where microstate sampling is no longer the limiting factor.

References

1. Kirkwood JG. (1935) Statistical mechanics of fluid mixture. *J Chem Phys* **3**(5): 300–313.
2. Zwanzig RW. (1954) High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J Chem Phys* **22**(8): 1420–1426.
3. Tembe BL, McCammon JA. (1984) Ligand-receptor interactions. *Comput Chem* **8**(4): 281.
4. Bash PA, Singh UC, Brown FK, Langridge R, Kollman PA. (1987) Calculation of the relative change in binding free energy of a protein-inhibitor complex. *Science* **235**(4788): 574–576.
5. Gilson M, Zhou H. (2007) Calculation of protein-ligand binding affinities. *Ann Rev Biophys Biomol Struct* **36**: 21–42.
6. Rodinger T, Pomès R. (2005) Enhancing the accuracy, the efficiency and the scope of free energy simulations. *Curr Opin Struct Biol* **15**(2): 164–170.

7. Brandsdal BO, Österberg F, Almlöf M, *et al.* (2003) Free energy calculations and ligand binding. *Adv Protein Chem* **66**: 123–158.
8. van Gunsteren WF, Daura X, Mark AE. (2002) Computation of free energy. *Helv Chim Acta* **85**: 3113.
9. Simonson T, Archontis G, Karplus M. (2002) Free energy simulations come of age: protein-ligand recognition. *Acc Chem Res* **35**: 430–437.
10. van Gunsteren WF, Beutler TC, Fraternali F, *et al.* (1993) Computation of free energy in practice: choice of approximations and accuracy limiting factors. In WF van Gunsteren, PK Weiner, AJ Wilkinson (eds.), *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*, pp. 315–348. Escom Science Publishers, Leiden, The Netherlands.
11. Kollman P. (1993) Free energy calculations: applications to chemical and biochemical phenomena. *Chem Rev* **93**: 2395–2417.
12. Chipot C, Pohorille A. (2007) *Free Energy Calculations, Theory and Applications in Chemistry and Biology.* Springer, Berlin.
13. Trzesniak D, Kunz A-PE, van Gunsteren WF. (2007) A comparison of methods to compute the potential of mean force. *Chem Phys Chem* **8**: 162–169.
14. Ytreberg FM, Swendsen RH, Zuckerman DM. (2006) Comparison of free energy methods for molecular systems. *J Chem Phys* **125**: 184114.
15. Shirts MR, Pande VS. (2005) Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *J Chem Phys* **122**: 144107.
16. Rodriguez-Gomez D, Darve E, Pohorille A. (2004) Assessing the efficiency of free energy calculation methods. *J Chem Phys* **120**(8): 3563–3578.
17. Frenkel D, Smit B. (2002) *Understanding Molecular Simulation: from Algorithms to Applications.* Academic Press, San Diego.
18. Bennett CH. (1976) Efficient estimation of free energy differences from Monte Carlo data. *J Comput Phys* **22**(2): 245–268.
19. Pitera JW, van Gunsteren WF. (2002) A comparison of non-bonded scaling approaches for free energy calculations. *Mol Simul* **28**(1): 45–65.
20. Beutler TC, Mark AE, van Schaik RC, Gerber PB, van Gunsteren WF. (1994) Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem Phys Lett* **222**(6): 529–539.
21. Straatsma TP, McCammon JA. (1991) Multiconfiguration thermodynamic integration. *J Chem Phys* **95**(2): 1175–1188.
22. Fasnacht M, Swendsen RH. (2004) Adaptive integration method for Monte Carlo simulations. *Phys Rev E* **69**: 056704.
23. Tembre BL, McCammon JA. (1984) Ligand-receptor interactions. *Comput Chem* **8**(4): 281.
24. Michielin O, Karplus M. (2002) Binding free energy differences in a TCR-peptide-MHC complex induced by a peptide mutation: a simulation analysis. *J Mol Biol* **324**(3): 547–569.

25. Oostenbrink C, van Gunsteren WF. (2005) Free energies of ligand binding for structurally diverse compounds. *PNAS* **102**(19): 6750–6754.
26. Gilson MK, Given JA, Bush BL, McCammon JA. (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J* **72**: 1047–1069.
27. Boresch S, Tettinger F, Leitgeb M. (2003) Absolute binding free energies: a quantitative approach for their calculation. *J Phys Chem B* **107**: 9535–9551.
28. Jorgensen WL, Buckner JK, Boudon S, Tirado-Rives J. (1988) Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *J Chem Phys* **89**(6): 3742–3746.
29. Roux B, Nina M, Pomes R, Smith JC. (1996) Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study. *Biophys J* **71**: 670–681.
30. Woo H-J, Roux B. (2005) Calculation of absolute protein-ligand binding free energy from computer simulations. *PNAS* **102**(19): 6825–6830.
31. Wang J, Deng Y, Roux B. (2006) Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys J* **91**: 2798–2814.
32. Torrie GM, Valleau JP. (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J Comput Phys* **23**(2): 187–199.
33. Northrup SH, Pear MR, Lee C-Y, McCammon JA, Karplus M. (1982) Dynamical theory of activated processes in globular proteins. *Proc Am Math Soc* **79**: 4035.
34. Roux B. (1995) The calculation of the potential of mean force using computer simulations. *Comput Phys Commun* **91**: 275–282.
35. Ferrenberg AM, Swendsen RH. (1989) Optimized Monte Carlo data analysis. *Phys Rev Lett* **63**(12): 1195–1198.
36. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM. (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The methods. *J Comput Chem* **13**(8): 1011–1021.
37. Gullingsrud JR, Braun R, Schulten K. (1999) Reconstructing potentials of mean force through time series analysis of steered molecular dynamics simulations. *J Comput Phys* **151**(1): 190–211.
38. Boczko EM, Brooks III CL. (1993) Constant-temperature free energy surfaces for physical and chemical processes. *J Phys Chem* **97**: 4509–4513.
39. Bernèche S, Roux B. (2001) Energetics of ion conduction through the K+ channel. *Nature* **414**: 73–77.
40. Bartels C, Karplus M. (1998) Probability distributions for complex systems: adaptive umbrella sampling of the potential energy. *J Phys Chem B* **102**: 865–880.

41. Haydock C, Sharp JC, Prendergast FG. (1990) Tryptophan-47 rotational isomerization in variant-3 scorpion neurotoxin. A combination thermodynamic perturbation and umbrella sampling study. *Biophys J* **57**(6): 1269–1279.
42. Souaille M, Roux B. (2000) Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput Phys Commun* **135**: 40–57.
43. Kästner J, Thiel W. (2005) Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "umbrella integration". *J Chem Phys* **123**: 144104.
44. Carter EA, Ciccotti G, Hynes JT, Kapral R. (1989) Constrained reaction coordinate dynamics for the simulation of rare events. *Chem Phys Lett* **156**(5): 472–477.
45. Sprik M, Ciccotti G. (1998) Free energy from constrained molecular dynamics. *J Chem Phys* **109**(18): 7737–7744.
46. den Otter WK, Briels WJ. (1998) The calculation of free-energy differences by constrained molecular-dynamics simulations. *J Chem Phys* **109**(11): 4139–4146.
47. Ryckaert JP, Ciccotti G, Berendsen HJC. (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**(3): 327–341.
48. Christen M, van Gunsteren WF. (2007) On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: a review. *J Comput Chem*, **in print**.
49. Darve E, Pohorille A. (2001) Calculating free energies using average force. *J Chem Phys* **115**(20): 9169–9183.
50. Darve E, Wilson MA, Pohorille A. (2002) Calculating free energies using a scaled-force molecular dynamics algorithm. *Mol Simul* **28**(1): 113–144.
51. Hénin J, Chipot C. (2004) Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J Chem Phys* **121**(7): 2904–2914.
52. Rosso L, Minary P, Zhu Z, Tuckerman ME. (2002) On the use of adiabatic molecular dynamics to calculate free energy profiles. *J Chem Phys* **116**: 4389–4402.
53. Laio A, Parrinello M. (2002) Escaping free-energy minima. *PNAS* **99**(20): 12562–12566.
54. Huber T, Torda AE, van Gunsteren WF. (1994) Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J Comput Aided Mol Des* **8**: 695–708.
55. Evans DJ, Cohen EGD, Morriss GP. (1993) Probability of second law violations in shearing steady states. *Phys Rev Lett* **71**(15): 2401–2404.
56. Evans DJ, Searles DJ. (2002) The fluctuation theorem. *Adv Phys* **51**(7): 1529–1585.

57. Bustamante C, Liphardt J, Ritort F. (2005) The nonequilibrium thermodynamics of small systems. arXiv:cond-mat/0511629.
58. Jarzynski C. (1997) A nonequilibrium equality for free energy differences. *Phys Rev Lett* **78**(14): 2690–2693.
59. Jarzynski C. (1997) Equilibrium free energy differences from nonequilibrium measurements: a master equation approach. *Phys Rev E* **56**: 5018–5035.
60. Crooks GE. (1999) Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys Rev E* **60**(3): 2721–2726.
61. Jarzynski C. (2000) Hamiltonian derivation of a detailed fluctuation theorem. *J Stat Phys* **98**(1): 77–102.
62. Evans DJ. (2003) A nonequilibrium free energy theorem for deterministic systems. *Mol Phys* **101**(10): 1551–1553.
63. Schurr JM, Fujimoto BS. (2003) Equalities for the nonequilibrium work transferred from an external potential to a molecular system. Analysis of single-molecule extension experiments. *J Phys Chem B* **107**(50): 14007–14019.
64. Cuendet MA. (2006) Statistical mechanical derivation of Jarzynski's identity for thermostated non-Hamiltonian dynamics. *Phys Rev Lett* **96**(12): 120602.
65. Cuendet MA. (2006) The Jarzynski identity derived for general non-Hamiltonian and Hamiltonian dynamics generating the NVT or NPT ensembles. *J Chem Phys* **125**: 144109.
66. Liphardt J, Dumont S, Smith SB, Tinoco Jr I, Bustamente C. (2002) Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's identity. *Science* **296**: 1832–1835.
67. Park S, Schulten K. (2004) Calculating potentials of mean force from steered molecular dynamics information. *J Chem Phys* **120**(13): 5946–5961.
68. Hummer G, Szabo A. (2005) Free energy surfaces from single-molecule force spectroscopy. *Acc Chem Res* **38**(7): 504–513.
69. Hummer G, Szabo A. (2001) Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *PNAS* **98**(7): 3658–3661.
70. Ytreberg FM, Zuckerman DM. (2004) Efficient use of nonequilibrium measurement to estimate free energy differences for molecular systems. *J Comput Chem* **25**(14): 1749 –1759.
71. Sun SX. (2003) Equilibrium free energies from path sampling of nonequilibrium trajectories. *J Chem Phys* **118**(13): 5769–5775.
72. Jarzynski C. (2006) Rare events and the convergence of exponentially averaged work values. *Phys Rev E* **73**: 046105.
73. Zuckerman DM, Woolf TB. (2004) Systematic finite-sampling inaccuracy in free energy differences and other nonlinear quantities. *J Stat Phys* **114**(5): 1303.
74. Oberhofer H, Dellago C, Geissler PL. (2005) Biased sampling of nonequilibrium trajectories: can fast switching simulations outperform conventional free energy calculation methods? *J Phys Chem B* **109**: 6902.

75. Procacci P, Marsili S, Barducci A, Signorini GF, Chelli R. (2006) Crooks equation for steered molecular dynamics using a Nosé-Hoover thermostat. *J Chem Phys* **125**: 164101.
76. Chelli R, Marsili S, Barducci A, Procacci P. (2007) Recovering the Crooks equation for dynamical systems in the isothermal-isobaric ensemble: a strategy based on the equations of motion. *J Chem Phys* **126**: 044502.
77. Collin D, Ritort F, Jarzynski C, Smith SB, Tinoco Jr I, Bustamante C. (2005) Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature* **437**(8): 231–234.
78. Crooks GE. (2000) Path-ensemble averages in systems driven far from equilibrium. *Phys Rev E* **61**: 2361.
79. Shirts R, Bair E, Hooker G, Pande VS. (2003) Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods. *Phys Rev Lett* **91**(14): 140601.
80. Swanson JM, Henchman RH, McCammon JA. (2004) Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys J* **86**(1 Part 1): 67–74.
81. Åqvist J, Medina C, Samuelsson JE. (1994) A new method for predicting binding affinity in computer-aided drug design. *Protein Eng* 7(3): 385–391.
82. Srinivasan J, Cheatham TE, Cieplak P, Kollman PA, Case DA. (1998) Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate — DNA helices. *J Am Chem Soc* **120**(37): 9401–9409.
83. Kollman PA, Massova I, Reyes C, *et al.* (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res* **33**(12): 889–897.
84. Hansson T, Marelius J, Åqvist J. (1998) Ligand binding affinity prediction by linear interaction energy methods. *J Comput Aided Mol Des* **12**(1): 27–35.
85. Åqvist J, Luzhkov VB, Brandsdal BO. (2002) Ligand binding affinities from MD simulations. *Acc Chem Res* **35**(6): 358–365.
86. Carlson HA, Jorgensen WL. (1995) An extended linear-response method for determining free-energies of hydration. *J Phys Chem* **99**(26): 10667–10673.
87. Lee FS, Chu ZT, Bolger MB, Warshel A. (1992) Calculations of antibody antigen interactions — microscopic and semimicroscopic evaluation of the free-energies of binding of phosphorylcholine analogs to McPC603. *Protein Eng* **5**(3): 215–228.
88. Åqvist J, Hansson T. (1996) On the validity of electrostatic linear response in polar solvents. *J Phys Chem* **100**(22): 9512–9521.
89. Ljungberg KB, Marelius J, Musil D, Svensson P, Norden B, Åqvist J. (2001) Computational modeling of inhibitor binding to human thrombin. *Eur J Pharm Sci* **12**(4): 441–446.

90. Zhou RH, Friesner RA, Ghosh A, Rizzo RC, Jorgensen WL, Levy RM. (2001) New linear interaction method for binding affinity calculations using a continuum solvent model. *J Phys Chem B* **105**(42): 10388–10397.
91. Åqvist J, Marelius J. (2001) The linear interaction energy method for predicting ligand binding free energies. *Comb Chem High Throughput Screen* **4**(8): 613–626.
92. Foloppe N, Hubbard R. (2006) Towards predictive ligand design with free-energy based computational methods? *Curr Med Chem* **13**(29): 3583–3608.
93. Almlof M, Brandsdal BO, Åqvist J. (2004) Binding affinity prediction with different force fields: examination of the linear interaction energy method. *J Comput Chem* **25**(10): 1242–1254.
94. Carlsson J, Ander M, Nervall M, Åqvist J. (2006) Continuum solvation models in the linear interaction energy method. *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys* **110**(24): 12034–12041.
95. Almlof M, Åqvist J, Smalas AO, Brandsdal BO. (2006) Probing the effect of point mutations at protein-protein interfaces with free energy calculations. *Biophys J* **90**(2): 433–442.
96. Brandsdal BO, Åqvist J, Smalas AO. (2001) Computational analysis of binding of P1 variants to trypsin. *Protein Sci* **10**(8): 1584–1595.
97. Gilson MK, Honig BH. (1988) Energetics of charge-charge interactions in proteins. *Proteins* **3**(1): 32–52.
98. Gilson MK, Honig BH. (1988) Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins* **4**(1): 7–18.
99. Still WC, Tempczyk A, Hawley RC, Hendrickson T. (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* **112**: 6127–6129.
100. Gohlke H, Kiel C, Case DA. (2003) Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J Mol Biol* **330**(4): 891–913.
101. Lee MS, Feig M, Salsbury Jr. FR, Brooks III CL. (2003) New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J Comput Chem* **24**(11): 1348–1356.
102. Lee MS, Salsbury Jr. FR, Brooks III CL. (2002) Novel generalized Born methods. *J Chem Phys* **116**(24): 10606–10614.
103. Zoete V, Meuwly M, Karplus M. (2005) Study of the insulin dimerization: binding free energy calculations and per-residue free energy decomposition. *Proteins* **61**(1): 79–93.
104. Amidon GL, Yalkowsky SH, Anik ST, Valvani SC. (1975) Solubility of nonelectrolytes in polar solvents. V. Estimation of the solubility of aliphatic

monofunctional compounds in water using a molecular surface area approach. *J Phys Chem* **79**: 2239–2246.

105. Hermann RB. (1972) Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area. *J Phys Chem* **76**: 2754–2759.
106. Sitkoff D, Sharp KA, Honig B. (1994) Accurate calculation of hydration free energies using macroscopic solvent models. *J Phys Chem* **98**: 1978–1988.
107. Jayaram B, Sprous D, Beveridge DL. (1998) Solvation free energy of biomacromolecules: parameters for a modified generalized Born model consistent with the amber force field. *J Phys Chem B* **102**: 9571–9576.
108. Gohlke H, Kuhn LA, Case DA. (2004) Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* **56**(2): 322–337.
109. McQuarrie DA. (1976) *Statistical Mechanics.* Harper & Row, New York.
110. Tidor B, Karplus M. (1994) The contribution of vibrational entropy to molecular association. *J Mol Biol* **238**: 405–414.
111. Moreira IS, Fernandes PA, Ramos MJ. (2007) Computational alanine scanning mutagenesis — an improved methodological approach. *J Comput Chem* **28**: 644–654.
112. Rizzo RC, Toba S, Kuntz ID. (2004) A molecular basis for the selectivity of thiadiazole urea inhibitors with stromelysin-1 and gelatinase-A from generalized Born molecular dynamics simulations. *J Med Chem* **47**(12): 3065–3074.
113. Pearlman DA. (2005) Evaluating the molecular mechanics Poisson-Boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. *J Med Chem* **48**(24): 7796–7807.
114. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M. (2005) Validation and use of the MM-PBSA approach for drug discovery. *J Med Chem* **48**(12): 4040–4048.
115. Huo S, Massova I, Kollman PA. (2002) Computational alanine scanning of the 1 : 1 human growth hormone-receptor complex. *J Comput Chem* **23**(1): 15–27.
116. Massova I, Kollman PA. (1999) Computational alanine scanning to probe protein-protein interactions: a novel approach to evaluate binding free energies. *J Am Chem Soc* **121**(36): 8133–8143.
117. Zoete V, Michielin O. (2007) Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein-protein association using MM-GBSA: application to the TCR-p-MHC complex. *Proteins* **67**(4): 1026–1047.
118. Lafont V, Schaefer M, Stote RH, Altschuh D, Dejaegere A. (2007) Protein-protein recognition and interaction hot spots in an antigen-antibody

complex: free energy decomposition identifies “Efficient amino acids”. *Proteins* **67**(2): 418–434.

119. Zoete V, Meuwly M. (2006) Importance of individual side chains for the stability of a protein fold: computational alanine scanning of the insulin monomer. *J Comput Chem* **27**(15): 1843–1857.
120. Yip YL, Zoete V, Scheib H, Michielin O. (2006) Structural assessment of single amino acid mutations: application to TP53 function. *Hum Mutat* **27**(9): 926–937.
121. Michalik L, Zoete V, Krey G, Grosdidier A, Gelman L, Chodanowski P, Felge JN, Desvergne B, Wahli W, Michielin O. (2007) Combined simulation and mutagenesis analyses reveal the involvement of key residues for peroxisome proliferator-activated receptor alpha helix 12 dynamic behavior. *J Biol Chem* **282**(13): 9666–9677.

Chapter 20

Structure-based Computational Pharmacology and Toxicology

Angelo Vedani*,†,‡ and Martin Smiesko†,‡

20.1 Introduction

Structure-based design — the tailoring of a small-molecule ligand to the three-dimensional topology of the binding pocket of a target protein — is doubtless a powerful concept in drug discovery. The prerequisite is the availability of the 3D structure of the macromolecular target at atomic resolution, preferably with a bound ligand molecule. Although a wealth of computational approaches exist to perform the individual steps of the task (e.g. see Refs. 1, 2 and selected chapters of this book), two obstacles would still seem to jeopardize the otherwise sound approach. First, the difficulty in quantifying the binding affinity of a ligand-protein complex from its 3D structure, particularly when metal ions are involved in the binding process or when solvent stripping contributes significantly. Second, induced-fit — the ligand-induced conformational adaptation of the protein to the topology of the ligand molecule — is still far from being accessible to simulations, particularly when the phenomenon exceeds local dimensions

*Corresponding author.

†Biographics Laboratory 3R, Friedensgasse 35, 4056 Basel, Switzerland. Email: angelo@biograf.ch.

‡Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50, 4056 Basel, Switzerland.

and involves larger domains or affects the quaternary structure. The "insolvable problem" underlying both aspects is the need to extract small differences from large numbers (energies), the calculation of which is afflicted with errors. Those are particularly associated with solvation phenomena (ligand desolvation, solvent stripping, change in protonation status), binding to metal centers, and most important, the quantification of induced fit. Unfortunately, these errors are in the range of 5–10 kcal/mol; for induced-fit exceeding local dimensions, most likely significantly higher. In drug discovery, however, structurally similar molecules must be energetically discriminable, i.e. the accuracy in the calculated binding affinity should be within 0.9–1.4 kcal/mol, corresponding to a factor of 5–10 in the K_i or IC_{50} value.

The quantification of binding affinities has been approached from several ends. Free-energy perturbation techniques allow for precise estimates of the quantity, but are limited to small structural changes of two molecules.[1] Scoring functions are very effective for a fast (but at best qualitative) classification of a larger series of potential lead candidates.[3] Quantitative structure-activity relationships (QSAR) are the most frequently used approach as it allows for a fast and quantitative determination of the binding affinity based on linear or multiple-regression techniques.[4–6]

QSAR is an area of computational research that builds atomistic or virtual models to predict quantities such as the binding affinity or the toxic potential of existing or hypothetical molecules. In drug discovery, quantitative structure-activity relationships are widely used to identify affine ligands for a given macromolecular target. QSAR concepts are ubiquitously applied in pharmaceutical R&D worldwide — particularly for lead identification and optimization. More recently, the technology has been extended to predict ADMET properties (adsorption, distribution, metabolism, elimination, toxicity)[7] or the oral bioavailability.[8] In the context of the REACH legislation (Registration, Evaluation and Authorization of Chemicals) of the European Union, the prediction of the toxic potential of a drug or environmental chemical *in silico* has spawned much interest.[9–10]

Half a decade after its introduction, QSAR has matured into a computational tool, substantially contributing to the drug-discovery process. Originally based on the philosophy that compounds with similar physico-chemical properties trigger comparable biological effects, QSARs are nowadays frequently employed to establish a correlation between structural properties of potential drug candidates and their binding affinity (K_i, IC_{50}) towards a common macromolecular target. With the large number of 3D structures of proteins available from X-ray diffraction studies,[10] structure-based design has become a powerful tool. With an appropriately parameterized force field, it became possible to identify the binding mode of any given existing or hypothetical molecule to a macromolecular bioregulator of interest. Unfortunately, this classic approach is limited by the aforementioned problems. In addition, ligand desolvation (solvent stripping) can turn out to be a dominant factor — as a small molecule typically binds from an aqueous environment (blood and body fluid) to a target protein, which is at least partially hydrophobic in nature — particularly if the dielectric properties of the binding pocket trigger a change in the ligand's protonation state. Apart from the cost of induced-fit and internal strain (a ligand, frequently binds in a conformation different from its low-energy state in aqueous solution), entropic contributions to the binding energy are difficult to estimate.[11]

The introduction of *Comparative Molecular Field Analysis* (CoMFA) in 1988 represents a milestone in QSAR, as for the first time, structure-activity relationships were based on the three-dimensional structure of the ligand molecules (3D-QSAR). In CoMFA, steric and electrostatic properties of the target protein were mapped onto a surface or grid, surrounding a series of compounds superimposed in their bioactive conformation.[12] This surface or grid represents a surrogate of the binding site of the true biological receptor; in a pharmacological context frequently referred to as pharmacophore. Apart from the diversity of the employed data set, the quality of the map depends on the correct superposition of the ligands, the identification of which is almost impossible in the absence of structural information for the target protein. While this problem has long been recognized, only the more recently developed 4D-QSAR technologies would seem to provide decent solutions.[13–15] The calculation of binding energies in QSAR studies is by no means simple, as the determining quantities can hardly be calculated in the absence of the 3D structure of the target protein. Of course, sufficient variables given, any quantity can be "reproduced," but what would the predictive power of such a model be?

20.2 Multi-Dimensional QSAR

In 3D-QSAR, bioactive conformation and relative orientations of the ligand molecules must be unambiguously identified in order to generate a predictive model. Particularly, in the absence of structural information on the target protein, the identification of both bioactive conformation and orientation of the ligand molecule is all but simple. If the 3D structure of the macromolecular target is known, it can be used for this very purpose. Ligand docking is preferably achieved using automated, flexible fitting, as this allows for local induced fit, the rearrangement of the protein side-chains lining the binding pocket. More recent approaches allow for dynamic solvation, i.e. the protein-ligand complex is subjected to a solvation algorithm — explicitly, by placing water molecules in geometrically feasible positions, and hence, allowing for solvent-mediated protein-ligand hydrogen bonds[16] or implicitly, by calculating the solvation effect based on the solvent-accessible surface area.[17]

In 4D-QSAR, energetically feasible binding modes are composed into a 4D data set, including different ligand conformations, poses, and protonation states.[13–15] The true binding mode (or the bioactive conformation) is then identified by the algorithm underlying the QSAR concept (genetic, neural networks or combinations thereof). Particularly in the absence of structural information on the macromolecular target, a 4D representation of the ligand molecules reduces the bias associated with their alignment. But even with the 3D structure of the target protein at hand, the identification of the binding mode is not trivial. First, the electron density at modest resolution (2.0 Å or higher) is not sharp enough to reveal all the details. Here, the deposited coordinates represent a mixture of experiment and modeling.[18] Second, when docking compounds not too anisotropic in shape, different poses with comparable energies will always be identified.

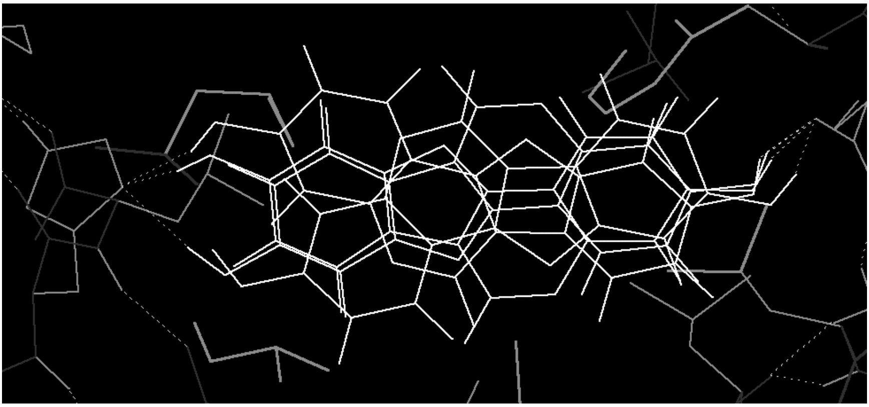


Fig. 20.1 Multiple binding modes of coumestrol to the ERα (PDB code 3ERD), as identified by automated, flexible docking.[16] The ligands are shown in white, the protein in gray; hydrogen bonds are indicated as dashed lines. Key interactions include the hydrogen bonds to His524 (on the left) and Glu353 (on the right).

Figure 20.1 shows four possible arrangements of coumestrol at the estrogen receptor α as identified by flexible docking: all arrangements are within 0.8 kcal/mol of the lowest-energy pose.[16]

Even if the structure of the ERα-coumestrol complex (Fig. 20.1) is determined by X-ray diffraction techniques, it is unlikely that the true pose(s) could be identified, as the root-mean-square (rms) deviation for two poses is smaller than 1.0 Å and smaller than 2.0 Å for the others. In addition, coumestrol could well bind in different ways. Such systems indicate a serious limitation of 3D-QSAR approaches, where a ligand molecule can only be represented by a single entity. This conceptual flaw was only corrected by the more recently developed 4D-QSAR technologies,[13–15] where each ligand molecule may be represented by an ensemble of conformations, poses (different orientations), tautomeric forms, protonation states, and stereoisomers.

4D-QSAR can be interpreted as a feasible extension of 3D-QSAR to address the uncertainties during the alignment process. It has, however, fundamental biological relevance, when dealing with multimode binding targets. Cytochrome P450 enzymes, for example, are known to accommodate a ligand in various binding poses, yielding different metabolic products of a given compound. 4D-QSAR technologies can explicitly account for different ligand configurations in a single simulation. Recently, this has been successfully applied to simulate binding of structurally diverse compounds to cytochrome P450 3A4, representing each small molecule with, on average, four different binding poses identified by an automated docking procedure (10 and Chapter 21).

Induced fit — the ligand-induced adaptation of the protein structure — may not only alter the topology of the binding pocket, but also its character: hydrophobic or hydrophilic, dielectric properties, solvent accessibility (Fig. 20.2). While a local manifestation of the phenomenon may be simulated by means of MD, the rearrangement of larger domains, or changes in the quaternary structure are not yet computationally accessible. Moreover, induced-fit energies estimated from MM/MD simulations are not suited for quantitative aspects, as they are associated with errors larger than the objective (see above). Here, the combination of protein modeling (identification of the binding mode and simulation of the induced fit) and mQSAR (multi-dimensional QSAR; quantification) would seem to offer an appropriate solution.

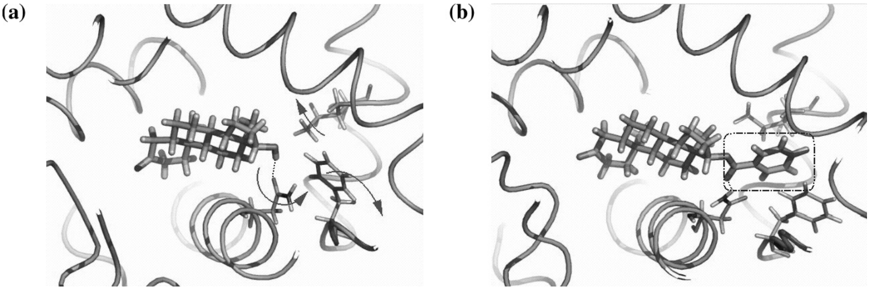


Fig. 20.2 Manifestation of induced fit: the steroid dihydrotestostosterone binding to the androgen receptor (left; PDB code 1I37). In order for the bulkier benzoate derivative to bind (right), the binding pocket undergoes a major rearrangement. Asn705 (previously accepting a hydrogen bond from the dihydrotestostosterone's OH group) flips by 180° and now acts as a H-bond donor towards the benzoate's carbonyl O atom; Phe891 and Leu873 both change from a trans to gauche conformation, thereby generating a small hydrophobic pocket, accommodating the aromatic portion of the benzoate.[20]

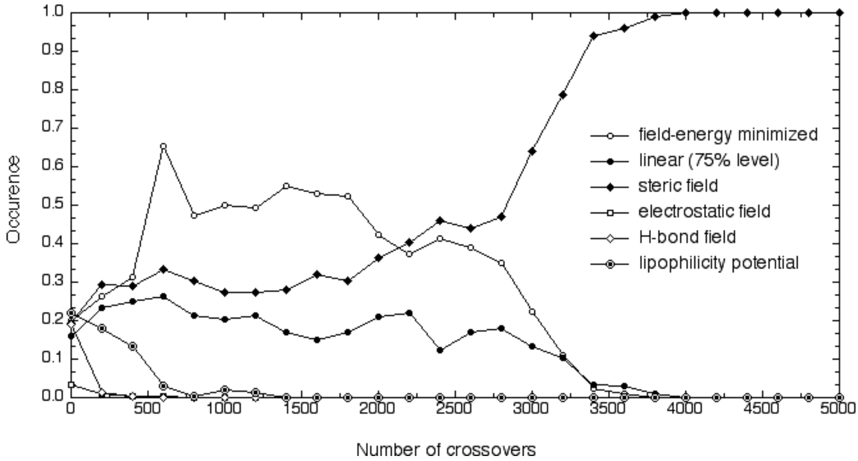


Fig. 20.3 Induced-fit crossover as observed in the simulation of dopamine β-hydroxylase. At 2000 crossovers (corresponding to 10 generations in a 200-model population), the prevailing induced-fit changes from a scenario based on a minimization protocol (loose fit) to one based on the steric potential of the compounds (tight fit).[29]

Particularly in the absence of structural information on the target protein (e.g. for GPCRs), the most realistic induced-fit scenario cannot be unambiguously identified. Consequently, a fifth dimension has been introduced (5D-QSAR), allowing for the simultaneous consideration of different manifestations thereof. Such models may not only accurately mimic the 3D topology of the binding pocket but also identify realistic induced-fit protocols, e.g. based on the steric, electrostatic, hydrogen-bonding, or lipophilic potential.[21] As the induced-fit scenario may well change throughout a simulation (Fig. 20.3), 5D-QSAR allows for a less biased approach. The employment of such protocols in receptor-surface modeling[21–23] yields surrogates of high predictive power for several proteins of biomedical interest (see below and in Refs. 24–28).

To simulate compounds that bind to different sub-pockets of the binding site as a consequence of induced fit and, hence, experience different fields, a dual-shell representation — able to anisotropically simulate induced fit (see Fig. 20.7) — has been devised.[23] Variations in the distribution of properties between the inner and outer shell are allowed. The adaptation of both field and topology of the receptor surrogate to each ligand is achieved by combining a steric adjustment to the 3D structure of the ligand and a component due to the attraction or repulsion between ligand and receptor model. The latter is obtained by correlating their physicochemical properties (hydrophobicity and hydrogen-bond propensity) in 3D space.[23,29]

Apart from the interaction with the target protein (including induced fit), the binding of a small-molecule ligand to a macromolecular target is strongly affected by solvation phenomena: ligand desolvation, solvent stripping, and proton transfer (Fig. 20.4). Here, 6D-QSAR — where different solvation models are considered simultaneously — allows for an even more realistic simulation of the binding process.[16] This can either be achieved explicitly where parts of the surface area are mapped with solvent properties, whereby position and size are optimized by the genetic algorithm, or implicitly. Here, the solvation terms (ligand desolvation and solvent stripping) are independently scaled for each different model within the surrogate family, reflecting varying solvent accessibility of the binding pocket.[16,29] A classification of QSAR concepts based on their dimensionality is given in Table 20.1.

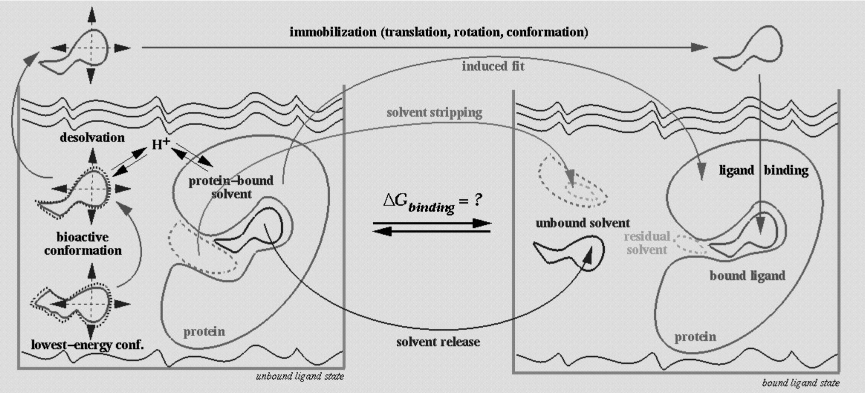


Fig. 20.4 Schematic view of the solvation phenomena associated with ligand binding.

Table 20.1 Classification of QSAR Concepts Based on their Dimensionality

Dimension	Method	Protein	References
1D-QSAR	Affinity is correlated with bulk properties of ligands (pK_a, log P, etc.)	no	5
2D-QSAR	Affinity is correlated with structural patterns (connectivity, 2D structure)	no	5
3D-QSAR	Affinity is correlated with the three-dimensional structure of the ligands	possible	4–6, 12
4D-QSAR	Ligands are represented as an ensemble of configurations	possible	13–15
5D-QSAR	as 4D-QSAR + representation of different induced-fit models or dual-shell representation of the receptor model	yes	22
6D-QSAR	as 5D-QSAR + representation of different solvation scenarios	yes	16

Quasar, a receptor-modeling concept developed at the Biographics Laboratory, is based on 6D-QSAR and explicitly allows for the simulation of induced fit.[16,22] It generates a family of quasi-atomistic receptor surrogates that are optimized by means of a genetic algorithm. The hypothetical receptor site is characterized by a 3D surface that surrounds the ligand molecules at van der Waals distance, and which is characterized by atomistic properties mapped onto it. The topology of this surface mimics the three-dimensional shape of the binding site and the mapped features represent other properties of interest, such as hydrophobicity, electrostatic potential and hydrogen-bonding propensity.[28] The fourth dimension in *Quasar* offers the possibility to represent each ligand molecule as an ensemble of conformations, orientations, tautomeric forms and protonation states.[15] Within this ensemble, the contribution of an individual entity to the total energy is determined by a normalized Boltzmann weight. As manifestation and magnitude of induced fit may vary for different ligands, the fifth dimension in *Quasar* allows for the simultaneous evaluation of up to six induced-fit protocols.[22] The most recent extension of the *Quasar* concept to six dimensions[16] allows for the simultaneous consideration of different solvation scenarios.

This can be achieved explicitly when parts of the surface area are mapped with solvent properties whereby position and size are optimized by the genetic algorithm. Alternatively, the solvation terms (ligand desolvation and solvent stripping) can be independently scaled for each different model within the surrogate family, reflecting varying solvent accessibility of the binding pocket (implicit approach). In *Quasar*, the binding energy is calculated according to Equation 20.1:

$$E_{\text{binding}} = E_{\text{ligand–receptor}} - E_{\text{ligand desolvation}} - E_{\text{ligand strain}} - T\Delta S - E_{\text{induced fit}} \quad (20.1)$$

where $E_{\text{ligand–receptor}} = E_{\text{electrostatic}} + E_{\text{van der Waals}} + E_{\text{hydrogen bonding}} + E_{\text{polarization.}}$

Raptor, an alternative receptor-modeling technology more recently developed at the Biographics Laboratory, is based on a fundamentally different scoring function and features a dual-shell representation of the receptor surrogate, which allows for an anisotropic simulation of induced fit.[23] During model generation, each shell is independently mapped with physicochemical properties (hydrophobic character and hydrogen-bonding propensity), a concept that permits changing the character of the binding pocket, e.g. upon binding of agonists or antagonists.[29] Induced fit is not limited to steric aspects but includes the variation of the fields — spawned by the physico-chemical properties mapped onto the two surfaces — along with it. The underlying scoring function for evaluating ligand-receptor interactions includes directional terms for hydrogen bonding, hydrophobicity, and thereby treats solvation effects implicitly. This makes the approach independent from a partial-charge model, and as a consequence, allows to model ligand molecules binding to the receptor with different net charges in a straightforward fashion. In *Raptor*, the binding energy is determined according to Equation 20.2:

$$E_{\text{binding}} = E_{\text{ligand–receptor}} - T\Delta S - E_{\text{induced fit}} \quad (20.2)$$

where $E_{\text{ligand–receptor}} = E_{\text{hydrogen bonding (shell 1)}} + E_{\text{hydrophobic (shell 1)}} + E_{\text{hydrogen bonding (shell 2)}} + E_{\text{hydrophobic (shell 2)}}$

20.3 Computational Pharmacology: Modeling GPCRs (Neurokinin-1, CCR-3, Bradykinin B_2 receptor)

The modeling of enzymes and receptors with known 3D structure (structure-based design), modeling by homology, and ligand-based concepts to computational pharmacology are covered in various chapters of this book. In this section, we present an approach for modeling G-protein coupled receptors (GPCRs) for which no experimental structures are presently available. Here, a technique, referred to as receptor modeling (formerly: receptor mapping) allows for the generation and validation of a three-dimensional receptor surrogate,[21–23] subsequently to be used in a structure-based design context.

Using the mQSAR technologies *Quasar*[16,22,28] and *Raptor*,[23,29] models for a series of G-protein coupled receptors have been validated at the 4D-level (Neurokinin-1 receptor, Ref. 15), 5D-level (Chemokine receptor-3, Ref. 30), and 6D-level (Bradykinin B_2 receptor, Ref. 31). As the 3D structure of these GPCRs is not available at atomic resolution, the ligand alignment was based on scaffold mapping for the NK-1 and CCR3 receptor, respectively, and using 3D/4D pharmacophore generation for the BB_2 receptor (software *Symposar*, Ref. 32).

The three-dimensional structures of all ligand molecules were generated using *MacroModel*[17] and optimized in aqueous solution by means of the AMBER* force field.[32] Atomic partial charges (MNDO/ESP) were calculated using the MOPAC package.[33] Next, the compounds were split into n training (NK1: $n = 50$, CCR3: $n = 106$, BB_2: $n = 147$) and m test ligands (NK1: $m = 15$, CCR3: $m = 35$, BB_2: $m = 139$) with the aim to obtain maximal structural diversity in the training set combined with a wide range in K_i or IC_{50}. First, the compounds were sorted according to their K_i value and the most active and the weakest-binding ligand defined as training compounds. Then — in descending order — test ligands were identified as those having each of their functional groups represented in the already defined subset of training ligands.

The *Quasar* simulations were based on a family of i receptor models (NK-1: $i = 500$, CCR3: $i = 250$, BB_2: $i = 250$) and evolved over j

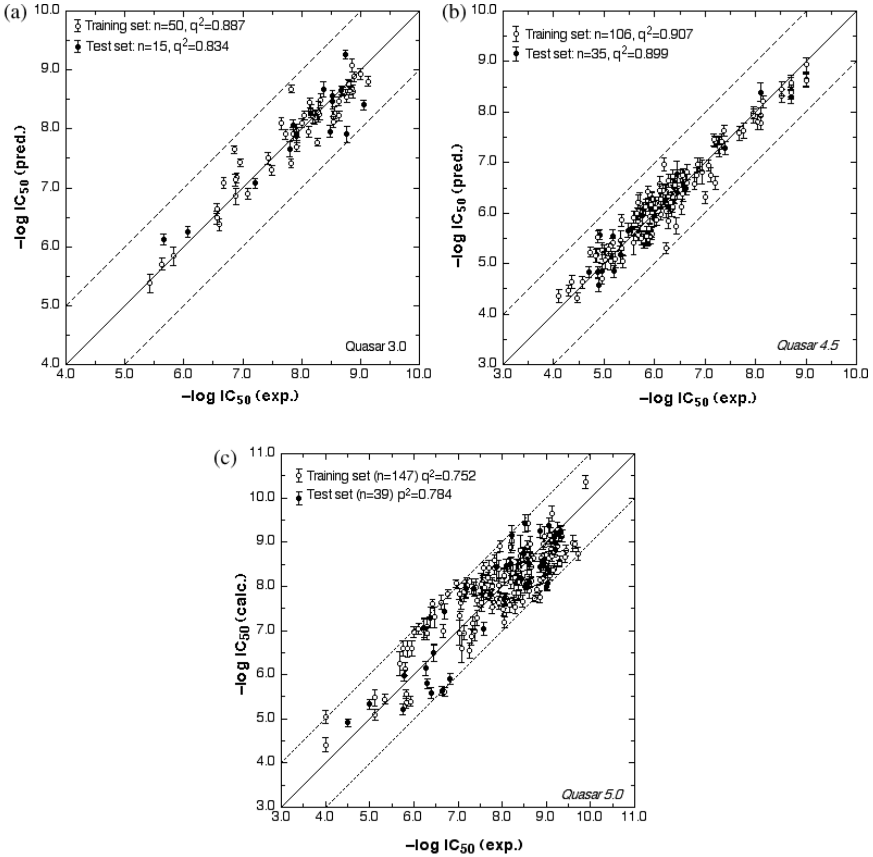


Fig. 20.5 Comparison of predicted and experimental IC_{50} values for the Neurokinin-1 receptor (left), the Chemokinne receptor-3 (center) and the Bradykinin B_2 receptor (right). The ligands of the training set are shown as open circles, while those of the test set are depicted as full circles. Dashed lines indicate the false-positive (upper) and false-negative threshold (lower), respectively.

crossover cycles (NK-1: $j = 40\,000$, CCR3: $j = 18\,000$, BB_2: $j = 50\,000$), corresponding to k generations (NK-1: $k = 80$, CCR3: $k = 72$, BB_2: $k = 200$). Predicted and experimental K_i values are compared in Fig. 20.5; key parameters are given in Table 20.2. Subsequently, a series of five to 10 scramble tests demonstrated the robustness of the models; details are given in Refs. 15, 30 and 31.

Table 20.2 Summary of the *Quasar* and *Raptor* Simulations for the GPCRs. q^2 = Cross-Validated r^2, p^2 = Predictive r^2; the rms and Maximal Deviation from the Experimental Binding Affinity is Given as a Factor in K_i.

System		Number of Compounds	q^2	rms Training	max. Training	p^2	rms Test	max. Test	Reference
NK-1:	*Quasar*	65	0.887	1.9	7.1	0.834	2.4	7.2	15
CCR3:	*Quasar*	141	0.907	1.0	7.1	0.899	0.8	3.5	30
BB_2:	*Quasar*	186	0.752	2.6	14.1	0.784	2.8	9.6	31
	Raptor		0.815	2.4	11.7	0.853	2.1	14.9	31

20.4 Computational Toxicology: Modeling Nuclear Receptors (Aryl Hydrocarbon, Estrogen α/β, Androgen, Thyroid α/β, PPARγ, Glucocorticoid Receptor)

Toxic agents, particularly those that exert their actions with a great deal of specificity, sometimes act via receptors to which they bind with high affinity. This phenomenon is referred to as receptor-mediated toxicity. Examples of soluble intracellular receptors, which are important in mediating toxic responses, include the glucocorticoid receptor, which is also involved in mediating toxicity associated effects such as apoptosis of lymphocytes as well as neuronal degeneration as a response to stress, the peroxisome proliferator-activated receptor, which is associated with hepatocarcinogenesis in rodents, and the aryl hydrocarbon receptor, which is involved in a whole range of toxic effects.[34] Harmful effects of drugs and chemicals can often be associated with their binding to other than their primary target — macromolecules involved in biosynthesis, signal transduction, transport, storage, and metabolism.[35–41]

Nuclear receptors comprise a family of ligand-dependent transcription factors that transform extra- and intra-cellular signals into cellular responses by triggering the transcription of target genes. In particular, they mediate the effects of hormones and other endogenous ligands to regulate the expression of specific genes. Among other members, this family includes receptors for the various steroid hormones, e.g. the estrogen, androgen, progesterone, and glucocorticoid receptor. Unbalanced production or cell insensitivity to specific hormones may result in diseases associated with human endocrine dysfunction.[42] The presence of hormonally active compounds — endocrine disruptors — in the biosphere has become a worldwide environmental concern. It has been concluded that such compounds elicit a variety of adverse effects in both humans and wildlife, including the promotion of hormone-dependent cancers, reproductive tract disorders, and a reduction in reproductive fitness. A number of receptor-mediated hormonal responses to toxicity are known, including xenobiotic effects on the thyroid hormone receptor, the epidermal growth factor receptor, the aryl hydrocarbon receptor as well as effects mediated by the androgen and the estrogen receptor, respectively. A variety of compounds in the environment have been shown to display agonistic or antagonistic activity towards the *ER*, including both natural products and synthetic compounds.[43–49] The concern over xenobiotics binding to the *ER* has created a need to both screen and monitor compounds that can modulate endocrine effects.[50–52]

Using the mQSAR technologies *Quasar*[16,22,28] and *Raptor*,[23,29] models for a series of eight nuclear receptors have been validated at the 6D-level. Except for the aryl hydrocarbon receptor, where the 3D structure of the protein is not available, the alignment was obtained using automated, flexible docking (software Yeti, see Refs. 50, 51). Again, the three-dimensional structures of all ligand molecules were generated using *MacroModel*[17] and optimized in aqueous solution by means of the AMBER* force field.[32] Atomic partial charges (MNDO/ESP) were calculated using the MOPAC package.[33] Next, the compounds were split into n training (Ah: $n = 105$, AR: $n = 86$, ER$\alpha\beta$: $n = 80$, TR$\alpha\beta$: $n = 66$, PPARγ: $n = 75$, GR: $n = 82$) and m test ligands (Ah: $m = 35$, AR: $n = 26$, ER$\alpha\beta$: $n = 26/23$, TR$\alpha\beta$: $n = 16$, PPARγ: $n = 20$, GR: $n = 28$). The splitting of the training and test set was performed as described above for the GPCRs.

The *Quasar* simulations were based on a family of i receptor models (Ah: $i = 250$, AR: $i = 200$, ER$\alpha\beta$: $i = 200$, TR$\alpha\beta$: $i = 200$, PPARγ: $i = 200$, GR: $i = 200$) and evolved over j crossover cycles (Ah: $j = 50\,000$, AR: $j = 10\,000$, ER$\alpha\beta$: $j = 32\,000$, TR$\alpha\beta$: $j = 20\,000$, PPARγ: $j = 20\,000$, GR: $j = 40\,000$), corresponding to k generations (Ah: k = 200, AR: $k = 50$, ER$\alpha\beta$: $k = 160$, TR$\alpha\beta$: $k = 100$, PPARγ: $k = 100$, GR: $k = 100$). Predicted and experimental K_i values are compared in Fig. 20.6, and

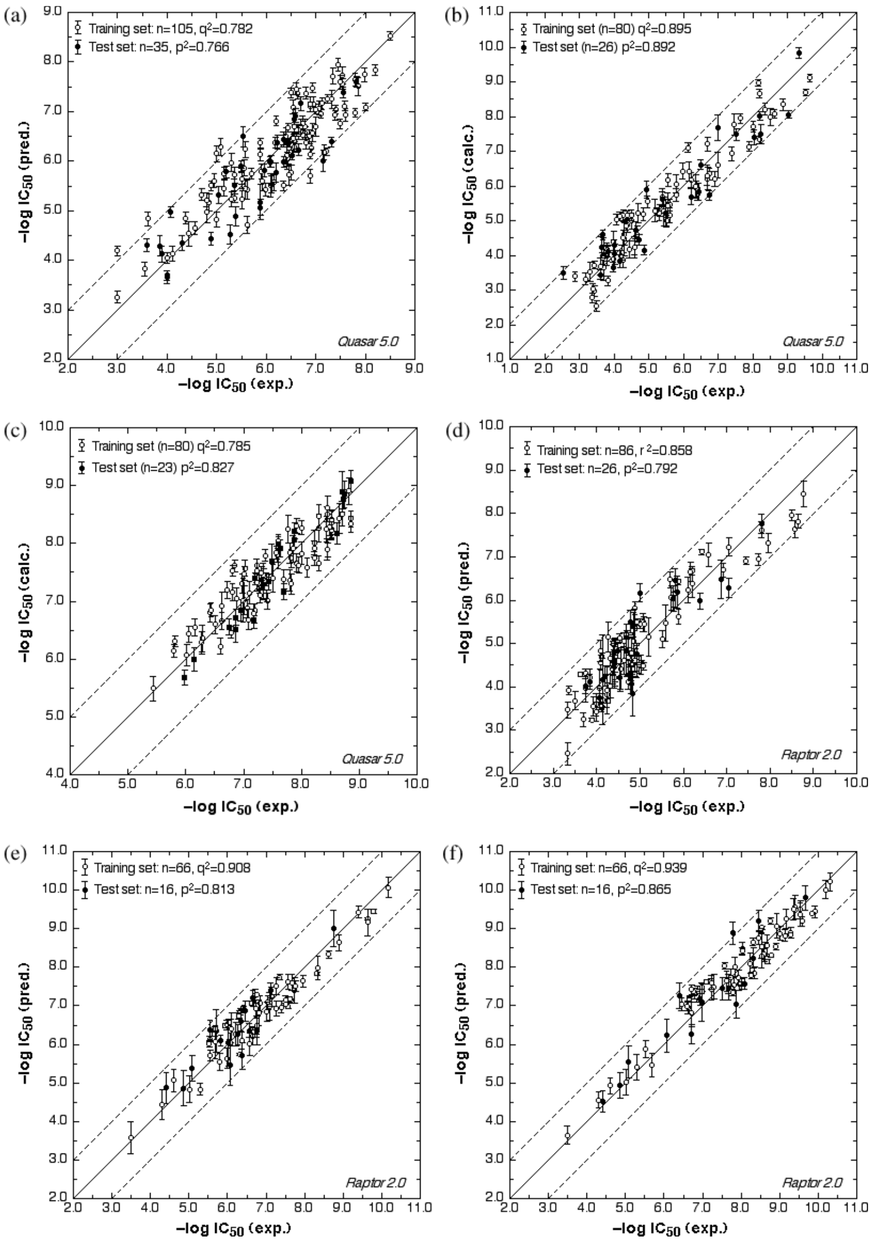


Fig. 20.6 Comparison of experimental and predicted binding affinities for the aryl hydrocarbon, estrogen α, estrogen β, androgen, thyroid α, thyroid β, peroxisome

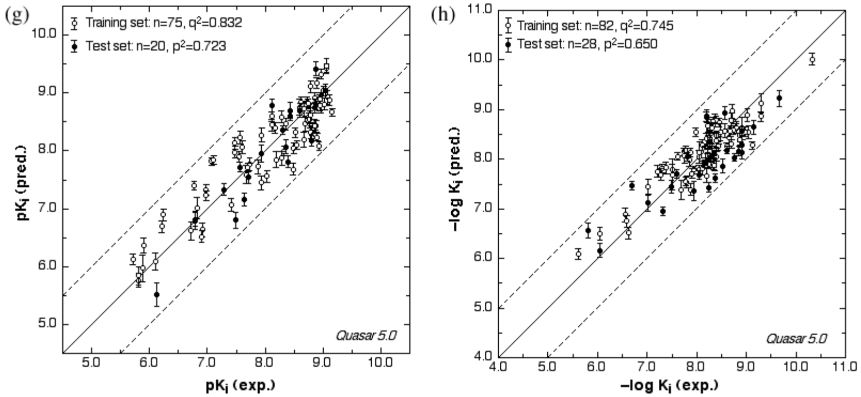


Fig. 20.6 (*Continued*) proliferator-activated γ, glucocorticoid receptor (top to bottom and left to right). Ligands of the training set are shown as open circles, and those of the test set as filled circles. Dashed lines indicate the false-positive (upper) and false-negative threshold (lower), respectively.

representative models are shown in Fig. 20.7; key parameters are compiled in Table 20.3. Subsequently, a series of five to 10 scramble tests demonstrated the robustness of the models (details are given in Refs. 16, 20, 24–27).

Other valuable modeling studies on nuclear receptors include those of Lukacova and Balaz[52] for the aryl hydrocarbon receptor; of van Lipzig *et al.*, Akahori *et al.*, da Cunha *et al.*, Kurunczi *et al.*, Sippl and Asikainen *et al.*[52–58] for the estrogen receptor, and of Hong *et al.*[59] for the androgen receptor.

20.5 Modeling Toxicity — The *VirtualToxLab* Concept

The receptor models for the eight nuclear receptors — along with the surrogate for the enzyme CYP450 3A4 (see (19) or Chapter 21) — represent the "virtual test kits" of the *VirtualToxLab* currently under compilation at the Biographics Laboratory 3R.[24] A pilot project using these models and a representative selection of 798 compounds (thereof, 188 substances were used to test the models) suggested that

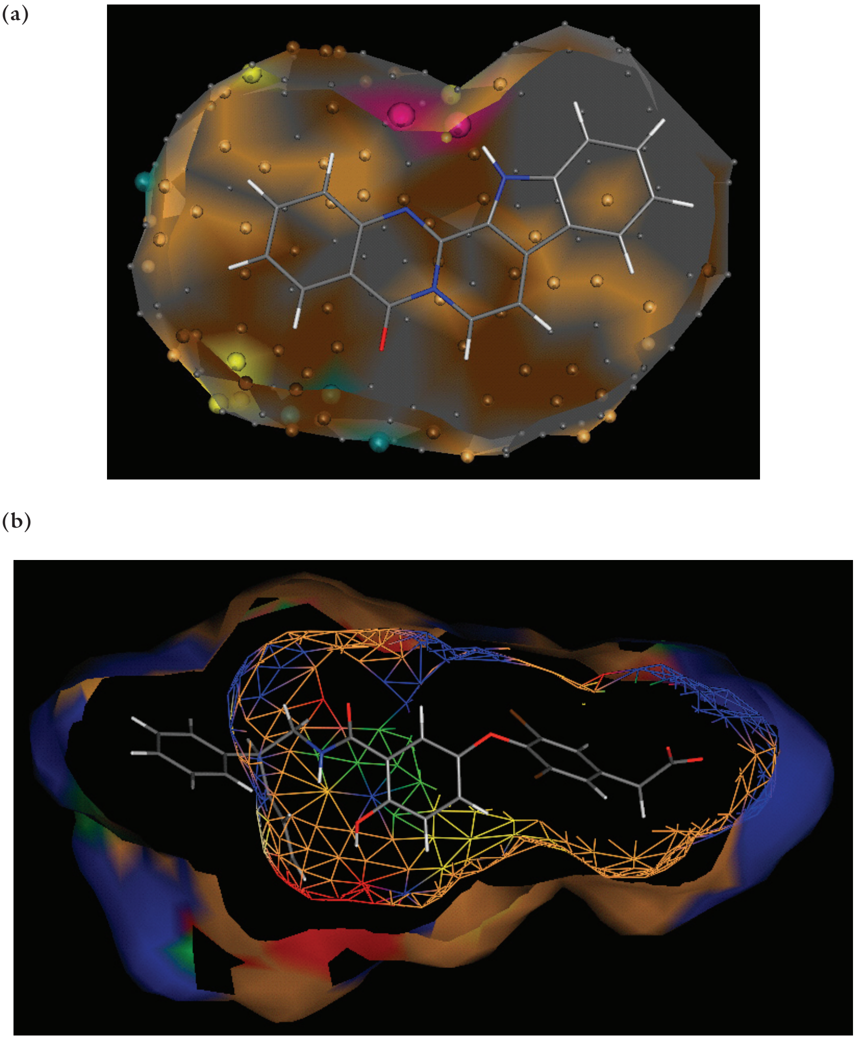


Fig. 20.7 *Quasar* model (6D-QSAR, left) and *Raptor* surrogate (dual-shell 5D-QSAR, right). The bound ligand is shown as a stick representation (atom coloring: gray = carbon, white = hydrogen, red = oxygen, blue = nitrogen). The quasi-atomistic properties of the receptor are mapped onto the surface(s): blue = positively charged salt bridge, red = negatively charged salt bridge; brownish colors = hydrophobic properties, pink = hydrogen-bond flip flop. The *Quasar* model represents the Aryl hydrocarbon receptor;[27] the *Raptor* model depicts the thyroid β receptor.[26]

Table 20.3 Summary of the *Quasar* and *Raptor* Simulations for the Nuclear Receptors. q^2 = Cross-Validated r^2, p^2 = Predictive r^2; the rms and Maximal Deviation from the Experimental Binding Affinity is Given as a Factor in K_i.

System		Number of Compounds	q^2	rms Training	max. Training	p^2	rms Test	max. Test	Reference
AhR:	*Quasar*	140	0.824	1.8	10.2	0.769	2.3	13.5	27
AR:	*Raptor*	114	0.858	1.7	7.8	0.792	1.6	13.9	20
ERα:	*Quasar*	106	0.895	2.0	8.6	0.892	2.9	9.5	16
ERβ:	*Quasar*	103	0.785	1.1	4.8	0.827	0.8	2.4	24
TRα:	*Raptor*	82	0.919	1.8	4.3	0.814	2.5	10.0	26
TRβ:	*Raptor*	82	0.909	2.0	7.7	0.796	2.7	8.8	26
PPARγ:	*Quasar*	95	0.832	1.4	6.2	0.723	1.4	3.9	25
GR:	*Quasar*	110	0.745	1.2	5.9	0.650	2.2	5.5	24

our approach is suited for the *in silico* identification of adverse effects triggered by drugs and chemicals: only six of the test compounds are predicted by more than a factor of 10 off the experimental binding affinity, with the maximal individual deviation not exceeding a factor of 15.[24,26] The flowchart of the database is depicted in Fig. 20.8; its validation status is continuously updated on our website.[60]

Up to date, our concept has not produced any truly false-positive results. At the current level, however, false-negative predictions are still obtained, as a compound of interest cannot be tested against all potential receptors it may bind to *in vivo* (some macromolecular targets will remain unknown, while for others no experimental structure exists, or too few affinity data are available to establish a QSAR). It is planned to extend the current concept by implementing a set of virtual filters, which can recognize benign compounds. Among others, criteria include the molecular weight, drug-like properties, and the presence or absence of characteristic structural motifs.

20.6 Future Outlook

The recent decade has experienced a subtle change in the focus of molecular-modeling approaches to both drug discovery and environmental issues, as poor pharmacokinetics and toxicity are not only

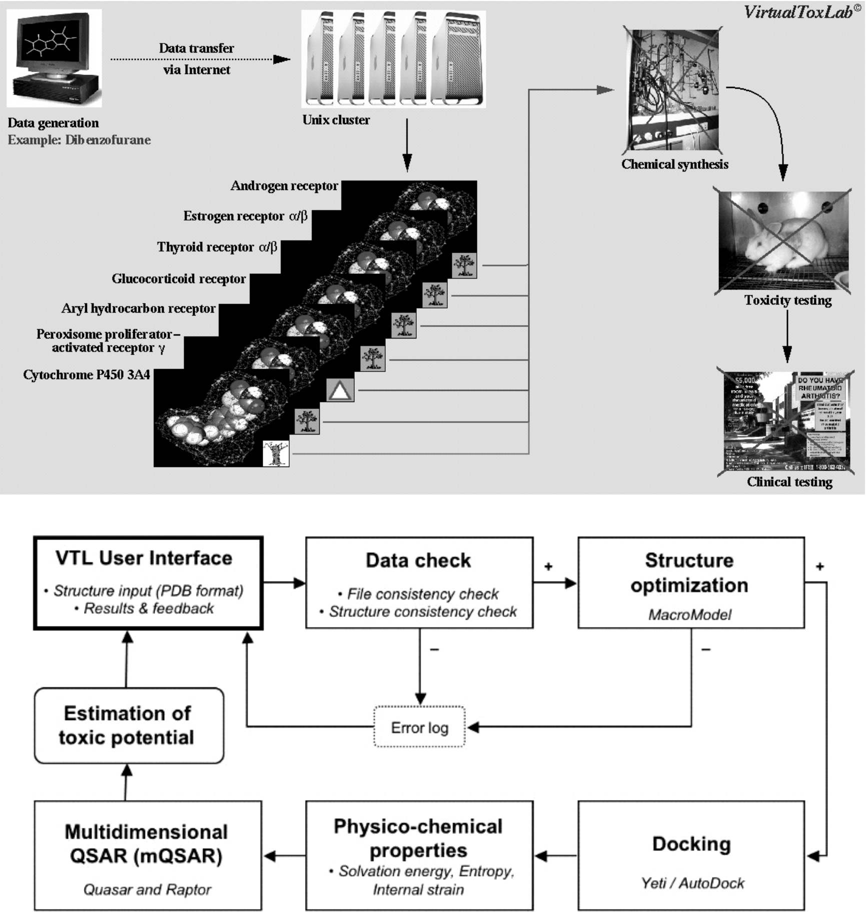


Fig. 20.8 Flowchart of the *VirtualToxLab*: pictorial (top) and schematic (bottom).

frequent causes of late-stage failures in drug development but also a source for unnecessary animal tests. It has been recognized in pharmaceutical R&D that ADMET (Adsorption, Delivery, Metabolism, Elimination, Toxicity) plays a key role in identifying safe drugs. The REACH initiative of the EU, which calls for the re-testing of some 30 000 chemicals with respect to their toxic potential, requires the scientific community to develop *in silico* concepts, allowing for fast and reliable screening of larger batches of drugs and environmental chemicals. While computational models for a dozen of proteins triggering or mediating adverse effects may be considered validated today, a marketable *in silico* approach must include more bioregulators, e.g. CYP2A13 and other representatives of this enzyme class, the pregnan-X, liver-X, mineralocorticoid, and the constituive androstane receptor. Metabolic products should also routinely be included in simulations addressing ADMET properties — a most difficult task as the formation of the various metabolic products depends on the physiological conditions. Aiming for a better prediction of the binding affinity of small-molecule ligands to macromolecular targets calls for a more rigorous treatment of effects frequently neglected in computational studies: induced fit, solvation, and entropy. Efforts should be undertaken to agree on *good modeling practices*, i.e. what criteria should be fulfilled to consider the model validated, and reaching farther, to have it accepted by the regulatory bodies (e.g. within REACH). Finally, the developed models should be shared — at least, if emerging from the academic community. The Biographics Laboratory 3R is prepared to provide free access to its *VirtualToxLab* for academic institutions and NPOs. Currently, the technology is under peer testing in selected laboratories; free access is planned for 2008.

References

1. Leach AL. (2001) *Molecular Modeling — Principles and Applications.* Pearson Education Ltd., Harlow.
2. Schlick T. (2002) *Molecular Modeling and Simulation*. Springer, Berlin.
3. Ferrara P, Gohlke H, Price D, Klebe G, Brooks CL. (2004). Assessing scoring functions for protein-ligand interactions. *J Med Chem* **47**: 3023–3047.
4. Kubinyi H. (1993) *3D-QSAR in Drug Design. Theory, Methods, and Applications.* ESCOM Science Publishers B.V., Leiden.
5. Kubinyi H. (2002) From narcosis to hyperspace: the history of QSAR. *Quant Struct-Act Relat* **21**: 348–356.
6. Kubinyi H, Folkers G, Martin YC. (1998) 3D-QSAR in drug design. Vol. 3. Recent advances, Kluwer/ESCOM, Dordrecht; also published in *Persp Drug Dis Des* **12–14**: 1–352.
7. Penzotti JE, Landrum GA, Putta S. (2004) Building predictive ADMET models for early decisions in drug discovery. *Curr Opin Drug Discov Devel* 7: 49–61.

8. Yoshida F, Topliss JG. (2000) QSAR model for drug human oral bioavailability. *J Med Chem* **43**: 2575–2585.
9. http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm.
10. Cronin M. (2003) A toxic gamble. *Chem Ind* **4**: 13–14.
11. Dunitz JD. (1994) The entropic cost of bound water in crystals and biomolecules. *Science* **264**: 670.
12. Cramer RD III, Patterson DE, Bunce JD. (1988) Comparative Molecular Fields Analysis (CoMFA). I. Effect of shape on binding of steroids to carrier proteins. *J Am Chem Soc* **110**: 5959–5967.
13. Hopfinger A, Wang S, Tokarski JS, *et al.* (1997). Construction of 3D-QSAR models using 4D-QSAR analysis formalism. *J Am Chem Soc* **119**: 10509–10524.
14. Ekins S, Bravi G, Binkley S, *et al.* (1999) Three- and four-dimensional quantitative structure-activity relationship analyses of cytochrome P450 3A4 inhibitors. *J Pharmacol Exp Ther* **290**: 429–438.
15. Vedani A, Briem H, Dobler M, Dollinger K, McMasters, DR. (2000) Multiple conformation and protonation-state representation in 4D-QSAR: the neurokinin-1 receptor system. *J Med Chem* **43**: 4416–4427.
16. Vedani A, Dobler M, Lill MA. (2005) Combining protein modeling and 6D-QSAR — simulating the binding of structurally diverse ligands to the estrogen receptor. *J Med Chem* **48**: 3700–3703.
17. Mohamadi F, Richards NGJ, Guida WC, *et al.* (1990) MacroModel — an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J Comput Chem* **11**: 440–467.
18. Perola E, Charifson PS. (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem* **47**: 2499– 2510.
19. Lill MA, Dobler M, Vedani A. (2006) Prediction of small-molecule binding to cytochrome P450 3A4: flexible docking combined with multidimensional QSAR. *ChemMedChem* **1**: 73–81.
20. Lill MA, Winiger F, Vedani A, Ernst B. (2005) Impact of induced fit on the ligand binding to the androgen receptor: a multidimensional QSAR study to predict endocrine-disrupting effects of environmental chemicals. *J Med Chem* **48**: 5666–5674.
21. Hahn M. (1995) Receptor-surface models. 1. Definition and construction. *J Med Chem* **38**: 2080– 2090.
22. Vedani A, Dobler M. (2002) 5D-QSAR: the key for simulating induced fit? *J Med Chem* **45**: 2139–2149.
23. Lill MA, Vedani A, Dobler M. (2004) *Raptor* — combining dual-shell representation, induced-fit simulation and hydrophobicity scoring in receptor modeling: application towards the simulation of structurally diverse ligand sets. *J Med Chem* **47**: 6174–6186.

24. Vedani A, Dobler M, Spreafico, M, Peristera O, Smiesko M. (2007) VirtualToxLab — *in silico* prediction of the toxic protential of drugs and environmental chemicals: validation status and Internet access protocol. *ALTEX* **24**: 153–161.
25. Vedani A, Decloux AV, Spreafico M, Ernst B. (2007). Predicting the toxic potential of drugs and chemicals *in silico*: a model for the peroxisome proliferator-activated receptor γ. *Toxicol Lett* **173**: 17–23.
26. Vedani A, Zumstein M, Lill MA, Ernst B. (2007) Simulating α/β specificity at the thyroid receptor: consensus scoring in multidimensional QSAR. *Chem Med Chem* **2**: 78–87.
27. Vedani A, Dobler M, Lill MA. (2006) The challenge of predicting drug toxicity *in silico*. *Pharmacol Toxicol* **99**: 195–208.
28. http://www.biograf.ch/downloads/quasar.pdf.
29. http://www.biograf.ch/downloads/raptor.pdf.
30. Vedani A, Dollinger H, Hasselbach KM, Dobler M, Lill MA. (2005) Novel ligands for the chemokine receptor-3 (CCR3): a receptor modeling study based on 5D-QSAR. *J Med Chem* **48**: 1515–1527.
31. Lill MA, Vedani A. (2006) Combining 4D pharmacophore generation and multidimensional QSAR: modeling ligand binding to the Bradykinin B_2 receptor. *J Chem Inform Model* **46**: 2135–2145.
32. Weiner SJ, Kollmann PA, Case DA, *et al.* (1984) A new force field for molecular-mechanical simulation of nucleic acids and proteins. *J Am Chem Soc* **106**: 765–784.
33. Stewart JJP. (1990) MOPAC — a semi-empirical molecular orbital program. *J Comput Aided Mol Des* **4:** 1–105.
34. Gustaffson JA. (1995) Receptor-mediated toxicity. *Toxicol Lett* **135**: 465–470.
35. Rihova B. (1998) Receptor-mediated targeted drug or toxin delivery. *Adv Drug Deliv Rev* **29**: 273–289.
36. Fischer B. (2000) Receptor-mediated effects of chlorinated hydrocarbons. *Andrologia* **32**: 279–283.
37. Hestermann EV, Stegemann JJ, Hahn ME. (2000) Relative contribution of affinity and intrinsic efficacy to aryl hydrocarbon receptor ligand potency. *Toxicol Appl Pharmacol* **168**: 160–172.
38. Lukasink K, Pitkanen A. (2000) $GABA_A$-mediated toxicity of hippocampal neurons *in vitro*. *J Neurochem* **74**: 2445–2454.
39. Rymer DL, Good TA. (2001) The role of protein activation in the toxicity of amyloidogenic Aβ (1–40), Aβ (25–35), and bovine calcitonin. *J Biol Chem* **276**: 2523–2530.
40. Hampson AJ, Grimaldi M. (2002) 12-hydroxyeicosatetrenoate (12-HETE) attenuates AMPA receptor-mediated neurotoxicity: evidence for a G-protein coupled HETE receptor. *J Neurosci* **22**: 257–264.

41. Oliver JD, Roberts RA. (2002) Receptor-mediated hepatocarcinogenesis: role of hepatocyte proliferation and apoptosis. *Pharmacol Toxicol* **91**: 1–7.
42. Zubay GL, Parson WW, Vance, DE (1995) *Principles of Biochemistry*, Brown Communications, Inc., Dubuque, USA.
43. Dibb S. (1995) Swimming in a sea of estrogens. *Ecologist* **25**: 27–31.
44. McLachlan JA, Arnold SF. (1996) Environmental estrogens. *Am Sci* **84**: 452–461.
45. Guillette LJ, Crain DA, Rooney AA, Pickford DB. (1995) Organization versus activation: the role of endocrine disrupting contaminants EDCs during embryonic development in wildlife. *Environ Health Perspect* **103**: 157–164.
46. Colborn T, von Saal FS, Soto AM. (1993) Developmental effects of endocrine-disrupting chemicals in wildlife and humans (see comments). *Environ Health Perspect* **101**: 378–384.
47. Hoare SA, Jobling S, Parker MG, Sumpter JP, White, R. (1994) Environmental persistent alkylphenolic compounds are estrogenic. *Endocrinology* **135**: 175–182.
48. Korach KS, Levy LA, Sarver PJ. (1987) Estrogen receptor stereochemistry: receptor binding and hormonal responses. *J Steroid Biochem* **27**: 281–290.
49. US Government (1996) *Safe Drinking Water Act Amendment*, Public Law 104–182 (Section 136): http://www.epa.gov.safewater/sdwa/index.html; *Food Quality Protection Act*. Public Law 104–170 (Section 408) http://www.fda.gov/opacom/ laws/foodqual/fqpatoc.htm.
50. Vedani A, Huhta DW. (1990) A new force field for modeling metalloproteins. *J Am Chem Soc* **112**: 4759–4767; Vedani A, Huhta DW. (1991) An algorithm for the systematic solvation of proteins based on the directionality of hydrogen bonds. *J Am Chem Soc* **113**: 5860– 5862.
51. http://www.biograf.ch/downloads/yeti.pdf.
52. Lukacova V, Balaz S. (2003) Multimode ligand binding in receptor-site modeling: implementation in CoMFA. *J Chem Inform Comput Sci* **43**: 2093–2105.
53. van Lipzig MMH, Laak AM, Jongejan A, *et al.* (2004) Prediction of ligand binding affinity and orientation of xenoestrogen to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. *J Med Chem* **47**: 1018–1030.
54. Akahori Y, Nakai M, Yakaba Y, *et al.* (2005) Two-step models to predict estrogen receptor α by 3D-QSAR using receptor-ligand docking simulation. *SAR QSAR Environ Res* **16**: 323–337.
55. da Cunha EFF, Martins RCA, Albuquerque MG, de Alencastro RB. (2004) LIV-3D-QSAR model for estrogen-receptor ligands. *J Mol Model* **10**: 297–304.
56. Kurunczi L, Seclaman E, Oprea TI, Crisan L, Simon A. (2005) MTD-PLS: a PLS variant of the minimal topologic difference method. Mapping interactions between estradiol derivatives and the alpha estrogenic receptor. *J Chem Inform Model* **45**: 1275–1282.

57. Sippl W. (2002) Binding-affinity prediction of novel estrogen receptor ligands using receptor-based 3D-QSAR methods. *Bioinorg Med Chem* **10**: 3741–3755.
58. Asikainen AH, Ruuskanen J, Tuppurainen KA. (2004) Consensus kNN: a versatile method for predicting estrogenic activity of organic compounds *in silico*. A comparative study with five estrogen receptors and a large, diverse set of ligands. *Environ Sci Tech* **38**: 6724–6729.
59. Hong H, Fang H, Xie Q, *et al.* (2003) CoMFA model using a large diverse set of natural, synthetic, and environmental chemicals for binding to the androgen receptor. *SAR QSAR Environ Res* **14**: 373–388.
60. http://www.biograf.ch/index.php?id=projects.

Chapter 21

Structure-based Computational Approaches to Drug Metabolism

M. A. Lill*

21.1 Introduction

Drug metabolism plays a huge role in disposition and the adverse reactions of drug candidates, and can critically affect the drug discovery process. Consideration of pharmacokinetic properties in the early phase of drug design has become increasingly mandatory for the efficient development of new drugs. Since the publication of the first X-ray structure of a mammalian cytochrome P450 enzyme (CYP) in the year 2000, structure-based computational approaches have become an important tool to help in understanding and predicting drug metabolism, focusing predominantly on phase I metabolism by CYPs. This chapter will cover the application of a diverse set of computational approaches to metabolism, including the induction of drug-metabolizing enzymes by nuclear receptors. It includes the discussion of methods for docking, predicting binding affinities, modeling of entry and exit paths to the binding site, as well as simulating protein flexibility, water molecules, and entropic effects upon ligand binding.

*Department of Medicinal Chemistry and Molecular Pharmacology, School of Pharmacy and Pharmaceutical Sciences, Purdue University, Heine Pharmacy Building, 575 Stadium Mall Drive, West Lafayette, IN 47907, USA. Email: mlill@purdue.edu.

Metabolism of drugs and other xenobiotics (e.g. environmental pollutants or industrial chemicals), also called biotransformation, is an essential detoxification mechanism in animal organisms, including humans. Typically, a compound is chemically transformed into a more hydrophilic compound, increasing its solubility, and consequently, its rate of elimination from the body. Metabolism is a major factor determining the systemic exposure of a drug, and represents the major clearance mechanism for about 73% of the top 200 selling drugs.[1]

Furthermore, xenobiotics can interact with the metabolic system by inhibiting metabolizing enzymes, or by activating or deactivating receptors transcribing the corresponding genes. Co-administration of two or more drugs can yield to increased plasma concentrations of one of these drugs, hence causing serious adverse drug-drug interactions complicating drug therapy. Moreover, metabolizing enzymes are also able to activate compounds to reactive intermediates that can form covalent adducts with macromolecules, e.g. DNA, leading to toxic reactions.

Biotransformation is typically divided into two main phases, phases I and II metabolism, although a clear temporal separation does not always exist.[2] Phase I reactions are mainly oxidative (although reductive biotransformation also occurs), and typically involve a member of the cytochrome P450 (CYP) enzyme family. In phase II reactions, molecules are enzymatically conjugated with polar functionalities (e.g. with glucuronic acid, glutathione, or sulfate) resulting in soluble entities, which are then easily excreted from the body. Although approximately a quarter of all compounds are biotransformed by enzymes involved in phase II drug metabolism, and crystal structures for these human enzymes have been resolved, e.g. glutathione S-transferases or UDP-glucuronosyltransferase, only anecdotal computational studies have been published that focus on understanding their enzyme mechanism, for example, in Ref. 3. Rarely have structure-based methods been applied towards predicting phase II metabolism for a series of compounds, e.g. docking to glutathione S-transferase.[4] This chapter will focus on CYPs, since a diverse set of computational approaches has deepened our understanding of how these enzymes work, and will also address the problems that still exist to reliably predict metabolism of compounds *in silico*.

21.2 Metabolism by Cytochrome P450 Enzymes

CYPs comprise a large superfamily of proteins, which is classified into families (minimum 40% sequence similarity), subfamilies (55%) and individual members that vary by at least 3% in their sequence. This classification is enumerated by an Arabic number for the family, a roman letter for subfamily, and a second number for the member. In humans, the family of P450 enzymes has 57 members, but only six (3A4/5, 2C9, 2C19, 2D6, and 1A2) are responsible for more than 90% of all phase I metabolic events of drug molecules.[1] CYPs are further involved in the formation of endogenous molecules, including hormones.

A variety of ligand-based computational approaches[5,6] have been applied to predict metabolism by P450 enzymes, including pharmacophore and QSAR models. Here, we will focus on the application of structure-based techniques to CYPs, which have gained importance since the publication of the first X-ray structure of a mammalian CYP in the year 2000.

Predicting drug metabolism by CYPs is a difficult task, as the metabolic feasibility is dependent on a variety of factors: the observed metabolic reaction and rate of biotransformation for a substrate is determined by its affinity towards the enzyme, its orientation towards the reactive center and the intrinsic reactivity of the chemical group in close proximity to the catalytic center. In addition to these mechanistic factors, abundance of the various enzymes involved in biotransformation and their genetic polymorphism can lead to a pronounced variability in metabolizing capacity, and account for individual differences in drug response.

21.2.1 *Structural Data for P450 Enzymes*

CYPs are mainly α-helical in their fold [Fig. 21.1(a)] with some β-sheet elements. Mammalian CYPs are membrane-anchored by an

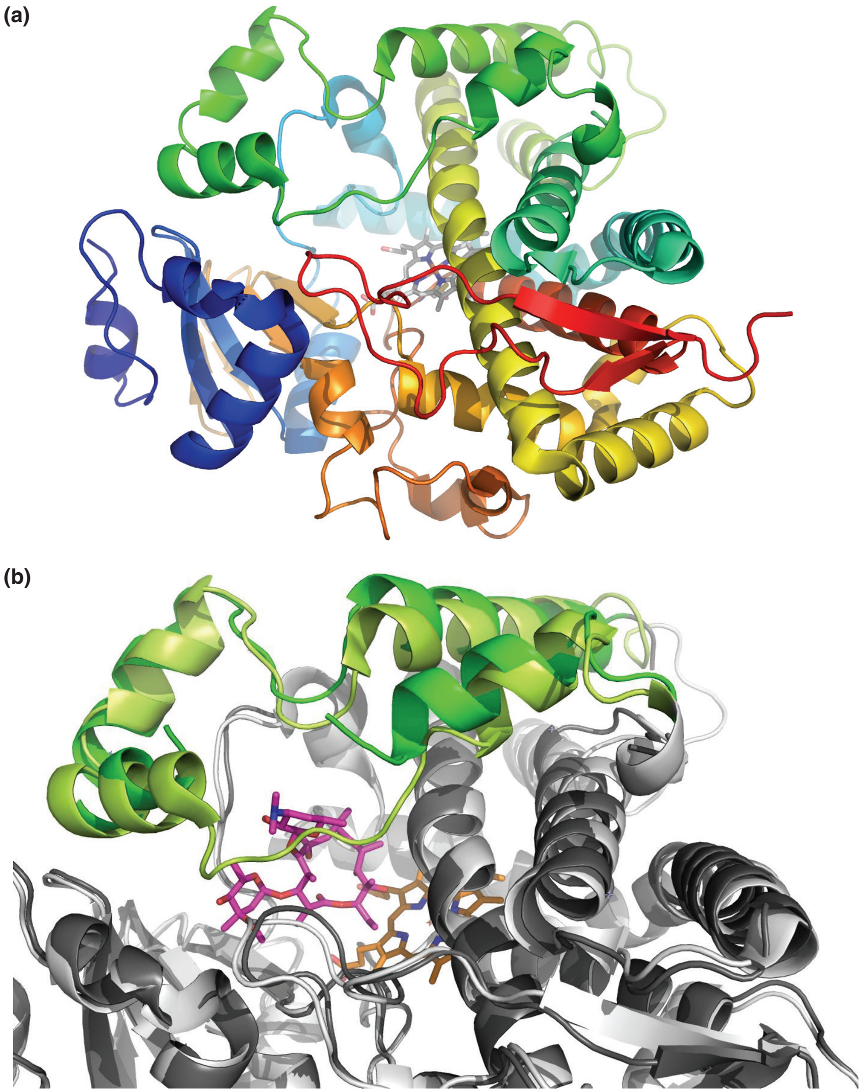


Fig. 21.1 **(a)** Crystal structure of the apo-form of CYP3A4.[43] **(b)** Superposition of crystal structures of apo-(light gray) and erythromycin (dark gray) bound CYP3A4.[40] CYP3A4 performs a significant induced fit in the F to G portion of the protein (apo: lime, erythromycin bound: green) to accommodate erythromycin; part of the F-F′ loop becomes disordered. The solvent accessible volume of the binding site is increased by a factor of approximately two. Figure was created with PyMol.[44]

N-terminal helix, which is removed in all X-ray crystallographic studies in order to increase solubility. CYPs comprise of 12 helices plus additional helical elements in some of the CYPs (e.g. B′, F′, and G′ helices as part of the flexible structure locking the binding site). CYPs enclose a common iron-containing heme *b* cofactor coordinated to a proximal cysteine residue and a loosely bound water molecule on the distal site of the heme, which is replaced by a ligand upon binding to the active site. Currently (by June 2007), there are 25 X-ray structures of 10 mammalian (20 human structures of 1A2, 2A6, 2A13, 2C8, 2C9, 2D6, 3A4, 2R1) CYPs available.

X-ray structures with several different compounds complexed with the same CYP member have often identified different conformations of the protein. This demonstrates the importance of protein flexibility for binding and metabolizing structurally different molecules [Fig. 21.1(b)]. As ligand diversity can be profound for specific CYPs (Fig. 21.2), induced protein fit is a fundamental factor underlying its biological function. Thus, it seems to be important to include protein flexibility in all computational studies on P450 enzymes, at least when dealing with non-congeneric data sets.

Prior to the publication of the first mammalian X-ray structures, homology models for all-important human CYPs (e.g. 3A4, 2D6, 2C) were derived based on the experimental structures of bacterial CYPs.[7] Validation strategies include NMR-derived ligand-heme distance restraints, mutagenesis data, and the comparison between predicted ligand orientation towards the catalytic center with the experimentally determined site of metabolism, suggesting which chemical group of a ligand binds in close proximity to the catalytic center. At least for the less promiscuous enzymes like CYP2D6, acceptable agreement between experiment and computationally predicted binding modes was observed.[8] It should, however, be noted that predicting the correct binding modes of structurally diverse ligands in promiscuous binding pockets of CYPs is still not fully resolved, even when an experimentally resolved X-ray structure is available (see Section 21.2.2).

Since the first X-ray structures for mammalian CYPs were resolved, templates with higher sequence similarity are available, providing

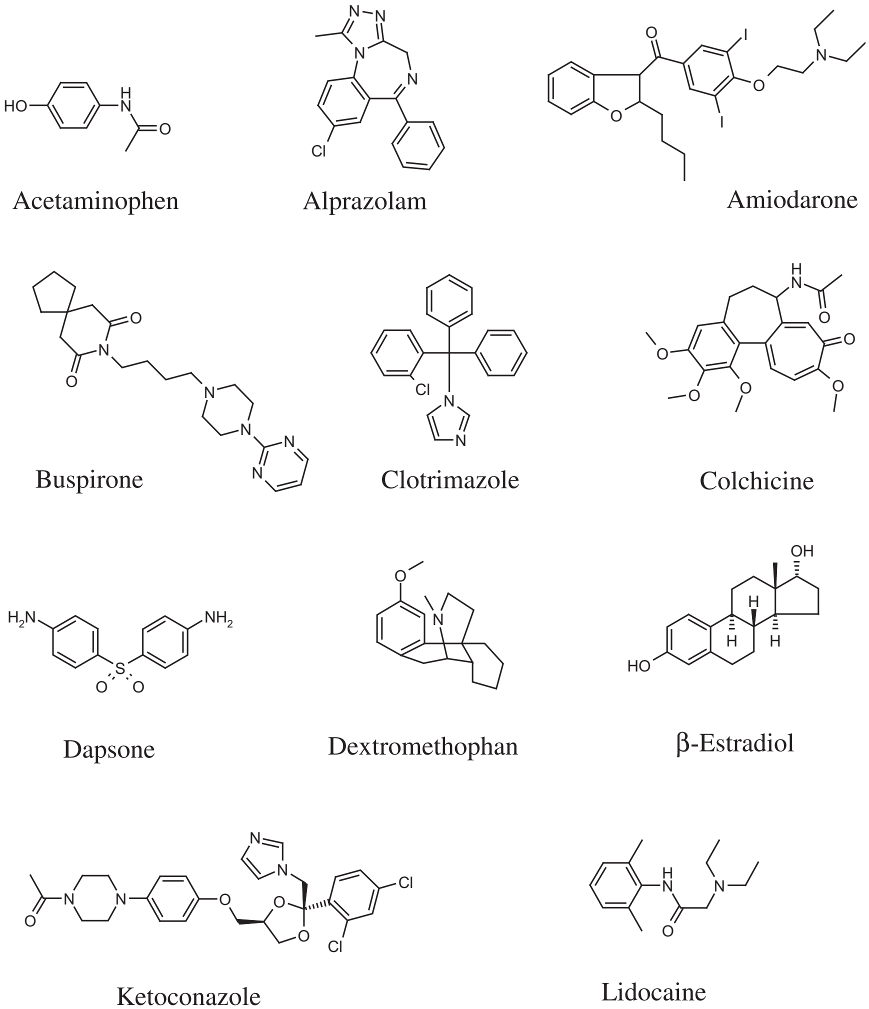


Fig. 21.2 A selection of structurally diverse xenobiotics binding to CYP3A4.

a basis for better homology models for human CYPs. Furthermore, microbial and mammalian CYPs differ in important structural details, e.g. positions of the F-G structural unit[9] or in the substrate-recognition site,[10] which can have a significant influence on xenobiotics binding.

21.2.2 *Prediction of Binding Modes*

Both crystallographic protein structures and homology models have been utilized for studying protein-ligand interactions using computational methods. The dominant factors determining if and at which position a substrate is metabolized by a specific CYP are the structural elements of the substrate that binds close to the catalytic heme center, and the intrinsic reactivity of this chemical group. Automated docking has been used to virtually screen compound libraries aimed at identifying energetically feasible docking poses, and especially to predict the chemical portion of a possible substrate that lies close to the reaction center. This allows the prediction of the regioselective component of drug metabolism, a prerequisite for catalytic site prediction. However, the promiscuity of most CYPs allows for the metabolism of structurally diverse sets of xenobiotics, making the prediction a difficult task.

A further consequence of the structural variability of the binding compounds is that one needs to take into account the effect of water molecules that often bind in the active site in addition to the ligand. These waters may directly mediate hydrogen-bonds between the ligand and protein, or might be considered as part of the ligand's solvation shell, which is not completely stripped during binding.[11] De Graaf *et al.* have demonstrated that the explicit inclusion of pre-equilibrated water molecules in docking yields better agreement with experimental binding poses.[12] Depending on the docking software and protocol, they have reached an improvement by about 10–30% in catalytic site prediction for CYP2D6 when water was explicitly taken into account. Through docking with four different docking programs (AutoDock, FlexX, GOLD-Goldscore, and GOLD-Chemscore) and rescoring the best docking poses with the scoring function SCORE, they have achieved a successful prediction of the experimentally observed site of metabolism in 80% of all substrates.

In several X-ray structures, large non-reactive substrate-heme distances have been observed, e.g. (*S*)-warfarin-heme distance ~10 Å in CYP2C9[13] or progesterone-heme distance ~17 Å in CYP3A4.[14] Until recently, two possible explanations have been discussed that address this data: electron transfer triggers a conformational change necessary for the ligand to move towards the heme group. Alternatively, homotropic or heterotropic cooperativity of a molecule located at the observed distant site might trigger conformational changes of the protein, allowing a second compound to bind to the same binding site, but in closer proximity with the heme.

Recently, an alternative scheme was postulated pointing out the importance of entropy upon ligand binding to CYPs. Levy and co-workers[15] performed replica-exchange molecular dynamics simulations on *N*-palmitoylglycine binding to P450 BM3. 24 MD simulations (called replicas) were preformed in parallel, each at a different temperature, ranging from 260 K to 463 K. At regular intervals, replicas can exchange temperatures when fulfilling the Metropolis transition criteria with respect to their actual potential-energy difference. This allows alternative configurations, sampled at high temperatures, to be trapped. Using a temperature-weighted histogram analysis method, it is further possible to estimate the relative population and free energies of these states. This study identified two distinct stable states: one in agreement with the X-ray structure, in which the substrate carbon atom, which is hydroxylated, is ~8 Å away from the heme iron atom; and another configuration where the distance is reduced to ~4.5 Å. The former is mostly populated by the low temperature replicas, which perfectly agrees with the experimental setup for the X-ray studies performed at 110 K. The later protein-substrate configuration was populated 70% of the simulation time at 302 K. The reason is that the former configuration has the lowest enthalpy value, whereas the latter complex is entropically favored, thus not observed at the low temperature conditions of the X-ray studies; this complex, however, allows the ligand to temporarily approach the catalytic center close enough to be metabolized. The study clearly demonstrated the significance of ligand and protein entropy in predicting the correct binding mode of a CYP substrate, a factor that is typically neglected in automated docking.

21.2.3 *Prediction of Binding Affinities*

Quantifying protein-ligand interactions is a further prerequisite for *in silico* prediction of drug metabolism and adverse drug-drug interactions. Standard scoring functions employed so far have not shown significant correlation with experimental binding affinities for mammalian CYPs.[16,17]

Molecular-dynamics (MD) based free-energy methods (see Chapter 19), on the other hand, show good agreement between predicted binding affinities and their corresponding experimental values. Helms and Wade[18] were able to predict the absolute binding free energy of camphor to P450 cam with 0.8 kcal/mol deviation from the experiment using a multi-configuration variant of thermodynamic integration. Linear-interaction analysis (LIE)[19] was applied on 11 compounds binding to P450 cam[20] and three substrates to CYP1A1,[21] reproducing the experimental binding affinity within a mean error of 0.55 and 0.8 kcal/mol, respectively. In LIE, the electrostatic and van der Waals interaction energies between the ligand and surrounding water is averaged over a MD simulation of the compound in solvent alone. The sum of these mean energies is subtracted from the interaction energies between the ligand and protein in a corresponding MD simulation of the complex:

$$\Delta G = \left\langle E^{elst}_{L-P} \right\rangle + \left\langle E^{vdW}_{L-P} \right\rangle - \left\langle E^{elst}_{L-S} \right\rangle - \left\langle E^{vdW}_{L-S} \right\rangle \tag{21.1}$$

Molecular-mechanics Poisson-Boltzmann surface area (MM/PBSA) method[22] typically uses snapshots of a single MD simulation of the complex to estimate the free energy of binding by subtracting the free energy of free ligand and protein from that of the complex:

$$\Delta G_{\text{binding}} = \Delta G_{P+L} - \Delta G_{L} - \Delta G_{P} \tag{21.2}$$

Each free energy is determined from the average electrostatic $\langle E^{elst} \rangle$ and van der Waals $\langle E^{vdw} \rangle$ energy of each entity plus terms for the solvation energy. These terms are computed in continuum solvent, using a finite difference Poisson-Boltzmann model $\langle G^{PB} \rangle$, and a non-polar solvation term, represented by the solvent-accessible surface area $\langle G^{SA} \rangle$. Finally, configurational entropy $\langle T\Delta S^{config} \rangle$ is estimated from normal-mode analysis and subtracted from the other contributions:

$$\Delta G = \left\langle E^{elst} \right\rangle + \left\langle E^{vdW} \right\rangle + \left\langle G^{PB} \right\rangle + \left\langle G^{SA} \right\rangle - \left\langle T\Delta S^{config} \right\rangle \qquad (21.3)$$

This method was applied to a set of compounds binding to CYP2B4 and predicted all binding affinities within 2 kcal/mol of their experimental values.

While these examples seem to suggest that these methods are able to reliably quantify protein-ligand interactions, it must be noted that the data sets in these studies were rather small and did not cover the experimentally known chemical space of ligands binding to CYPs. In addition to the structural promiscuity of CYP ligands, a further challenge arises from the fact that compounds with different net charges are able to bind to the same CYP. This leads to large differences in electrostatic interaction energies, presenting a hurdle to these methods. Further studies on large and structurally diverse sets of compounds binding to promiscuous human CYP enzymes, like CYP3A4 or 2C9, must demonstrate the potential of these methods as standard tools for accurate prediction of binding affinities in drug-discovery settings.

Lill *et al.*[23] combined structure-based and ligand-based design concepts, developing a computational model that predicts the inhibitory potential of structurally diverse molecules binding to CYP3A4. Possible binding modes were first sampled using docking that incorporated induced protein fit at the level of flexible side-chains, as well as on-the-fly solvation of the protein-ligand complexes. The predicted binding modes for most compounds were consistent with experimental data. CYP3A4 is known to accommodate a ligand in various binding poses, yielding different metabolic products of the compound. 4D-QSAR techniques (see Chapter 20) can explicitly handle different ligand configurations in a single simulation. On average, the four energetically most favorable docking orientations are used as input for the multidimensional QSAR concept[24] to quantify protein-ligand interaction. This approach has produced predictions of the binding affinity of training and test ligands, on average, within a factor of 2.7 and 3.8 from the experimental values, respectively.

21.2.4 *Modeling of Access/Exit Channels*

The active site of CYPs is located deep inside the protein, and the channels through which ligands access and egress the binding site can have a strong impact on substrate specificity and enzyme kinetics. Wade and coworkers applied a variety of different computational methods, based on X-ray structures, to identify and analyze possible paths between protein surface and binding site of CYPs. They developed a systematic survey of ligand paths covering all currently available CYP crystal structures, applying the CAVER program.[25] CAVER projects the protein structure on a grid, where a penalty is assigned to each grid point depending on the distance to a protein atom (smaller distance corresponding to higher penalty). CAVER identifies channels upon traversing from the binding site cavity to the surface by optimizing a cost function that is based on the assigned penalty values. However, entry and exit channels may exist in a closed state in the crystal structures, as opening and closing of channels is often a dynamic process. Thus, substrate specificity may need to be determined both by structural and dynamic properties of the entry and egress channels. Along this line, Wade and coworkers also used thermal motion pathway (TMP) analysis to identify chains of atoms with above-average temperature factors connecting the active site and protein surface.[26]

Random expulsion molecular dynamics (REMD) simulations have been performed to identify exit routes of specific ligands, including protein dynamics explicitly.[27] As the time scale for ligand access or exit is several log-units above that accessible to standard MD simulations, spontaneous ligand access to or exit from the active site cannot be observed by this type of simulations. In order to enhance the probability of ligand exit, an artificial force is imposed on the ligand with a randomly chosen direction. The magnitude and direction of this force are kept constant for a given time period, Δt. If the compound encounters relatively rigid parts of the binding site along the direction of expulsion force, the average velocity over Δt of the ligand projected in this direction will become smaller than a predefined threshold value. In this case, a new direction of the expulsion force is randomly chosen. This process is repeated until the molecule is expelled along an exit channel, yielding average velocities above the given threshold value.

Although more time consuming compared to CAVER or TMP, REMD explicitly allows for the inclusion of protein dynamics and provides mechanistic insight into the process of channel opening. Standard MD simulation in conjunction with essential dynamics analysis (see Chapter 11) further allows the identification of dominant modes along the opening route. REMD yields a semi-quantitative estimate of the energetic profile and kinetic rates for ligand exit from the active site. The energetic profile was studied in more detail using adiabatic mapping based on steered MD simulations.[28] Herein, a constantly increasing external force directed along the identified paths was added to accelerate the ligand repulsion from the active site.

These studies have identified the location of several channels that share common structural elements in CYPs, but which, however, deviate significantly among CYPs in their specific structural topology (Fig. 21.3). Variations in sequence, and observation of different side-chain and backbone configurations along the channels makes homology-modeling without extensive simulations questionable. Important structural elements include the F-G structural block (including the F-G helix, connecting loop, and F′-,G′-helices when present), and the B-C loop, both of which border most of the identified channels. Significant, and often correlated motions of these structural elements are observed in MD simulations. Exit and access channels are all oriented in the distal direction opposite to those elements of the protein to which P450 reductase or cytochrome b_5 binds. Consequently, these electron donating proteins cannot block the channels, however, they might regulate their dynamic behavior.

In membrane-bound mammalian CYPs the channels can be classified into those ending in the membrane and those parallel to it. The paths of the former group are suggested to represent the entry route of hydrophobic compounds, which can easily penetrate into the

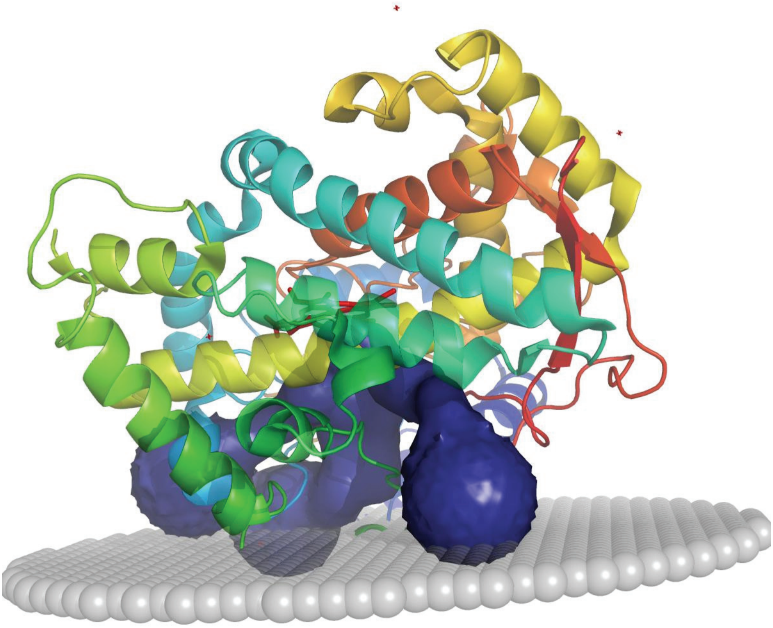


Fig. 21.3 Possible entry and exit channels as predicted using CAVER[45] for CYP3A4 (pbd-code: 1TQN). Membrane position and orientation as predicted by Lomize *et al.*[46] Figure was created with PyMol.[44]

membrane. The latter group is thought to play an important role in substrate channeling. It is postulated that the channels are ideally oriented towards the dimerization interface with other biotransformation enzymes, including CYPs, allowing subsequent metabolic events in different enzymes without intermediate release to the solvent. Solvent accessible channels were also identified in most CYPs, which might be important for the release of more hydrophilic products from the metabolic machinery.

Typically, results from simulations have identified several distinct but spatially close channels, which might merge into wide-open channels or even funnels. This topology might be responsible for combining ligand specificity with the observed large variety in the size of compounds binding to a specific CYP. It should also be noted that the natural environment, where CYPs are partially embedded in the membrane, which was neglected in all X-ray structures, might have a significant effect on the structural and dynamic properties of the channels ending in or close to the membrane. Future computational studies might be predestinated to study these effects.

21.2.5 *Reaction Mechanism*

Computational studies have contributed strongly to our current understanding of the complex mechanism involved in the catalytic reactions of CYPs. Most detailed investigations have focused on the bacterial P450 cam enzyme, which is the first sequenced and crystallographic resolved CYP, but the overall mechanism is believed to be shared by different CYPs. The consensus mechanism of CYPs' monooxygenase function (not including alternative pathways, e.g. uncoupling reactions) is depicted in Fig. 21.4.

Structures for each trappable intermediate state have been obtained for P450 cam using time-resolved crystallographic studies,

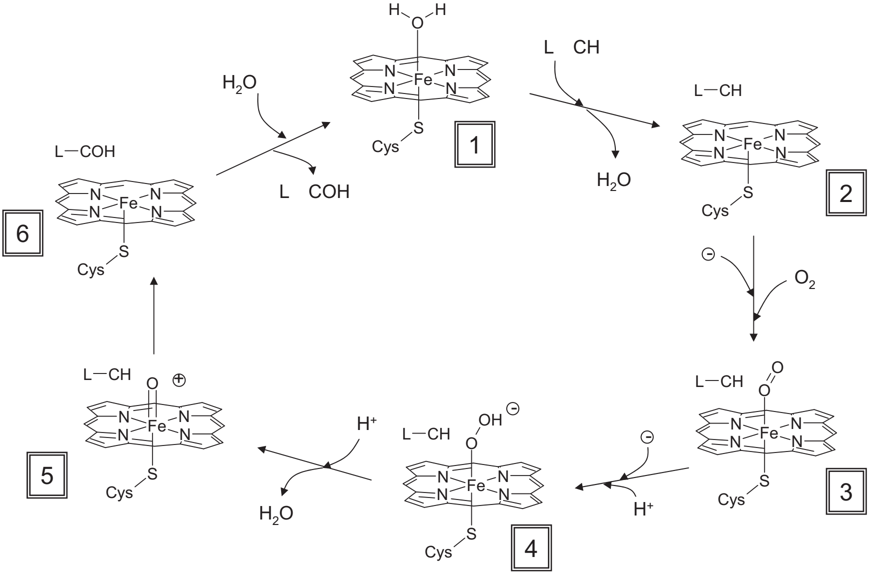


Fig. 21.4 Common mechanism for hydroxylation reaction of CYPs.

which show conformational changes along the pathway.[29] Based on these structures, Friesner, Thiel, and coworkers[30–32] have used mixed quantum mechanics/molecular mechanics (QM/MM) approaches including the full protein structure, and compared the results with parallel calculations only on active site models. This has allowed them to study the influences of the heme environment on different states in the reaction pathway.

In its ligand-free resting state (**1**), CYPs heme iron is in its ferric state (Fe^{III}), typically in low spin state (although some CYPs are in high spin state). The surrounding protein residues can strongly influence the spin state: in P450 cam the low spin state is significantly favored by protein stabilization of the antibonding interaction among Fe and the axial cysteine and water ligand. After substrate binding, the axial water ligand is replaced, typically leading to a high spin state (quartet or sextet) (**2**). The surrounding amino acids appear to have less effect on the spin state of (**2**). The change in spin state for P450 cam, accompanied by a significant increase in reduction potential, results in an electron transfer to the heme group typically from NADPH via NADPH-P450 reductase. The resulting ferrous CYP then binds O_2 (**3**). After a second electron transfer from NADPH-P450 reductase or cytochrome b_5, a proton is transferred (in P450 cam from Thr252, see Fig. 21.5) to the dioxygen producing a ferric hydroperoxide species (**4**). Using a simulated annealing protocol coupled to QM/MM to equilibrate the resulting structures, Guallar *et al.*[32] have shown that in P450 cam, a water channel from Glu366 to Thr252 provides the missing proton to Thr252, and after a reorientation of the hydrogen-bond network, an additional proton is transferred from Thr252 to the hydroperoxide group. Lastly, the O-O bond is cleaved, generating a water molecule and an oxyferryl group (compound I) (**5**). Generally, compound I is viewed as Fe^{IV} and has a one-electron deficiency in the porphyrin ring. In a hydroxylation reaction, this reactive species abstracts a hydrogen atom from the substrate, forming a radical intermediate, which rapidly collapses (rebound mechanism) forming the hydroxylated product (**6**). This product is released from the binding site, yielding the resting state (**1**) again. The energetic influence of the heme surrounding amino acids

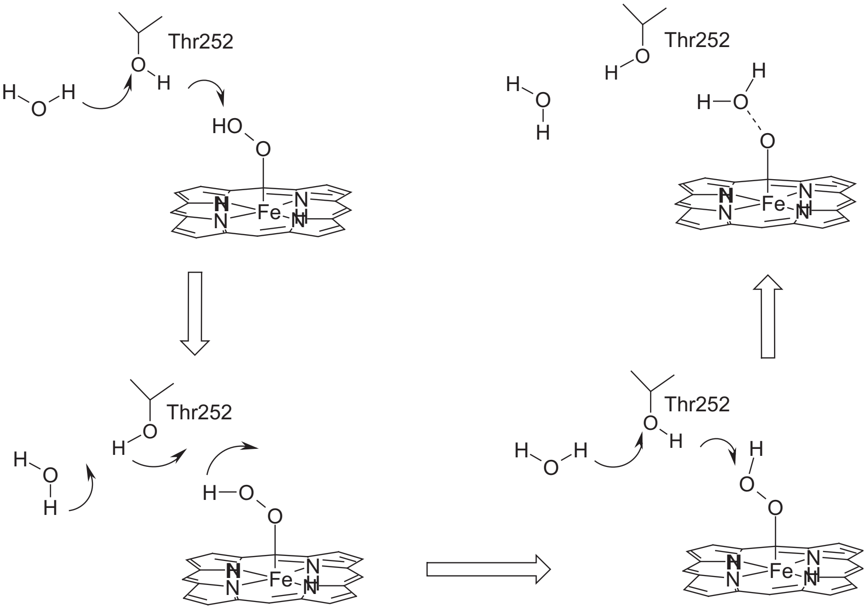


Fig. 21.5 Scheme for O–O cleavage process in P450cam (figure adapted from Ref. 32).

on the hydrogen-atom abstraction reaction is still controversial. In P450 cam, a water molecule seems to reduce the energy barrier involved.[30] Further stabilization of the intermediate state due to electrostatic interaction between the heme carboxylate substituents and nearby positive residues is also proposed.[32]

Quantum mechanical calculations are also used to predict the intrinsic reactivity of xenobiotics, a necessary component to predict CYP metabolism. Semiempirical AM1 calculations, for example, are used to compute the hydrogen abstraction energy for compounds, which are then correlated with descriptors such as hybridization, atomic element, aromatic character, number of non-hydrogen neighbors, etc.[33] Park and Harris combined docking and MD simulations to sample reasonable heme-ligand configurations, with hydrogen abstraction energies calculated by density functional theory.[34] However, only recently, an integrative high-throughput approach, combining electronic properties of a molecule with structural properties of the protein, was developed by Cruciani and coworkers.[35] In their technique, MetaSite, the GRID approach is utilized to screen the interaction fields of selected chemical probes with the protein atoms in the binding site. Amino-acid side-chains are allowed to rearrange in response to attractive or repulsive interactions with the probe atom. The interaction map is then transformed into histograms representing the distances between the reactive center and the different physicochemical moieties in the binding pocket. For the substrate, the distances between different atom types (represented by the chemical probes in GRID) are binned. The protein and ligand descriptors are then correlated to predict which position in the molecular skeleton of the ligand is accessible to the catalytic center, defining an accessibility parameter for each ligand atom.

A library of fragments was designed, representing a large portion of the chemical space accessible for drugs. Molecular orbital calculations on these fragments are performed to calculate the reactivity of each fragment atom in a specific reaction type, e.g. hydroxylation, dealkylations, and deamination. The combination between reactivity and accessibility parameters provides an estimate of the relative reaction probability of each fragment atom for a specific metabolic reaction. Validation studies on over 900 substrates for CYP1A2, 2C9, 2C19, 2D6, and 3A4 showed on average a rate of over 80% for successful prediction of metabolic sites.

21.3 Induction of Drug Metabolism

In addition to drugs directly binding to CYPs, adverse drug-drug interactions may occur when a drug binds to a transcription factor that regulates the P450 gene transcription. Inhibition of CYP-gene transcription may increase the concentration in the body of a coadministered drug due to slower metabolism, and can result in significant adverse effects. Induction of P450 genes can alter the metabolic profile of a drug by increasing metabolic rates or by creating an alternative pathway of metabolism with profound effects on the drug's toxicity profile. Drugs such as cisapride, terfenadine, and mibefradil are examples of compounds that have been withdrawn from the market because of adverse drug-drug interactions associated with CYP induction or inhibition.

The mechanism of regulating the induction process is primarily due to agonizing or antagonizing ligand-activated transcription factors such as the aryl-hydrocarbon receptor (AhR) for the CYP1A family, constitutive-androstan receptor (CAR) for the CYP2B family, and pregnane xenobiotics receptor (PXR), glucocorticoid receptor (GR) and vitamin D receptor (VDR) for the CYP3A family. Since there is some considerable degree of cross-talk among these receptors and to other proteins, the process of induction is not a single receptor-protein relationship. Among those transcription factors, PXR plays a key role as it binds structurally diverse xenobiotics and endogenous compounds and regulates, in addition to CYP3A4, the genes for CYP2B6, CYP2C8, CYP2C9, CYP2C19, phase II enzymes like UDP-glucoronosyltransferase, and efflux pumps like multidrug-resistance proteins (MDR) 1 and 2.

3D structures of PXR[36] obtained by X-ray crystallography display a large hydrophobic ligand-binding domain with few polar residues. The structures further reveal that significant flexibility and disorder of the topological elements lining the binding site enable the binding of structurally diverse ligands [Fig. 21.6(a)].

Before the first crystal structures became available, pharmacophore models were derived to attempt to understand the key features for ligand binding to PXR.[37] In addition to several ligand-based pharmacophore models using the software Catalyst, Schuster and Langer[38] manually annotated an additional hypothesis for PXR activation based on the X-ray structure of PXR with the ligand SR12813. The results suggested that hydrogen bonding to Gln285 is critical for PXR activation, while a second hydrogen bond to His407 could be identified for most ligands [Fig. 21.6(b)]. Further hydrophobic interactions contribute to ligand affinity, where highly active compounds share up to five hydrophobic features, allowing the ligand to occupy the predominantly hydrophobic binding pocket.

Gao *et al.*[39] used docking with the support of induction experiments to identify important interactions for PXR activation. Based on their studies, they were able to hypothesize that hydrogen-bonding

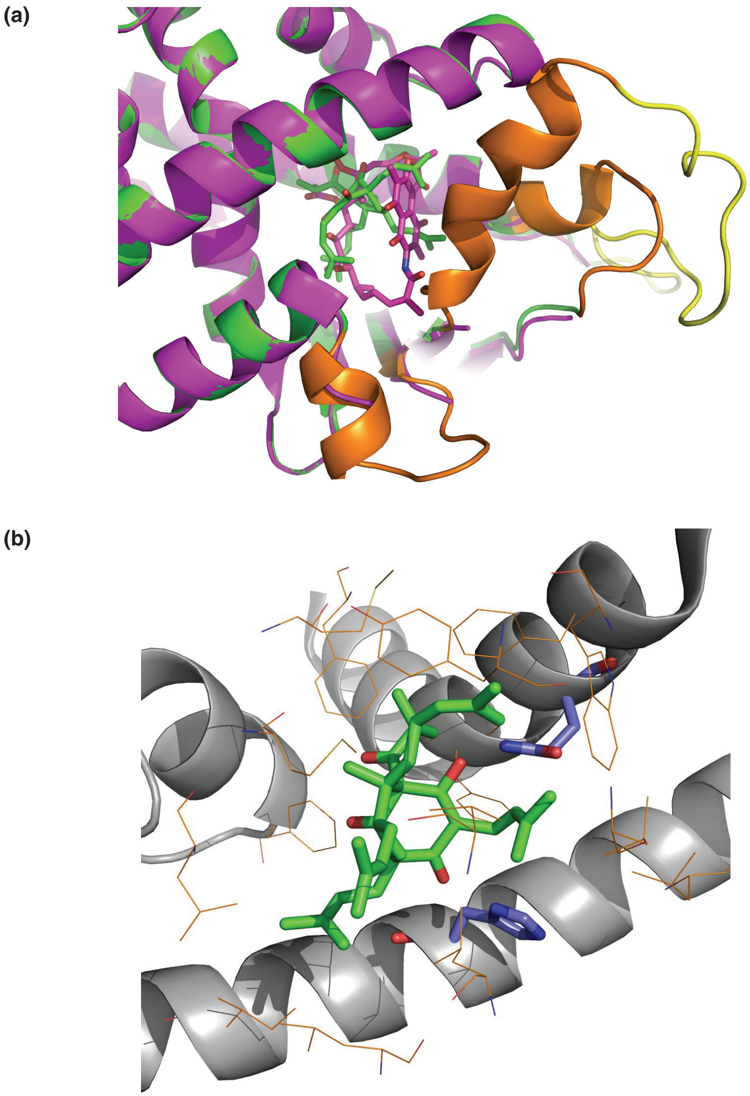


Fig. 21.6 **(a)** Superposition of crystal structures of PXR with bound hyperforin (protein and ligand's carbon atoms: green) and rifampicin (purple). In the hyperforin-PXR structure a large region (L178 to A197) is disordered: a possible representative of the expected ensemble of loop structures was modeled with the program loopy[47] (colored in yellow). When the larger rifampicin binds, additional helical portions become unwinded (Ia, IIIa) or disordered (Ib, II, IIIb); **(b)** X-ray structure of PXR with bound hyperforin. Figure was created with PyMol.[44]

with H407 in helix 11 and hydrophobic groups occupying at least two of the three hydrophobic areas [Fig. 21.6(b), purple carbon atoms: region 1 containing F288, W299, Y306, M246; yellow carbon atoms: region 2 containing F429, F251, F281; brown carbon atoms: region 3 containing I414, L240, L206, L209, M243 L324, L308, V211) are essential for activation. Interaction with H407 (helix 11) and with F429 (helix 12) might be involved in stabilizing the agonist state, as helices 11 and 12 form the activation function 2 (AF2) moiety responsible for co-activator binding.

In general, it is not obvious how future docking studies can handle the structural promiscuity of compounds binding to PXR. Furthermore, the profound flexibility and partial disorder of the binding site, which seems to be size dependent on the bound ligand, provides a real challenge for computational studies on this protein.

21.4 Future Outlook

Drug metabolism, by cytochrome P450 enzymes in particular, is an extensively studied subject. An interesting mixture of different computational applications has brought insight to many facets of xenobiotic metabolism, but various questions still remain. Due to the structural variability of substrates and inhibitors, and the observed flexibility of the protein, identifying the naturally existing binding modes is still challenging, especially for the very promiscuous isoenzymes such as CYP3A4 and CYP2C9. Various crystal structures and computational studies on protein-ligand complexes have recently suggested that induced protein fit, water molecules mediating interactions between ligand and protein, as well as entropic contributions, both of the ligand and the protein, might play significant roles in ligand binding to CYPs. While these factors might influence ligand binding to many protein targets, they seem to be particularly important for CYP-ligand interactions. For example, no current computational approach for docking or quantifying binding affinities seems to be capable of dealing with the observed size of induced fit in CYPs: the volume of the binding site varies by at least a factor of two in CYP3A4.[40]

All structure-based computational studies on mammalian CYPs are based on crystal structures of a solubilized form of the protein, i.e. the natural membrane environment is neglected. Part of the F-G helical region, however, is known to be embedded in the membrane. This portion of the protein is directly involved in ligand binding, and is a constituting part of the entry and exit channels. Future computational studies might investigate on an atomistic scale how the inclusion of the membrane in the simulation might change current detailed models for CYP metabolism. Atomistic simulation techniques also seem to be predestinated to investigate whether the binding of P450 reductase or cytochrome b_5 will have a structural and/or dynamic influence on CYP-substrate interaction, beyond their role as an electron donor.

Experimental studies have exhibited atypical steady-state kinetics, i.e. a sigmoidal relationship between reaction velocity versus substrate concentration. Possible interpretations of the data are a sequential binding step model[41] as well as homotropic or heterotropic cooperativity between substrates, i.e. the simultaneous binding of an additional effector molecule to the binding site favors the substrate's configurations in nearby distance to the catalytic center. Recent crystal structures with two molecules binding to the active site,[40] and the first MD simulation[42] of two diazepam molecules binding to CYP3A4 have suggested a possible mechanism for homotropic cooperativity. If cooperativity will prove to be a common phenomena for binding to CYPs like 3A4 or 2C9, current docking methods may need to be modified as the binding of one molecule might influence the energetic profile of the poses of the other compound and vice versa.

As a further complication, the majority of enzymes in phases I and II metabolism are polymorphic. This can cause quantitatively decreased or enhanced drug metabolism, or yield alternative metabolic products. Many examples exist where individuals or whole populations carrying certain alleles do not benefit from drug therapy due to ultra rapid metabolism, or suffer from adverse effects due to reduced metabolism causing toxic drug plasma concentrations. Computational methods, e.g. a combination of *in silico* mutations, homology modeling, MD simulations, and docking, might strongly contribute to a better understanding of polymorphic aspects of drug metabolism. Predicting these pharmacokinetic properties of compounds will be a cornerstone on the way towards the dream for individualized medicine.

References

1. Williams JA, Hyland R, Jones BC, *et al.* (2004) Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metab Dispos* **32**: 1201–1208.
2. Josephy DP, Guengerich FP, Miners J. (2005) "Phase I and Phase II" drug metabolism: terminology that we should phase out? *Drug Metab Rev* **37**: 575–580.
3. Bowman AL, Ridder L, Rietjens IM, Vervoort J, Mulholland AJ. (2007) Molecular determinants of xenobiotic metabolism: QM/MM simulation of the conversion of 1-chloro-2,4-dinitrobenzene catalyzed by M1-1 glutathione S-transferase. *Biochemistry* **46**: 6353–6363.
4. Zavodszky MI, Sanschagrin PC, Korde RS, Kuhn LA. (2002) Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *J Comput Aided Mol Des* **16**: 883–902.
5. de Groot MJ, Ekins S. (2002) Pharmacophore modeling of cytochromes P450. *Adv Drug Deliv Rev* **54**: 367–383.
6. Ekins S, de Groot MJ, Jones JP. (2001) Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites. *Drug Metab Dispos* **29**: 936–944.
7. de Graaf C, Vermeulen NP, Feenstra KA. (2005) Cytochrome p450 *in silico*: an integrative modeling approach. *J Med Chem* **48**: 2725–2755.
8. de Groot MJ, Ackland MJ, Horne VA, Alex AA, Jones BC. (1999) A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed N-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J Med Chem* **42**: 4062–4070.
9. Otyepka M, Skopalik J, Anzenbacherova E, Anzenbacher P. (2007) What common structural features and variations of mammalian P450s are known to date? *Biochimica et Biophysica Acta (BBA) — Gen Sub* **1770**: 376–389.
10. Gotoh O. (1992) Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *J Biol Chem* **267**: 83–90.
11. Wester MR, Johnson EF, Marques-Soares C, *et al.* (2003) Structure of mammalian cytochrome P450 2C5 complexed with diclofenac at

2.1 A resolution: evidence for an induced fit model of substrate binding. *Biochemistry* **42**: 9335–9345.

12. de Graaf C, Oostenbrink C, Keizers PH, *et al.* (2006) Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J Med Chem* **49**: 2417–2430.
13. Williams PA, Cosme J, Ward A, *et al.* (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature* **424**: 464–468.
14. Williams PA, Cosme J, Vinkovic DM, *et al.* (2004) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* **305**: 683–686.
15. Ravindranathan KP, Gallicchio E, Friesner RA, McDermott AE, Levy RM. (2006) Conformational equilibrium of cytochrome P450 BM-3 complexed with N-palmitoylglycine: a replica exchange molecular dynamics study. *J Am Chem Soc* **128**: 5786–5791.
16. Kemp CA, Flanagan JU, van Eldik AJ, *et al.* (2004) Validation of model of cytochrome P450 2D6: an *in silico* tool for predicting metabolism and inhibition. *J Med Chem* **47**: 5340–5346.
17. Tanaka T, Okuda T, Yamamoto Y. (2004) Characterization of the CYP3A4 active site by homology modeling. *Chem Pharm Bull (Tokyo)* **52**: 830–835.
18. Helms V, Wade RC. (1998) Computational alchemy to calculate absolute protein-ligand binding free energy. *J Am Chem Soc* **120**: 2710–2713.
19. Aqvist J, Medina C, Samuelsson JE. (1994) A new method for predicting binding affinity in computer-aided drug design. *Protein Eng* **7**: 385–391.
20. Paulsen MD, Ornstein RL. (1996) Binding free energy calculations for P450cam-substrate complexes. *Protein Eng* **9**: 567–571.
21. Szklarz GD, Paulsen MD. (2002) Molecular modeling of cytochrome P450 1A1: enzyme-substrate interactions and substrate binding affinities. *J Biomol Struct Dyn* **20**: 155–162.
22. Srinivasan J, Miller J, Kollman PA, Case DA. (1998) Continuum solvent studies of the stability of RNA hairpin loops and helices. *J Biomol Struct Dyn* **16**: 671–682.
23. Lill MA, Dobler M, Vedani A. (2006) Prediction of small-molecule binding to cytochrome P450 3A4: flexible docking combined with multidimensional QSAR. *Chem Med Chem* **1**: 73–81.
24. Lill MA, Vedani A, Dobler M. (2004) Raptor: combining dual-shell representation, induced-fit simulation, and hydrophobicity scoring in receptor modeling: application toward the simulation of structurally diverse ligand sets. *J Med Chem* **47**: 6174–6186.
25. Cojocaru V, Winn PJ, Wade RC. (2007) The ins and outs of cytochrome P450s. *Biochim Biophys Acta* **1770**: 390–401.

26. Ludemann SK, Carugo O, Wade RC. (1997) Substrate access to cytochrome P450cam: a comparison of a thermal motion pathway analysis with molecular dynamics simulation data. *J Mol Model* **3**: 369–374.
27. Ludemann SK, Lounnas V, Wade RC. (2000) How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *J Mol Biol* **303**: 797–811.
28. Ludemann SK, Lounnas V, Wade RC. (2000) How do substrates enter and products exit the buried active site of cytochrome P450cam? 2. Steered molecular dynamics and adiabatic mapping of substrate pathways. *J Mol Biol* **303**: 813–830.
29. Schlichting I, Berendzen J, Chu K, *et al.* (2000) The catalytic pathway of cytochrome p450cam at atomic resolution. *Science* **287**: 1615–1622.
30. Altun A, Shaik S, Thiel W. (2006) Systematic QM/MM investigation of factors that affect the cytochrome P450-catalyzed hydrogen abstraction of camphor. *J Comput Chem* **27**: 1324–1337.
31. Altun A, Guallar V, Friesner RA, Shaik S, Thiel W. (2006) The effect of heme environment on the hydrogen abstraction reaction of camphor in P450cam catalysis: a QM/MM study. *J Am Chem Soc* **128**: 3924–3925.
32. Guallar V, Friesner RA. (2004) Cytochrome P450cam enzymatic catalysis cycle: a quantum mechanics/molecular mechanics study. *J Am Chem Soc* **126**: 8501–8508.
33. Singh SB, Shen LQ, Walker MJ, Sheridan RP. (2003) A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. *J Med Chem* **46**: 1330–1336.
34. Park JY, Harris D. (2003) Construction and assessment of models of CYP2E1: predictions of metabolism from docking, molecular dynamics, and density functional theoretical calculations. *J Med Chem* **46**: 1645–1660.
35. Cruciani G, Carosati E, de Boeck B, *et al.* (2005) MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J Med Chem* **48**: 6970–6979.
36. Carnahan VE, Redinbo MR. (2005) Structure and function of the human nuclear xenobiotic receptor PXR. *Curr Drug Metab* **6**: 357–367.
37. Ekins S, Erickson JA. (2002) A pharmacophore for human pregnane X receptor ligands. *Drug Metab Dispos* **30**: 96–99.
38. Schuster D, Langer T. (2005) The identification of ligand features essential for PXR activation by pharmacophore modeling. *J Chem Inform Model* **45**: 431–439.
39. Gao YD, Olson SH, Balkovec JM, *et al.* (2007) Attenuating pregnane X receptor (PXR) activation: a molecular modeling approach. *Xenobiotica* **37**: 124–138.
40. Ekroos M, Sjogren T. (2006) Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc Natl Acad Sci USA* **103**: 13682–13687.
41. Isin EM, Guengerich FP. (2006) Kinetics and thermodynamics of ligand binding by cytochrome P450 3A4. *J Biol Chem* **281**: 9127–9136.

42. Fishelovitch D, Hazan C, Shaik S, Wolfson HJ, Nussinov R. (2007) Structural dynamics of the cooperative binding of organic molecules in the human cytochrome P450 3A4. *J Am Chem Soc* **129**: 1602–1611.
43. Yano JK, Wester MR, Schoch GA, *et al.* (2004) The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-A resolution. *J Biol Chem* **279**: 38091–38094.
44. DeLano WL. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, Palo Alto, CA, USA.
45. Damborsky J, Petrek M, Banas P, Otyepka M. (2007) Identification of tunnels in proteins, nucleic acids, inorganic materials, and molecular ensembles. *Biotechnol J* **2**: 62–67.
46. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. (2006) OPM: orientations of proteins in membranes database. *Bioinformatics* **22**: 623–625.
47. Xiang Z, Soto CS, Honig B. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci USA* **99**: 7432–7437.

This page intentionally left blank

Section IV

New Frontiers in Experimental Methods

This page intentionally left blank

Chapter 22

New Frontiers in X-ray Crystallography

C. U. Stirnimann[†] and M. G. Grütter[*,‡]

22.1 Introduction

Since the first recording of a diffraction pattern of the protein pepsin in 1934 by Bernal and Crowfoot,[1] progress in macromolecular crystallography has occurred in distinct intervals. With the solution of the phase problem for macromolecular diffraction data and the subsequent structure determination in 1959 of the first proteins myoglobin and hemoglobin by Kendrew and Perutz, respectively,[2,3] macromolecular crystallography has entered the field of modern molecular biology as an essential methodology. Until about the mid-1970s macromolecular crystallography was primarily practiced in relatively few specialized research laboratories around the world by physicists and chemists. Determining the three-dimensional structure of a macromolecule by X-ray structure analysis then represented a major effort and typically took several years to be completed. X-ray sources were weak, the data collection on films and data processing involved a lot of manual interventions, and the amount of data to be handled were at the limit of the storage and processing capacity of

[*] Corresponding author.

[†] Structural and Computational Unit, EMBL Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany. Email: christian.stirnimann@embl.de

[‡] Universität Zürich, Biochemisches Institut, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland. Email: gruetter@bioc.uzh.ch

computers. Refinement was only being developed and applied to relatively small proteins. Model building was a manual task and typically took several months. The achievements of that period can, for instance, be read in a review by Matthews.[4] The macromolecules analyzed then were those that were naturally occurring in large quantities and could easily be purified from natural sources.

Nevertheless, macromolecular crystallography did contribute enormously to modern molecular biology in helping to better understand the structure/function relationship of macromolecules. In the early 1980s, the recombinant DNA technology had completely changed the situation regarding macromolecular targets that could structurally be investigated. Now, essentially every protein independent of its natural occurrence can be overproduced in other host cells, preferably in bacteria. In parallel, over several decades, the computer technology has advanced, with the result that today almost all crystallographic computing can be performed on a personal computer. Another critical factor in the maturation of the field was the availability of synchrotron radiation that allowed the acquisition of data of smaller crystals at a much faster rate than with conventional X-ray sources.

This was the point when large pharmaceutical industries had started to establish macromolecular X-ray crystallography for structure-based drug design. Today, structure-based drug design is generally seen as an integral part in the drug development process.

During the last decade, X-ray crystallography has undergone further significant technological advances mainly as a consequence of the many structural genomics research programs that were started primarily in the United States of America and in Japan, but also to a lesser extent in Europe.[5–7] The latest developments, the way macromolecular crystallography is carried out today, and the future potential directions will be the scope of this review.

22.2 The Methods

The prerequisites for a successful macromolecular crystal structure determination are (i) the availability of sufficiently large amounts of highly purified material, (ii) crystals of sufficient diffraction quality for a high resolution structure determination, and (iii) phase information

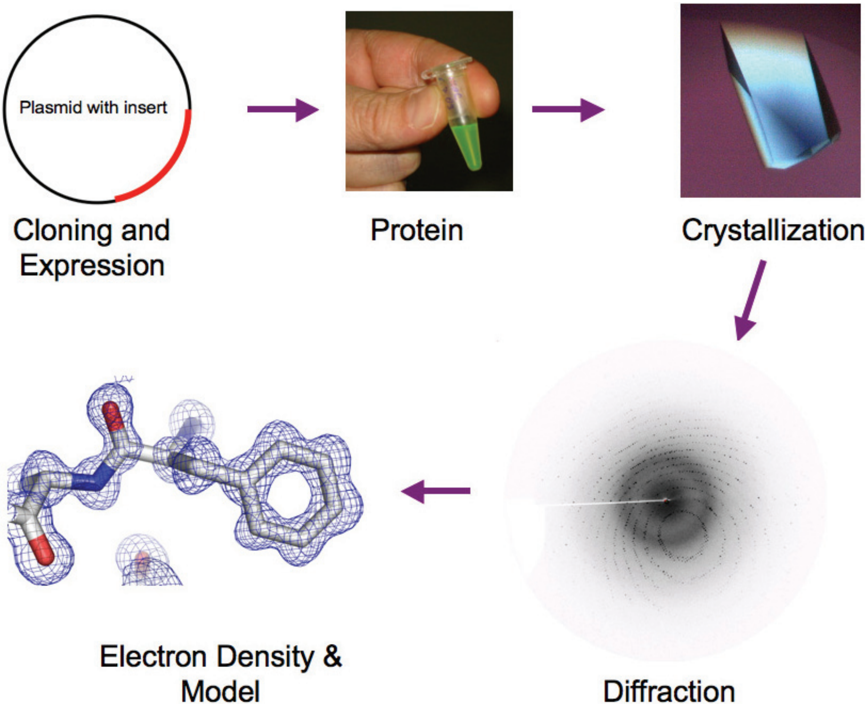


Fig. 22.1 Schematic overview of the process.

complementing the experimentally determined diffraction data for structure calculation. Today, well-diffracting crystal data collection using synchrotron radiation and structure determination is automated. Phase determination methods using synchrotron sources have been improved markedly (see below). The biggest challenge now towards a structure of a macromolecule is its expression, purification, and crystallization (Fig. 22.1).

22.2.1 *Protein Production*

An important prerequisite to be able to perform structural studies by X-ray crystallography or by other methods, such as NMR spectroscopy and electron diffraction, is the availability of sufficiently large amounts of highly purified macromolecular sample. Routine procedures in structural biology include cloning, expression in bacteria, yeast, insect or mammalian cells, or today, even in cell-free systems, as well as purification, biochemical and biophysical characterization of proteins before submitting them to crystallization experiments. For a current review, see Ref. 8.

22.2.2 *Crystallization*

When decreasing the solubility of a macromolecule by adding a precipitating agent, crystallization can be triggered. Crystallization was first done in batch mode: solutions of the purified protein and a precipitant such as ammonium sulfate were mixed in small vials and stored for days and weeks. Many alternative ways to crystallize proteins have been proposed, as reviewed in Refs. 9–11, with the most frequently used method being the vapor diffusion method in hanging or sitting drops. Other methods are the counterdiffusion methods in capillaries,[12] microdialysis,[13] or microbatch under oil.[14] The experiments were visually inspected using a stereo light microscope to analyze whether crystals appeared.

The process of crystal formation — nucleation, crystal growth, and growth cessation — is not understood in sufficient detail but depends on many physical and chemical parameters, such as temperature, pH, type and concentration of the precipitating agent, buffer and ion concentration, in addition to the properties of the macromolecule itself. Therefore, the crystallization conditions for a particular macromolecule can only be determined by screening many conditions. Most often, the precipitating agents used are inorganic salts, such as ammonium sulfate, and organic polyols, such as polyethylene glycols of various sizes. Crystallization screening conditions selected by sparse matrix sampling[15] have been proposed and are available commercially for the initial screening of crystallization conditions. Crystallization in the presence of detergents is necessary to crystallize membrane proteins. Here, the choice of the detergent that keeps the protein in its native conformation and does not disturb the regular assembly in a crystal lattice is critical.[16,17]

Crystallization experiments today are routinely carried out using the vapor diffusion method. To increase the throughput and to reduce consumption of purified protein, crystallization at the nanoliter scale is performed. This is possible with the availability of pipetting robots that can reproducibly dispense nl amounts of solution. With the enormous increase in the number of experiments set up, examination of the crystallization experiments is also automated. Individual crystallization experiments are photographed following a given protocol, and evaluation can be done either by the human eye or by software designed to recognize crystalline features, precipitates, or clear solution. The ultimate test to evaluate the quality of crystals is their diffraction in an X-ray beam. For this, crystals are either mounted in thin quartz or glass capillaries with enough mother liquor to prevent crystals from drying out. With the introduction of cryo-crystallography, the crystals are removed from the crystallization droplet by a small fiber loop and subsequently plunged into liquid nitrogen. The crystalline lattice is preserved and the liquid in the solvent channels of the macromolecular crystal is solidified in the form of amorphous glass. The crystals at the temperature of liquid nitrogen (100 K) are much more resistant to radiation damage by the X-ray beam than crystals exposed to X-rays at ambient temperatures.[18,19]

22.2.3 *Data Collection and Data Processing*

22.2.3.1 *Diffraction data collection*

The final experimental part in a structure determination is the collection of diffraction data from the crystals. Significant technological advances have helped to transform the data collection from an extremely tedious process to an almost automated one. Development of electronic detectors instead of X-ray films, synchrotron X-ray sources instead of sealed X-ray tubes or rotating anode X-ray instruments, as well as the development of software to automatically find the orientation of the crystal in the X-ray beam, and subsequent integration of the intensities of each diffraction spot, have made this part of the macromolecular structure determination extremely fast and automatic.

22.2.3.2 *X-ray sources*

Conventional X-ray sources used in the past and still used today are sealed tubes or rotating anode generators with a copper anode that provides a sufficient flux and monochromatic radiation from the CuKα transition at a wavelength of 1.54 Å after passing a Ni filter, graphite monochromator, or mirrors that remove unwanted polychromatic white radiation. Recently, the improved properties of the optical elements used with laboratory X-ray sources have dramatically increased the X-ray flux. The devices are double-reflecting mirrors and multilayered devices that provide a highly focused and intense X-ray beam with the intensity of a second generation bending magnet synchrotron beamline.[20,21] The radiation sources with the highest intensities are synchrotron sources. They are today the standard installations for collecting diffraction data from macromolecular crystals. Home sources are mostly used to perform preliminary characterization of the crystals before collecting data at a synchrotron. At present, the most favored beamlines are tunable wavelength lines with radiation from insertion devices such as wigglers or undulators that provide a highly parallel beam of excellent brightness and with a small size. Using such installations, collection of a full dataset of a well-diffracting macromolecular crystal typically takes less than 30 minutes, and structures can be solved in a few hours.[22] Currently, fourth generation X-ray sources are being designed and built at SLAC, Stanford, CA, USA (http://www.ssrl.slac.stanford.edu/lcls/) and DESY, Hamburg, Germany (http://www.hasylab.desy.de/facility/fel) based on the concept of the free electron laser (FEL). These installations will provide coherent X-ray beams of an intensity that is several orders of magnitude higher than available today. The use of such extremely powerful radiation sources for macromolecular structural biology is a new challenge but has the potential to revolutionize structural biology. Extracting diffraction data from extremely small crystalline samples or even from non-crystalline samples is theoretically possible.[23–25] One fundamental question exists here that can only be answered experimentally, which is the following: can diffraction data be acquired before the sample is destroyed due to the enormously strong radiation?

22.2.3.3 *Detectors*

Protein diffraction data was first recorded on conventional X-ray film using various geometric arrangements, and with diffractometers collecting data from individual diffraction spots in a sequential way. With the availability of automatic computer-controlled two-dimensional detectors such as multiwire proportional counters, TV cameras, image plates, and finally charge-coupled devices diffraction images are directly stored on computer storage devices for subsequent data reduction. The currently used CCD detectors are large to capture a wide diffraction angle, have fast readout times to keep pace with the short second exposure times per image collected, and provide diffraction data of much higher quality. Recently, very fast data transfer solid state pixel detectors allow the recording of diffraction images from constantly rotating crystals with exposures controlled by the precisely synchronized shutter.[26] This opens the possibility again to more accurately evaluate the profile and intensity of each diffraction spot.

22.2.3.4 *Data Processing Software*

The initial interpretation of the recorded diffraction patterns involves the calculation of the crystal orientation and the prediction of the location of the diffraction spots, the indexing of the spots, and the integration of the individual reflection profiles, the application of the necessary corrections, and the merging and scaling of the data. The result is a list of data points with the three indices (h, k, l), the intensity, and the standard deviation for each data point. The entire procedure of data processing can be performed by one of the available highly sophisticated software systems and is not further discussed here. For details, the descriptions of these software packages have to be consulted. The most widely used systems are Mosfilm,[27] HKL2000,[28] and XDS.[29]

22.2.4 *Phase Determination*

22.2.4.1 *Single and multiple anomalous sispersion methods*

Since the early days of crystallography, new crystal structures have been solved by incorporating heavy atoms (mainly heavy metals) into protein crystals. The first method ever used to solve a protein structure was in fact a heavy atom method, multiple isomorphous replacement (MIR).[30,31] Since then, MIR became the technique of choice to solve the phase problem for novel structures. The bottleneck of MIR is the need of at least two, but mostly more, isomorphous crystals (native and heavy atom-derivatized crystals), for obtaining unambiguous phases, and thus, an interpretable electron density map.[32,33]

Experimental difficulties in measuring precise amplitude values prevented for a long time the productive use of the anomalous dispersion effect of scattering heavy atoms. This effect is often in the range of 2 to 5% of the real scattering component, and thus, smaller than or comparable to the measured error.[34] The above difficulties became less severe during the 1980s and 1990s. Those two decades witnessed advances in the accuracy of diffraction data measurement and the appearance of dedicated synchrotrons sources, allowing for very accurate tuning of the X-ray beam wavelength. Those technical advances paved the way to the development and widespread use of the multiple anomalous dispersion (MAD) method. MAD typically involves the collection of three datasets from the same single crystal, thus obliterating the thorny issue of isomorphism between different crystals, which affects the MIR approach. In MAD, the first dataset is collected at the absorption edge of the heavy atom, a second one at its inflection point, and a third dataset at high- or low-energy remote wavelength. The MAD method requires that a K- or L-absorption edge of the heavy atom is located within the wavelength range of a synchrotron source that lies usually between 0.7 Å and 2.0 Å.[35,36] A much simpler approach to phase determination that gained recent popularity is the single anomalous dispersion (SAD) method, where a single, highly redundant dataset is measured from a single wavelength. SAD does not require the availability of synchrotron sources since it is not necessarily coupled to the absorption edges of heavy atoms.[37] Thus, iodide (copper anode ($\lambda = 1.54$ Å)[38]) or sulphur phasing (copper or chromium anode ($\lambda = 2.23$ Å)[37,39,40]) became viable phasing approaches at home X-ray sources.

Besides crystal soaking in heavy atom solutions, which often decreases or ruins the diffraction of crystals,[33] other techniques for incorporating heavy atom scatterers into proteins are now available: among these, the most popular is the replacement of methionines by selenomethionines by molecular biological means.[41] Another very elegant and fast approach is a quick crystal-soak (30–60s) in a cryo-solution that contains either bromide or iodide in concentrations between 0.3 and 1 M.[38] Also gaseous elements such as Xe or Kr can be incorporated into protein crystals, where they bind to hydrophobic cores, and were successfully used for phasing.[42,43]

Phasing using SAD or MAD data is typically carried out in a three-step procedure followed by automatic model building if the resolution and phases are sufficient. First, the position of the anomalous scattering heavy atoms has to be determined, by Patterson or by direct methods.[44–47] Once heavy atom positions are identified, initial phases are calculated, refined, and evaluated. In a final step, phases are improved by density modification. Several powerful programs are available for performing the above steps and are summarized in Table 22.1.

The programs mentioned in Table 22.1 are mostly part of phasing pipelines. In the earlier versions of the autoSHARP,[48] the phasing pipeline performed the heavy atom search using the direct methods program RANTAN.[44] With the integration of SHELXD[45] for heavy atom search, autoSHARP gained further effectiveness in structure solution. SHARP[49] refines the initial heavy atom positions found by SHELXD and calculates phases, which are then improved by solvent flattening using DM,[50] or solvent flattening and flipping using SOLOMON.[51] If the resolution is below 2.8 Å, automatic model

Table 22.1 Programs for Experimental Phasing

<table>
<tr><th>HA Search</th><th>Phasing</th><th>Density Modification</th><th>Automatic Model Building</th></tr>
<tr><td colspan="2">SOLVE</td><td colspan="2">RESOLVE</td></tr>
<tr><td>SHELXD</td><td colspan="2">SHELXE</td><td>ARP/wARP</td></tr>
<tr><td></td><td>SHARP</td><td>DM[a]</td><td></td></tr>
<tr><td></td><td></td><td>SOLOMON[a,b]</td><td></td></tr>
<tr><td>ACORN</td><td></td><td></td><td></td></tr>
</table>

[a] solvent flattening; [b] solvent flipping.

building in ARP/wARP (see Section 22.2.5 and Refs. 52, 53) is started. Other powerful pipelines to solve structures by experimental phasing are the SOLVE/RESOLVE[47] and HKL2MAP[54] packages, where the latter combines SHELXD with SHELXE.[55]

22.2.4.2 *Molecular replacement*

In molecular replacement, the phase problem is solved by correctly positioning one or several search models that are structurally related to the target structure in the asymmetric unit. Hence, a six-dimensional search is required. A first group of traditional molecular replacement programs is based on the suggestion by Rossmann and Blow[56] to subdivide the six-dimensional search in two steps: a first three-dimensional search is applied to find the correct model rotation, which is then followed by a three-dimensional translational search. This method is implemented in programs such as AmoRe,[57] CNS[58] or MOLREP.[59] This very successful strategy can nonetheless be problematic in the presence of tightly packed crystals, non-globular proteins, or when several molecules have to be placed in the asymmetric unit, as the rotational and translational variables are not optimized simultaneously.[60] A full six-dimensional search can be employed in those cases, with the drawback of very time-consuming calculations. However, with the steady increase of computer performance over the last decades, full six-dimensional searches are becoming more and more affordable and they are implemented in several molecular replacement programs, namely EPMR,[61] Queen of Spades,[62] and SoMoRe.[60]

The newest and currently most popular molecular replacement program is Phaser.[63,64] It addresses the molecular replacement problem in the traditional bi-three-dimensional way, but takes advantage of the implementation of maximum-likelihood methods. Phaser exhibits a higher success rate compared to the other molecular replacement programs, especially when only distantly related search models are available.

A trend towards automation is well-visible in the molecular replacement field. Several molecular replacement pipeline programs have been developed, among which BALBES[65] and MrBump[66] were recently released. BALBES requires only the target amino acid sequence and the reflection data file as an input. It searches for possible molecular replacement models in an aptly modified PDB database and performs molecular replacement trials. MrBump allows more user intervention. It first performs a FASTA search of target amino acid sequence against the PDB to find homologue structures. From those homologues, trial models are generated and optimized using various helper programs from the CCP4 suite.[67] The models are iteratively used in Phaser or Molrep for molecular replacement trials until a solution is found. The possible solution is finally used in a Refmac restrained refinement to check for its correctness.[66]

22.2.5 *Model Building and Refinement*

22.2.5.1 *Model building*

The first molecular graphics program conceived for model building purposes was INTER,[68] which was later followed by O.[69] Additional molecular graphics programs that are widely used in the community are XtalView[70] and MAIN.[71] Since 2004, when *Coot*[72] was introduced, a very powerful tool for model rebuilding, refinement and structure validation became available. Thanks to an intuitive user interface, the program is fast and easy to learn. Coot allows the use of real space refinement, followed by geometric regularization. In addition, refinement programs, such as Refmac v5[73] and SHELXL[74] can be directly launched from the Coot interface and the refined model with the corresponding electron density maps is automatically updated in the graphics window. These features dramatically accelerate the whole rebuilding and refinement procedure. Coot also includes additional structure validation tools, further improving the efficiency and reliability of model building and refinement.

22.2.5.2 *Refinement*

The structure factors observed from the diffraction experiment and those calculated from an unrefined macromolecular model largely diverge. To minimize the difference between observed and calculated structure factors, the coordinates and B-factors of the model are refined against the structure factors, with geometrical and stereochemical restraints being included in the refinement process. Thus, refinement is a process in which the structure factors calculated from the model are adjusted to those measured in an X-ray diffraction experiment.[32]

Nowadays, three refinement programs are most commonly present in the crystallographer's toolbox. Refinement with the first, CNS,[58] typically involves first rigid body refinement, followed by positional refinement and/or simulated annealing, and finally, B-factor refinement. To do so, CNS uses geometrical information from parameters derived by Engh and Huber in 1996,[75] CNS is not suitable at atomic resolution since it implements the fast Fourier transform method, whose error increases with resolution.[76] SHELXL[74] is designed for medium and high-resolution refinement (better than 2.5 Å) and uses a least-squares target. SHELXL is very powerful if noncrystallographic symmetries are present in the structure, as it allows deformation in the NCS-related structures, which is not the case in the other programs.[77] Additionally, it allows the refinement of anisotropic thermal parameters at high resolutions and includes refinement of twinned data. The third program, Refmac v5[73] implements a maximum likelihood target and allows the refinement of anisotropic thermal parameters. Refmac allows also for TLS-based (translation, libration, screw) B-factor refinement. This refines anisotropic atomic displacement parameters for pseudo rigid bodies, for which translation, libration, and screw-rotation displacements are refined. This can be described using only three matrix tensors per TLS group, which diminishes the number of parameters used in the B-factor refinement drastically.[77,78] The integration of Refmac within the Coot interface[72] makes the program very easy to use in structure refinement.

A new suite, successor to CNS, is PHENIX. Its refinement tools were recently developed with an emphasis on refinement automation. In a first step, the best refinement strategy is chosen and parameters are tuned. The refinement cycle starts with the optimization of the bulk-solvent model, followed by anisotropic scaling and error model estimation for the maximum likelihood target function. The ordered water model is then built automatically, followed by an optional simulated annealing step. The refinement cycle is completed by coordinate and B-factor refinement. If no convergence is reached, the program automatically re-launches the refinement steps described above.[79]

22.2.5.3 *Automation of model building and refinement*

The ARP/wARP software package[52,53] allows auto-building of atomic models either from experimental phases or from molecular replacement solutions. An auto-building step is based on an unrestrained dummy atom model that is generated for phase improvement.[80] In an iterative process, each auto-building step is followed by a number of refinement cycles. If the phases and resolution allow for it, the program can dock and fit the protein sequence to the electron density after having built the main chain. ARP/wARP is highly dependent on the diffraction data resolution. Earlier versions of the program had a minimal resolution limited to 2.3 Å, while in the newest release (v6), a resolution higher than 2.6 Å is required for model building.

A program complementary to ARP/wARP is Buccaneer.[81] Unlike ARP/wARP, it shows only marginally resolution-dependent behavior when auto-tracing main chains (lowest resolution tested: 3.2 Å). Success depends, however, on the quality of the initial phases. To improve those, the maximum likelihood-based program Pirate[82] is used for density modification. Buccaneer is still under development and does not yet implement refinement cycles, proofreading, and side-chain docking.

Besides automatic model building programs, it would be advantageous for many users to have a program that is able to build and refine protein structures in a completely user intervention-free fashion. A first step in this direction is the program LAFIRE.[83] The program performs user intervention-free model building and refinement using either CNS[58] or Refmac.[73] The program monitors the R_{free} value as a control criterion for model building and refinement. According to its authors, the program performs best in a resolution range between 1.65 Å and 3 Å and was able to produce a fully refined model without user intervention in 11 out of 14 test cases.[83]

22.3 Recent Achievements and Future Challenges

As a result of the enormous technological progress, spectacular achievements in the structure determination of supramolecular complexes and membrane proteins were possible. Moreover, due to automation of the methodology primarily facilitated through the structural genomics initiatives, X-ray crystallography of biological macromolecules has also become an integral part in the development of drugs against disease-related target proteins.

22.3.1 *Structure-based Drug Design*

Apart from membrane-bound receptors such as G-protein coupled receptors, any disease-relevant target protein can be subjected to structure determination, and the active site of enzymes can be exploited by experimental structure determination of ligands bound to the active site of the enzyme or by docking compounds in the active site by various computational methods. The main protein classes currently investigated in the pharmaceutical industry are kinases and proteases because enzymes of both classes are involved in numerous different signaling pathways and biological processes that are affected in diseases such as cancer, cardiovascular diseases, cell death deregulated diseases, or diseases of the nervous system. Successful structure-based drug design resulting in drugs actually on the market are human immunodeficiency virus protease inhibitors and the renin inhibitor, aliskiren,[84] with many other examples in the late stages of clinical trials. Similarly, the structural work on kinases and kinase-inhibitor complexes has contributed to the development of potent inhibitors with compounds in clinical trials and on the market. Clearly, in the future, the application of experimental structure determination in the drug design process can be expected to increase due to the continuous technological improvements in the experimental techniques.

22.3.2 *Supramolecular Complexes*

This area has seen a tremendous development over the last years with complexes of ever increasing complexity being analyzed at atomic resolution by X-ray crystallography. The most spectacular contributions with enormous impact on the understanding of biological processes are: (i) the RNA polymerase II, which provides the foundation for understanding transcription,[85] (ii) the structure of the ribosome, which shows the complexity of this particle and its function in the translation process,[86] (iii) the structure of an intact nucleosome, which shows how DNA is compacted and protected from harm by the histone octamer protein complex,[87] (iv) the 20S proteasome structure, which is a multienzyme complex of 28 protein subunits involved in the degradation of ubiquitinylated proteins,[88] and (v) the structure of viruses such as the foot-and-mouse disease virus-oligosaccharide receptor complex[89] with crystal cell dimensions close to 1000 Å. All these examples illustrate the future potential of X-ray crystallography that allows the tackling of molecular systems or molecular machines that will help us understand the interplay of individual proteins in living organisms.

22.3.3 *Membrane Proteins*

This class of proteins is still considered a major challenge to the field. The prime reason is that, before purification, membrane proteins have to be extracted from their bilayer and solubilized using detergents. Most experience has to date been accumulated with bacterial membrane proteins due to their higher stability compared to eukaryotic membrane proteins. The accumulating know-how in expressing sufficient amounts of membrane proteins in various host systems and in the solubilization and purification of functionally intact membrane proteins over the past years has resulted in an increasing number of membrane protein structures entering the protein structure databank.

Highlights in this area are the structures of various K$^+$ channels,[90] of complexes of the respiratory chain,[91] of the photosystem,[92] of the lac-permease,[93] and of ABC-transporters.[94]

22.4 Future Outlook

As described in the above sections, X-ray crystallography has undergone major developments over the last 70 years of its existence and is established as a key method in biology to describe the architecture of proteins and protein assemblies. This structural information often is the basis for a detailed understanding of the function of these molecules. With the advances already achieved and additional developments in the future, we will see an increase in complexity of the molecules studied (as long as the complexes are stable). For membrane proteins, the biochemical methods still lag about 40 years behind the one for soluble proteins, but it can be expected that further breakthrough developments will come in the near future due to the enormous support the field of structure determination of membrane proteins is experiencing at the moment. On the technology side, the most exciting is the development of the FEL, which has the potential of revolutionizing the field again, allowing maybe even snapshots of different states of macromolecules (fsec scale), which opens a new dimension to understanding their function.

Acknowledgement

The authors acknowledge financial support by the Schweizerische Nationalfonds and the University of Zurich within the framework of the National Center of Competence in Research (NCCR) Structural Biology program.

References

1. Bernal JD, Crowfoot DC. (1934) X-ray photographs of crystalline pepsin. *Nature* **133**: 794–795.
2. Kendrew JC, Dickerson RE, Strandberg BE, *et al.* (1960) Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution. *Nature* **185**: 422–427.

3. Perutz MF, Rossmann MG, Cullis AF, *et al.* (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature* **185**: 416–422.
4. Matthews BW. (2003) Transformations in structural biology: a personal view. *Meth Enzymol* **368**: 3–11.
5. Terwilliger TC. (2000) Structural genomics in North America. *Nat Struct Biol* 7(Suppl): 935–939.
6. Yokoyama S, Hirota H, Kigawa T, *et al.* (2000) Structural genomics projects in Japan. *Nat Struct Biol* (7): 943–945.
7. Heinemann U. (2000) Structural genomics in Europe: slow start, strong finish? *Nat Struct Biol* 7(Suppl): 940–942.
8. (2006) Structural proteomics in Europe. *Acta Crystallogr D Biol Crystallogr* **62**: 0–1285.
9. McPherson A. (1982) *Preparation and Analysis of Protein Crystals.* Wiley, New York.
10. Ducruix A, Giegé R. (1992) *Crystallization of Nucleic Acids and Proteins. A Practical Approach.* Oxford University Press, Oxford.
11. Bergfors TM. (1999) *Protein Crystallization: Techniques, Strategies and Tips.* International University Line, La Jolla.
12. Garcia-Ruiz JM. (2003) Counterdiffusion methods for macromolecular crystallization. *Meth Enzymol* **368**: 130–154.
13. Zeppezauer M, Eklund H, Zeppezauer ES. (1968) Micro diffusion cells for the growth of single protein crystals by means of equilibrium dialysis. *Arch Biochem Biophys* **126**: 564–573.
14. Chayen NE, Shaw Stewart PD, Maeder DL, Blow DM. (1990) An automated system for micro-batch protein crystallization and screening. *J Appl Cryst* **23**: 297–302.
15. Jancarik J, Kim S-H. (1991) Sparse matrix sampling: a screening method for crystallization of proteins. *J Appl Cryst* **24**: 409–411.
16. Michel H, Oesterhelt D. (1980) Three-dimensional crystals of membrane proteins: bacteriorhodopsin. *Proc Natl Acad Sci USA* 77: 1283–1285.
17. Garavito RM, Rosenbusch JP. (1980) Three-dimensional crystals of an integral membrane protein: an initial X-ray analysis. *J Cell Biol* **86**: 327–329.
18. Teng T-Y. (1990) Mounting of crystals for macromolecular crystallography in a free-standing thin film. *J Appl Cryst* **23**: 387–391.
19. Garman EF, Schneider TR. (1997) Macromolecular Cryocrystallography. *J Appl Cryst* **30**: 211–237.
20. Arndt UW, Duncumb P, Long JVP, Pina L, Inneman A. (1998) Focusing mirrors for use with microfocus X-ray tubes. *J Appl Cryst* **31**: 733–741.
21. Arndt UW. (2003) Personal X-ray reflections. *Meth Enzymol* **368**: 21–42.
22. Walsh MA, Dementieva I, Evans G, Sanishvili R, Joachimiak A. (1999) Taking MAD to the extreme: ultrafast protein structure determination. *Acta Crystallogr D Biol Crystallogr* **55**: 1168–1173.

23. Miao J, Sayre D, Chapman HN. (1998) Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects. *J Optic Soc Am A* **15**: 1662–1669.
24. Miao J, Hodgson KO, Sayre D. (2001) An approach to three-dimensional structures of biomolecules by using single-molecule diffraction images. *Proc Natl Acad Sci USA* **98**: 6641–6645.
25. Huldt G, Szoke A, Hajdu J. (2003) Diffraction imaging of single particles and biomolecules. *J Struct Biol* **144**: 219–227.
26. Brönnimann C, Bühler C, Eikenberry EF, *et al.* (2004) *Synchrotron Radiat News* **17**: 23–30.
27. Leslie AG. (1992) Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4/ ESF-EAMCB Newslett Protein Crystallogr* **26**.
28. Otwinowski Z, Minor W. (1997) Processing of X-ray diffraction data collected in oscillation mode. *Meth Enzymology* **276**: 307–326.
29. Kabsch W. (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J Appl Cryst* **26**: 795–800.
30. Kendrew JC, Bodo G, Dintzis HM, *et al.* (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**: 662–666.
31. Perutz MF. (1956) Isomorphous replacement and phase determination in non-centrosymmetric space groups. *Acta Crystallogr* **9**: 867–873.
32. Drenth J. (1999) *Principles of Protein X-Ray Crystallography*. Springer-Verlag, New York, Berlin, Heidelberg.
33. Garman E, Murray JW. (2003) Heavy-atom derivatization. *Acta Crystallogr D Biol Crystallogr* **59**: 1903–1913.
34. Walsh MA, Evans G, Sanishvili R, Dementieva I, Joachimiak A. (1999) MAD data collection — current trends. *Acta Crystallogr D Biol Crystallogr* **55**: 1726–1732.
35. Gonzalez A. (2003) Optimizing data collection for structure determination. *Acta Crystallogr D Biol Crystallogr* **59**: 1935–1942.
36. Hendrickson WA. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**: 51–58.
37. Hendrickson WA, Teeter MM. (1981) Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature* **290**: 107–113.
38. Dauter Z, Dauter M, Rajashankar KR. (2000) Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. *Acta Crystallogr D Biol Crystallogr* **56**: 232–237.
39. Dauter Z, Dauter M, de La Fortelle E, Bricogne G, Sheldrick GM. (1999) Can anomalous signal of sulfur become a tool for solving protein crystal structures? *J Mol Biol* **289**: 83–92.

40. Yang C, Pflugrath JW, Courville DA, Stence CN, Ferrara JD. (2003) Away from the edge: SAD phasing from the sulfur anomalous signal measured in-house with chromium radiation. *Acta Crystallogr D Biol Crystallogr* **59**: 1943–1957.
41. Ogata CM. (1998) MAD phasing grows up. *Nat Struct Biol* **5**(Suppl): 638–640.
42. Schoenborn BP, Watson HC, Kendrew JC. (1965) Binding of xenon to sperm whale myoglobin. *Nature* **207**: 28–30.
43. Schiltz M, Shepard W, Fourme R, *et al*. (1997) High-pressure krypton gas and statistical heavy-atom refinement: a successful combination of tools for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **53**: 78–92.
44. Yao JX. (1983) On the application of phase relationships to complex structures. XX. RANTAN for large structures and fragment development. *Acta Crystallogr A* **39**: 35–37.
45. Schneider TR, Sheldrick GM. (2002) Substructure solution with SHELXD. *Acta Crystallogr D Biol Crystallogr* **58**: 1772–1779.
46. Yao JX, Dodson EJ, Wilson KS, Woolfson MM. (2006) ACORN: a review. *Acta Crystallogr D Biol Crystallogr* **62**: 901–908.
47. Terwilliger T. (2004) SOLVE and RESOLVE: automated structure solution, density modification and model building. *J Synchrotron Radiat* **11**: 49–52.
48. Vonrhein C, Blanc E, Roversi P, Bricogne G. (2006) Automated structure solution with autoSHARP. *Meth Mol Biol* **364**: 215–230.
49. de La Fortelle E, Bricogne G. (1997) *Maximum-Likelihood Heavy-Atom Parameter Refinement for the Multiple Isomorphous Replacement and Multiwavelength Anomalous Diffraction Methods*. Academic Press, New York.
50. Cowtan K. (1994) ‘dm’: an automated procedure for phase improvement by density modification. *Joint CCP4 and ESF-EACBM Newslett Protein Crystallogr* **31**: 34–38.
51. Abrahams JP, Leslie AG. (1996) Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Crystallogr D Biol Crystallogr* **52**: 30–42.
52. Perrakis A, Morris R, Lamzin VS. (1999) Automated protein model building combined with iterative structure refinement. *Nat Struct Biol* **6**: 458–463.
53. Lamzin VS, Perrakis A, Wilson KS. (2001) The ARP/wARP suite for automated construction and refinement of protein models. In Rossmann, MG, Arnold, E (eds.), *International Tables for Crystallography*, pp. 720–722. Kluwer Academic Publishers, Dordrecht.
54. Pape T, Schneider TR. (2004) HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs. *J Appl Cryst* **37**: 843–844.
55. Sheldrick GM. (2002) Macromolecular phasing with SHELXE. *Z Kristallogr* **217**: 644–650.
56. Rossmann MG, Blow DM. (1962) The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr* **15**: 24–31.

57. Navaza J. (1994) AMoRe: an automated package for molecular replacement. *Acta Crystallogr A* **50**: 157–163.
58. Brünger AT, Adams PD, Clore GM, *et al.* (1998) Crystallography & NMR System: a new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **54**: 905–921.
59. Vagin A, Teplyakov A. (1997) MOLREP: an automated program for molecular replacement. *J Appl Cryst* **30**: 1022–1025.
60. Jamrog DC, Zhang Y, Phillips GN, Jr. (2003) SOMoRe: a multi-dimensional search and optimization approach to molecular replacement. *Acta Crystallogr D Biol Crystallogr* **59**: 304–314.
61. Kissinger CR, Gehlhaar DK, Fogel DB. (1999) Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr D Biol Crystallogr* **55**: 484–491.
62. Glykos NM, Kokkinidis M. (2001) Multidimensional molecular replacement. *Acta Crystallogr D Biol Crystallogr* **57**: 1462–1473.
63. Storoni LC, McCoy AJ, Read RJ. (2004) Likelihood-enhanced fast rotation functions. *Acta Crystallogr D Biol Crystallogr* **60**: 432–438.
64. McCoy AJ, Grosse-Kunstleve RW, Storoni LC, Read RJ. (2005) Likelihood-enhanced fast translation functions. *Acta Crystallogr D Biol Crystallogr* **61**: 458–464.
65. Long F, Vagin A, Young P, Murshudov GN. (2008) BALBES: a molecular replacement pipeline. *Acta Crystallogr D Biol Crystallogr* **64**: 125–132.
66. Keegan RM, Winn MD. (2007) Automated search-model discovery and preparation for structure solution by molecular replacement. *Acta Crystallogr D Biol Crystallogr* **63**: 447–457.
67. Collaborative Computational Project N. (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* **50**: 760–763.
68. Jones TA. (1978) A graphics model building and refinement system for macromolecules. *J Appl Cryst* **11**: 268–272.
69. Jones TA, Zou J-Y, Cowan SW, Kjeldgaard M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr A* **47**: 110–119.
70. McRee DE. (1999) XtalView/Xfit — a versatile program for manipulating atomic coordinates and electron density. *J Struct Biol* **125**: 156–165.
71. Turk D. (1996) *MAIN 96: An Interactive Software for Density Modifications, Model Building, Structure Refinement and Analysis.*
72. Emsley P, Cowtan K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**: 2126–2132.
73. Murshudov GN, Vagin AA, Dodson EJ. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53**: 240–255.

74. Sheldrick GM, Schneider TR. (1997) SHELXL: high resolution refinement. *Meth Enzymology*: 319–343.
75. Engh RA, Huber R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr A* **47**: 392–400.
76. Ten Eyck LF. (1977) Crystallographic fast Fourier transforms. *Acta Crystallogr A* **29**: 183–191.
77. Tronrud DE. (2004) Introduction to macromolecular refinement. *Acta Crystallogr D Biol Crystallogr* **60**: 2156–2168.
78. Schomaker V, Trueblood KN. (1968) On the rigid-body motion of molecules in crystals. *Acta Crystallogr B* **24**: 63–76.
79. Afonine PV, Grosse-Kunstleve RW, Adams PD. (2005) The Phenix refinement framework. *CCP4 Newslett July*.
80. Perrakis A, Sixma TK, Wilson KS, Lamzin VS. (1997) wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. *Acta Crystallogr D Biol Crystallogr* **53**: 448–455.
81. Cowtan K. (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* **62**: 1002–1011.
82. Cowtan K. (2000) General quadratic functions in real and reciprocal space and their application to likelihood phasing. *Acta Crystallogr D Biol Crystallogr* **56**: 1612–1621.
83. Yao M, Zhou Y, Tanaka I. (2006) LAFIRE: software for automating the refinement process of protein-structure analysis. *Acta Crystallogr D Biol Crystallogr* **62**: 189–196.
84. Rahuel J, Rasetti V, Maibaum J, *et al.* (2000) Structure-based drug design: the discovery of novel nonpeptide orally active inhibitors of human renin. *Chem Biol* 7: 493–504.
85. Cramer P, Bushnell DA, Kornberg RD. (2001) Structural basis of transcription: RNA polymerase II at 2.8 A resolution. *Science* **292**: 1863–1876.
86. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **289**: 905–920.
87. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**: 251–260.
88. Groll M, Ditzel L, Lowe J, *et al.* (1997) Structure of 20S proteasome from yeast at 2.4 A resolution. *Nature* **386**: 463–471.
89. Fry EE, Lea SM, Jackson T, *et al.* (1999) The structure and function of a foot-and-mouth disease virus-oligosaccharide receptor complex. *EMBO J* **18**: 543–554.
90. Long SB, Campbell EB, Mackinnon R. (2005) Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. *Science* **309**: 897–903.

91. Lange C, Nett JH, Trumpower BL, Hunte C. (2001) Specific roles of protein-phospholipid interactions in the yeast cytochrome bc1 complex structure. *EMBO J* **20**: 6591–6600.
92. Ferreira KN, Iverson TM, Maghlaoui K, Barber J, Iwata S. (2004) Architecture of the photosynthetic oxygen-evolving center. *Science* **303**: 1831–1838.
93. Abramson J, Smirnova I, Kasho V, *et al*. (2003) Structure and mechanism of the lactose permease of *Escherichia coli*. *Science* **301**: 610–615.
94. Hollenstein K, Frei DC, Locher KP. (2007) Structure of an ABC transporter in complex with its binding protein. *Nature* **446**: 213–216.

Chapter 23

New Frontiers in High-Resolution Electron Microscopy

A. Engel*

23.1 Introduction

The sentence, "seeing is believing," reflects the importance of electron microscopy in basic and applied research. Modern biology textbooks document this even better: micrographs are key for deriving pictorial representations to foster our understanding of tissues, cells, organelles, and biological nanomachines. Without such pictures, it would not be possible to bring all the biochemical and atomic scale structural information from crystallography and NMR into a cellular context.

This chapter concentrates on the progress in high-resolution electron microscopy techniques achieved in the past decade. Main advances came through novel sample preparation methods preserving specimens in their most native state, by the development of automated data acquisition, and through important refinements of the instruments. To prevent dehydration of samples that must reside in the vacuum of an electron optical system, macromolecular complexes are now vitrified in a solution layer by rapid freezing,[1] while cells and tissue samples are high-pressure frozen and cryo-sectioned.[2] Such

*M.E. Müller Institute, Biozentrum, Klingelbergstrasse 70, 4056 Basel, Switzerland. Email: andreas.engel@unibas.ch.

frozen specimens required the development of ultra-stable cold stages, maintaining a temperature below the phase transition temperature of vitrified water to hexagonal ice (138 K) — a goal reached by liquid Nitrogen cooling. The ultimate stage, however, will keep the sample at a few Kelvin through cooling with liquid helium, not mainly to prevent crystallization of vitrified water, but rather to minimize the beam-induced damage of the sample.[3] Along with cold stages came a significant improvement of the vacuum systems to inhibit the trapping of water or hydrocarbons by the cold sample. Moreover, observation of native, frozen specimens fostered the development of elegant low-dose data acquisition systems for single micrographs and also entire tilt-series as required for electron tomography.[4] To achieve optimum information transfer to atomic scale resolution, field emission electron guns combined with acceleration voltages of 200–300 kV are becoming popular, which is a great improvement for high-resolution electron microscopy. Such progress emerging from leading laboratories and from microscope manufacturers enables researchers to study the proteome of a cell by electron tomography[5] and resolve the structure of a membrane protein at a resolution of better than 2 Å by electron crystallography.[6]

Altogether, such progress makes high-resolution electron microscopes essential tools for the structural biologist who has an open mind and attempts to use all the available techniques to obtain a deeper understanding of the structures of life.

23.2 Sample Preparation Methods

For electron microscopy, samples need to be either dehydrated or quick-frozen and kept at temperatures close to liquid nitrogen for transfer to the vacuum of an electron optical system. Dehydration not only changes the native environment of biological samples, but also exposes them to surface tension forces. Embedding aqueous suspensions of macromolecular complexes or membrane fragments in a heavy metal salt solution provides support against surface tension induced compression of the biological structure during dehydration, and it produces a high contrast due to the strong scattering of the heavy atoms. Heavy metal salt solutions create unphysiological ion strength, and for uranyl salt, a pH of around 4. Embedding the biological macromolecules in a sugar solution provides similar support and maintains some hydration, yet the contrast produced by such samples is low, since sugar and protein have a rather similar electron scattering power. Therefore, this preparation method is mainly used for electron crystallography,[3] because only ordered structure contributes to the crystallographic signal but not randomly arranged sugar molecules.

Quick freezing, and hence, vitrification of thin solution layers containing biological macromolecules was developed by Dubochet and colleagues,[1,7] who demonstrated the strength of this approach by numerous examples, convincing others to apply it to their samples. Vitrification is now a routine method that is widely used. However, its limitation concerns the maximal dimensions of the sample to be vitrified: only small or flat cells can be vitrified in this way. For tissues and large cells, high pressure freezing is the method of choice. It has been developed in several laboratories and is mainly applied when combined with substitution of water with organic solvents for plastic embedding at low temperature. The ultimate approach, again developed by Dubochet and co-workers, is the preparation of thin sections from vitrified blocks.[2,8,9] Combining cryo-sectioning with electron tomography promises to give an insight into the native contents of a cell at nm-scale resolution.[10]

23.3 Information Transfer

23.3.1 *Image Formation*

Modern electron optical systems comprise field emission electron guns operated at 100–300 kV, which provide a highly coherent illumination and a wavelength λ of 0.04-0.02 Å. Magnetic lenses with spherical aberrations around 2–4 mm shape the illuminating beam and collect the electron scattered by the object to form an image at typically 50–100 000 fold magnification. Owing to the short wave-

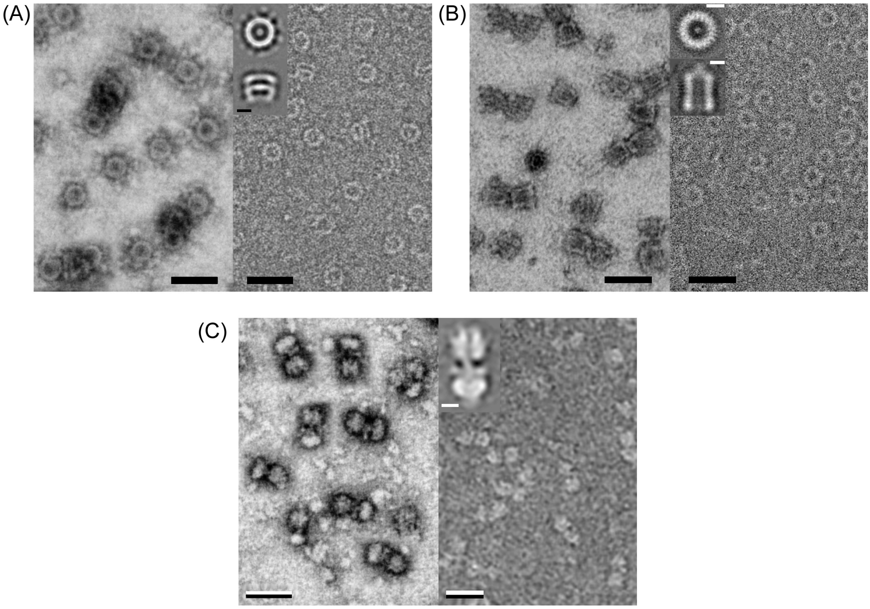


Fig. 23.1. Comparison of samples prepared by negative staining and vitrification. The latter warrants the preservation of the native structure, whereas the former allows structural details to be identified on single complexes. Negatively stained samples were recorded by an annular detector dark-field scanning transmission electron microscope operated at 100 kV and vitrified samples by a 200 kV transmission electron microscope equipped with a field-emission gun. Negatively stained samples are on the left, vitrified samples with an average shown in the inset on the right. **(A)** The bacterial outer membrane secretin PulD, a protein translocation apparatus. Flexible radial extensions are distinct after negative staining, while they are only visible in the projections of vitrified samples after averaging.[17] **(B)** The bacterial toxin ClyA is a cylindrical complex built of 13 subunits. Stripes running parallel to the cylinder axis visible in negatively stained complexes suggest elongated rod-shaped ClyA proteins. Significant flattening is indicated by the larger width/length ratio of cylinder side-views compared to that of vitrified samples. A 12 Å 3D map of the ClyA pore-forming complex has been calculated from projections of the vitrified preparation.[18] **(C)** Bovine V-ATPase comprises the membrane resident V0 part and the soluble V1 part, the latter often exhibiting a crown shape.[66] The scale bars in the micrographs correspond to 25 nm, whereas those in the insets represent 5 nm.

lengths, atomic scale resolution can be reached at a small collection angle, $\theta \approx 0.61\ \lambda/\mathrm{d}$, hence, electron optical systems exhibit a large depth of focus, as indicated by Equation 23.1:

$$D \leq \frac{d^2}{0.61\lambda} \tag{23.1}$$

Here, D is the depth of focus and d the resolution to be achieved. In addition, it is assumed that diffraction limited resolution can be reached after CTF correction (see below).

Electrons interact strongly with matter, making it possible to depict thin objects such as 2D protein crystals, viruses, or small cells. Electrons are elastically scattered by the nuclei of the atoms, which are orders of magnitude heavier than that of moving electrons. Electrons are inelastically scattered by the inner- and outer shell electrons, to which they transmit a fraction of their kinetic energy. While elastic electrons contribute to the coherent axial bright-field image that carries the high-resolution information on the 3D arrangement of the sample atoms, inelastic electrons carry interesting chemical information. However, inelastic scattering is directly related to the beam-induced specimen damage.

Since only the elastically scattered electrons contribute to a high-resolution image, the coherent phase contrast image formation is considered, also called an axial bright-field image. A thin object that comprises only light elements, and whose thickness is within the limits described in Fig. 23.2, is approximately described as a weak phase object:

$$t(x,y) = 1 + i\phi(x,y), \quad \phi(x,y) < \pi/4 \tag{23.2}$$

The function $t(x,y)$ represents the two-dimensional (2D) projection of the 3D object. The amplitude distribution in the image plane is the coherent superposition of the unscattered wave and the

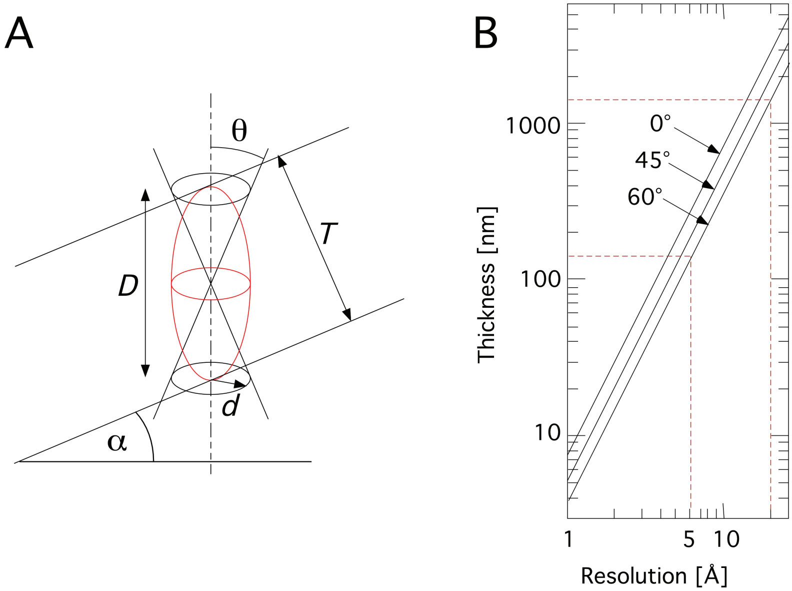


Fig. 23.2. Depth of focus and resolution. **(A)** The diffraction limit dictates the ultimate resolution of an optical system: $d = 0.61\ \lambda/\Theta$. The depth of focus D corresponds to the distance between locations where the diffraction disc and the geometric discs are equal. Scattering centers within the resolution volume indicated in red cannot be resolved. An optical system that collects waves emanating from scattering centers of a tilted 3D object whose projected thickness $T/\cos\,\alpha$ is smaller than D will produce the true 2D projection of the 3D object in the image plane. **(B)** Log-log plot of estimated resolution limit against the depth of focus D, as given by Eq. 23.1. The ordinate displays the maximal thickness T of the sample, which would be within the depth of focus D for a given resolution and a given tilt angle. For example, at a tilt of 60°, a resolution of 5 Å is achievable at 300 kV acceleration voltage with a sample thinner than 200 nm. If a resolution of 20 Å is to be reached, the sample slab can be even 2 μm thick (indicated by the dotted lines).

elastically scattered waves. For an optical system whose point spread function $h(x,y) = h_r(x,y) + i\ h_i(x,y)$ is space-invariant, this superposition is described as the convolution of $t(x,y)$ with $h(x,y)$:

$$a(x,y) = (h_r(x,y) + ih_i(x,y)) \otimes (1 + i\phi(x,y)) \qquad (23.3)$$

The intensity $|a(x,y)|^2$ is recorded on the film. Neglecting quadratic terms, the image can thus be approximated as:

$$|a(x,y)|^2 = 1 - 2h_i(x,y) \otimes \phi(x,y) \qquad (23.4)$$

The imaginary part, $h_i(x,y)$, is described by the inverse Fourier transform, FT^{-1}, of the phase contrast transfer function (CTF):

$$h_i(x,y) = FT^{-1}\left[A(p)\sin\{\pi(C_S\lambda^3 p^4/2 + \Delta f \lambda p^2)\}\right] \qquad (23.5)$$

where $A(p)$ describes the envelope of the CTF, C_s is the spherical aberration constant, Δf the defocus, λ the electron wavelength (about 0.02 Å for 300 kV electrons), and p the distance from the origin in the reciprocal space. The CTF for weak phase objects is displayed in Fig. 23.3. Since the electron optical system introduces a phase difference between the scattered and unscattered electrons of $\pi/2$, the axial bright-field mode corresponds to the Zernike phase contrast mode of light microscopy. The contrast is weak when the microscope is operated close to focus because the prominent low-resolution features of the specimen are transferred with small amplitude (Fig. 23.3, CTF labeled Scherzer, 56 nm). However, the contrast can be enhanced by moving out of focus, since frequency bands exhibiting a transfer coefficient >0.5 move towards lower resolution (Fig. 23.3, CTF labeled 1000 nm). Alternatively, electron optical phase plates akin to the Zernike phase plate have been explored a long time ago, but their value has been recognized only recently.[11] In any case, the phase shift introduced by the electron optical system has to be corrected to facilitate the image interpretation (see below). The phase shift of electrons scattered elastically by an atom is proportional to the coulomb-potential of this atom. Therefore, the ensemble of all electrons singly scattered by a specimen produce a projection of its

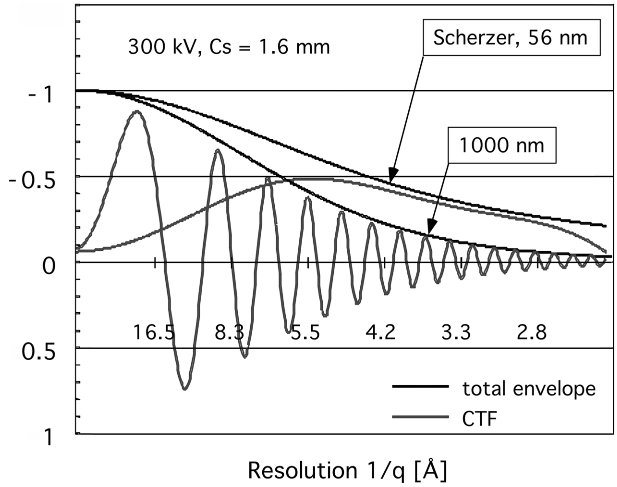


Fig. 23.3. The contrast transfer function of a 300 kV field emission transmission electron microscope. At an underfocus of 56 nm, the information is transferred without phase reversal, but low spatial frequencies exhibit a low contrast. At 1000 nm underfocus, information is transferred from about 30 Å up to a resolution of 3 Å, albeit with much reduced contrast above 5 Å compared to that at 30 Å. The phase reversal at such an underfocus is corrected computationally *a posteriori*.

coulomb potential, which is dominated by the atom's nuclei rather than the electron shells as in X-ray crystallography. It should be noted that this simplification is related to the large depth of focus provided by high-voltage transmission electron microscopes, as indicated by Equation 23.1.

Figure 23.4 shows the power spectrum of a 2D crystal image. This spectrum reveals (i) discrete spots representing the crystal information (see below), (ii) concentric rings (Thon-rings) with gaps in-between (the zero-crossings of the CTF, see Fig. 3), where no structural information is available, and (iii) the envelope modulation decreasing the signal intensity at higher resolutions. Thon rings and envelope function reflect the specific nature of the CTF. The decrease of contrast towards high resolution (envelope function) results from the partial incoherence of the electron beam. Modern electron

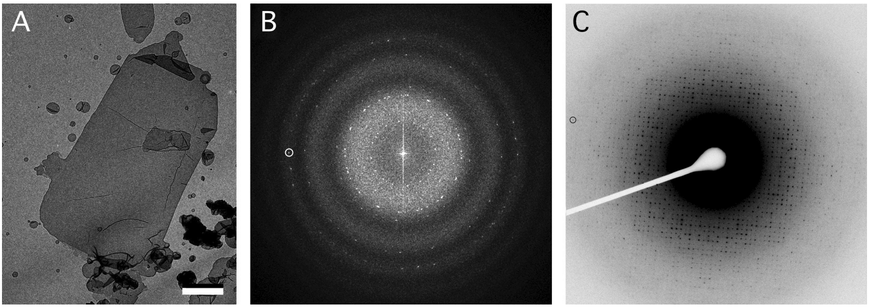


Fig. 23.4. The electron micrograph of a large 2D crystal does not reveal its crystallinity, but its homogeneity **(A)**. The scalebar represents 0.5 μm. The optical transform of such a micrograph **(B)** shows sharp diffraction maxima and the effect of the CTF. The reflection indicated is at 9.1 Å resolution. The electron diffraction pattern of a similar 2D crystal **(C)** exhibits diffraction maxima with a distinct four-fold symmetry. The reflection indicated is at 2 Å resolution.

microscopes are equipped with field emission guns exhibiting a high degree of coherence so that the contrast decrease is acceptable even at high resolution. The Thon-rings with alternating positive and negative contrast are the result of the phase shift introduced by the objective lens.

The great advantage of a modern field-emission electron microscope to directly acquire the phase information out to atomic scale resolution is watered down by several experimental difficulties. First, the instrument has to be stable and installed in a field- and vibration-free environment, prerequisites that are routinely reached. Second, beam-induced damage not only changes the specimen structure, but also leads to specimen charging. Since these charges act like an electrostatic lens, the focus changes during image recording. If the sample plane is perpendicular to the optical axis, the overall effect is quite small: the focus change occurring during image acquisition may have an influence only at very high resolution. In this case, it is possible to measure the zero-crossings of the CTF accurately [see Fig. 23.4(B)] and to correct the defects in the Fourier transform of the micrograph to retrieve both the phase and amplitude information. After this CTF-correction, the electron microscope performs close to a diffraction limited optical system. However, when the sample is tilted for collecting the 3D information (see below), the electrostatic lens building up during irradiation introduces an image shift, a problem that cannot always be solved satisfactorily.[12] Third, a further optical defect is related to the changing focus for tilted specimens: the point-spread function is not space invariant, and the commonly used CTF correction is only an approximate measure to eliminate the phase distortions of the electron optical system.[13]

23.3.2 *Electron Diffraction*

When a crystal is irradiated by a parallel beam and the diffracted beams are brought to focus in the diffraction plane by the optical system, the resulting pattern can be recorded by a CCD camera, for which the dynamic range is far better than that of photographic film. Electron diffraction is neither affected by the CTF, the envelope function, nor specimen charging. Moreover, the depth of focus is even larger than in the case of imaging, since a small spread of diffraction spots can be tolerated. Therefore, electron diffraction is much more effective for the collection of high-resolution information than imaging, although the phase information is not retrieved [Fig. 23.4(C)]. The directly measured amplitudes can be combined with the phases from the images during image processing. Electron diffraction is not an absolute requirement for determining a structure, but it allows a fast judgment of the crystal quality, helps in correcting the CTF, and provides suitable high-resolution information for molecular replacement methods.[6]

23.3.3 *Scanning Transmission Electron Microscopy*

The scanning transmission electron microscope (STEM) collects almost all scattered electrons, hence, making use of all the information transferred by electron scattering most efficiently.[14] As illustrated

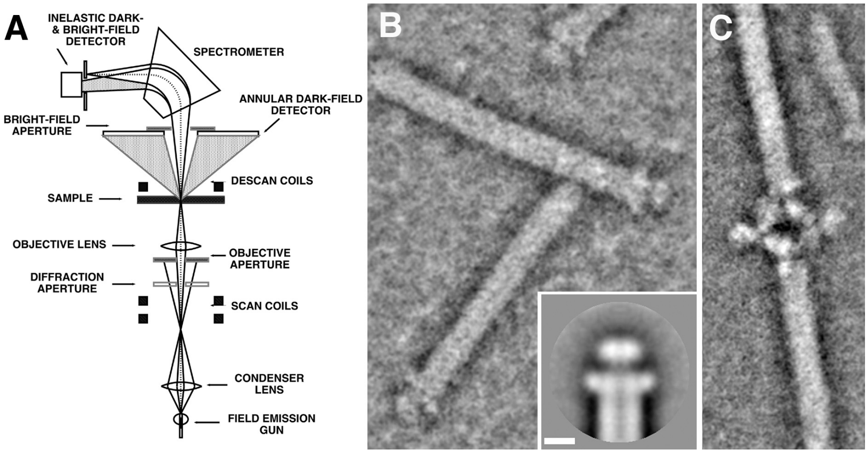


Fig. 23.5. The STEM is the ideal instrument to perform electron scattering experiments. **(A)** In the STEM, the majority of elastically scattered electrons are collected via an annular detector, and the inelastically scattered electrons via a spectrometer, generating an elastic and an inelastic dark-field image. Unscattered and forward scattered elastic electrons produce the same phase contrast image as the axial bright-field transmission electron microscope. **(B)** The superb contrast and lack of interference fringes make elastic dark-field images attractive, as they are easy to interpret. Negatively stained needles of the *Yersinia* injectosomes exhibit a particular tip complex. **(C)** Using antibodies, the protein assembling the tip complex was identified as LcrV, the V-antigen known to produce resistance against *Yersinia pestis* since seven decades.[20]

in Fig. 23.5, several signals can be collected in parallel, yielding an elastic dark-field, an inelastic dark-field, and an axial bright field image. Taking the same nomenclature as above, the elastic dark-field image intensity distribution is written as:

$$|a(x,y)|^2 = |h_i(x,y)|^2 \otimes |\phi(x,y)|^2 \tag{23.6}$$

Both dark-field images are of use in biological applications, as both provide a quantitative measure of the mass of the protein complex being imaged. Such measurements are attractive, as the STEM can determine the molecular weights of single complexes over a range of 50 kDa to 100 MDa, and this, even with rather heterogeneous preparations. Moreover, as the mass is derived from the dark-field image of a complex simply by integration over an area that includes the complex and subtraction of the background resulting from the carbon film, the STEM allows a link between the mass and shape of a protein complex. It is of particular interest to combine STEM and modern mass spectroscopy approaches for single-cell visual proteomics.[14] Since STEM mass analyses can be used for very large and heterogeneous assemblies, it has been the tool of choice to help in merging a large set of different measurements of synaptic vesicles towards producing an accurate description of this dynamic machinery.[15] To know the mass of a protein complex is also a great help to interpret their 3D maps[16] or to bootstrap the 3D reconstruction of homo-oligomers from the knowledge of the stoichiometry.[17–19]

The STEM is also producing images of exceptional contrast and clarity of negatively stained samples. This is often key to straightforward identification of proteins within a complex using antibody labeling.[20]

23.4 Electron Crystallography

To exploit the capacity of the electron microscope to acquire amplitude and phase information for crystallographic measurements, the primordial prerequisite is the availability of highly ordered thin crystals. They should exhibit lateral dimensions of several microns over which the crystallinity should be perfect. Electron crystallography is of special interest for membrane proteins that are crystallized in the presence of lipids, which reconstitutes the proteins in their native environment, the lipid bilayer. Provided that the protein of interest is available in mg quantities, it can be crystallized in a functional state, as demonstrated for the water channel AQP1.[21]

23.4.1 *Different Methods for 2D Crystallization*

Two-dimensional crystals consisting of membrane proteins and lipids can be produced in three different ways.[22] The first method involves the induction of regular packing of a highly abundant protein in its native membrane. This is achieved by eliminating interspersed lipids using lipases[23] or by extracting lipids with specific detergents.[24] Although this is the most gentle 2DX method, because it does not require the solubilization of the membrane protein, it is not generally applicable.

The second method reconstitutes the purified membrane protein into a lipid bilayer at high protein density.[25] The detergent solubilized protein is mixed with solubilized lipids to form a homogenous solution of mixed protein-detergent and lipid-detergent micelles. Detergent removal then results in the formation of protein aggregates in the worst case, and in the progressive formation of proteoliposomes with large 2D crystalline regions in the best case. Reconstitution begins once the detergent concentration reaches the critical micellar concentration (CMC).[26] The respective affinities between the components of the ternary mixture dictate the progress of the reconstitution process. Ideally, a starting condition should be established where mixed detergent-protein and mixed detergent-lipid micelles have exchanged their constituents to the extent that the mixture consists mainly of ternary detergent-protein-lipid micelles. Assuming the protein remains in its native, properly folded state during the solubilization and isolation steps, this ideal situation is likely to foster perfect reconstitution and possibly 2DX of a functional membrane protein.

The third method concerns the reconstitution of the membrane proteins at the water-air interface by attaching the solubilized membrane protein to an active lipid monolayer prior to detergent removal.[27] In this process, membrane proteins are concentrated at the monolayer, brought into a planar configuration, and finally squeezed together during detergent removal. This approach is useful for membrane proteins that are present in small amounts and are stably solubilized only in low CMC detergents.

What all the methods summarized in Fig. 23.6 have in common is that the detergent is brought below its CMC to foster the assembly of a bilayer, into which the membrane protein should integrate. Generally used methods to bring the detergent concentration below the CMC include dialysis,[25,26] adsorption of the detergent to Bio-Beads,[28] and

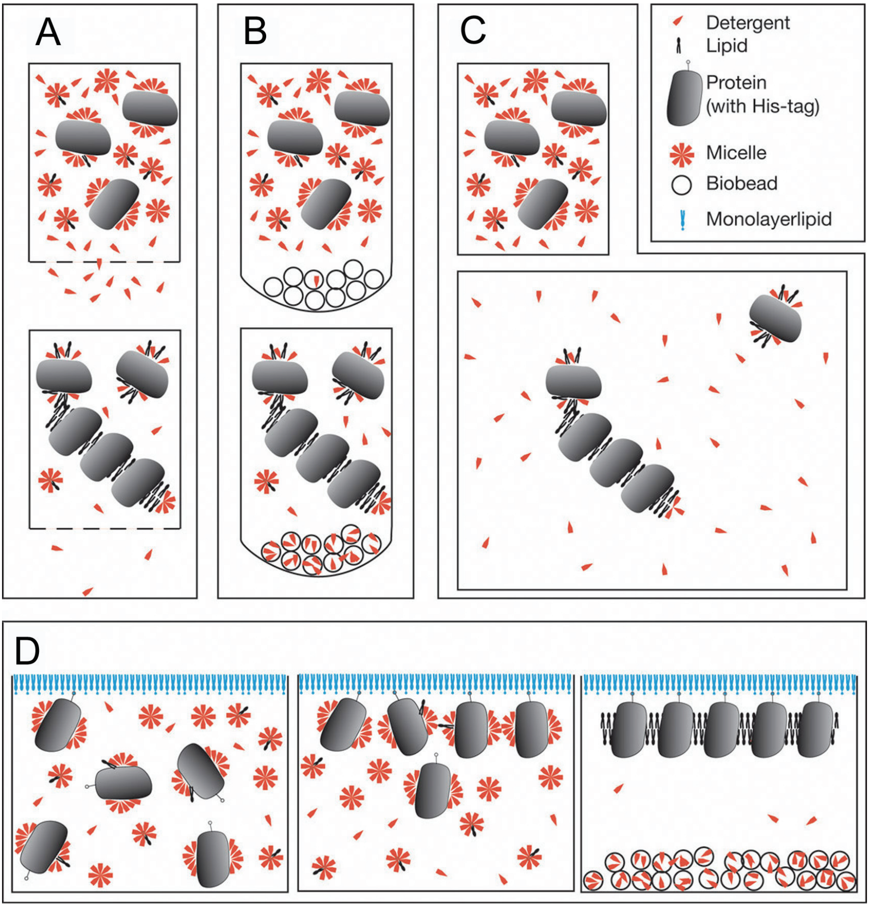


Fig. 23.6. 2D crystallization methods. All methods are based on the principle to bring the detergent concentration in the aqueous phase below the critical micelle concentration (CMC), forcing the detergent in the mixed micelles to partition in the aqueous phase. As a result, mixed micelles merge to form larger structures, and ultimately, 2D crystals. **(A)** Dialysis can be used to remove the detergent provided its CMC is >1 mM. **(B)** Bio-Beads adsorb detergent molecules and can be used for all detergents. Bio-Beads driven 2D crystallization is particularly successful with low CMC detergents. **(C)** Dilution is a well-known method for functional reconstitution of membrane proteins. In spite of dilution, it is also suitable for 2D crystallization, because the protein is highly concentrated after integration in the bilayer. **(D)** The monolayer technique combines the Bio-Beads method with crystallization at the air-water interface. This method works only with low CMC detergents because of the necessity to preserve the lipid monolayer. The latter incorporates special lipids having a high affinity for the solubilized protein, e.g. by recognition of a specific tag. (By courtesy of Thomas Braun.)

the dilution of the ternary mixture.[29] Moreover, in all the methods, the amount of interspersed lipid must be minimized to ensure regular interactions between the membrane proteins. The pertinent interactions depend on the shape and surface charges of the components. For a given protein, the lipid-detergent mixture, pH, counter ions, and temperature must all be optimized. In addition, the concentration, the ratio of the respective components, and the detergent removal rate are critical. This gives a multidimensional parameter space that needs to be experimentally sampled, a similar task to that carried out in 3D crystallization screens. The difficulty of such experiments is the management of the screens and the assessment of results. With 2DX, the latter is particularly cumbersome because 2D crystals cannot be detected by light microscopy and screening by electron microscopy is time consuming.

23.4.2 *Data Acquisition*

Beam damage induced by inelastic interactions of impinging electrons with sample atoms dictates the maximum electron dose a 2D crystal may take before discernable structural changes occur. This dose depends on the sample temperature: typically recording doses of 20 electron/Å^2 can be applied if the sample is kept below 10 K.[3] Hence, not only the best possible electron optical system is required, but ideally the sample should be cooled to the temperature of liquid helium.

2D crystals are adsorbed to flat thin carbon films and dried in the presence of sugar solutions. To prevent charging, the crystals can also be sandwiched between two thin films.[12] Such samples are loaded to the cold-stage, and images are acquired after the low temperature is reached and the stage is equilibrated. To minimize beam damage, crystals of potential high quality are identified at low magnification by their uniform thickness and characteristic shape [Fig. 23.4(A)]. Often, the entire grid is rapidly scanned in this mode and positions of interest are stored. The microscope is then adjusted for recording the high-resolution data. Either diffraction patterns or images are recorded, the former with a high resolution CCD camera, the latter preferably with a photographic film. In special cases, it is advantageous to record both, a diffraction pattern, and thereafter, an image. This allows the respective diffraction pattern to be properly classified based on the phases retrieved from the image.[30]

23.4.3 *Data Processing*

Each image represents a projection of the crystal, or after Fourier transformation, a central section through the 3D Fourier transform (3DFT) of the crystal (see Fig. 23.7). Since 2D crystals are periodic in (x, y), but have a single unit cell thickness in (z), the molecular transform is a smooth function, which is sampled along z^* on lattice lines (h, k). Each diffraction pattern represents the central section through the intensity $|3\mathrm{DFT}|^2$. Therefore, data processing proceeds along the following general scheme:

(i) Foremost is the CTF correction of the image, which is achieved after the image has been Fourier transformed. As documented by Fig. 23.3, this is a critical step for retrieving the phase information out to high resolution, since a small error in the determination of the CTF will lead to a wrong correction since the CTF oscillates rapidly. A complication arises with images of tilted samples, such as that required for extracting the 3D information. Since the point spread function $h(x,y)$ is not space independent, the linear systems theory yielding Equation 23.3 cannot be applied. Methods to correct the optical defects in this case are being developed, but they are computationally costly.[13] An interesting correction algorithm has been developed early,[31] while a CTF correction scheme working on stripes of the image that are parallel to the tilt axis is a robust practical approach.

(ii) The lattice parameters are measured from the Fourier transform of an image (2DFT), in fact from its intensity $|2\mathrm{DFT}|^2$, or directly from the electron diffraction pattern. The lattice is fitted by a least square distance minimization to the diffraction peaks identified. This information is required to calculate the z^* for each reflection measured.

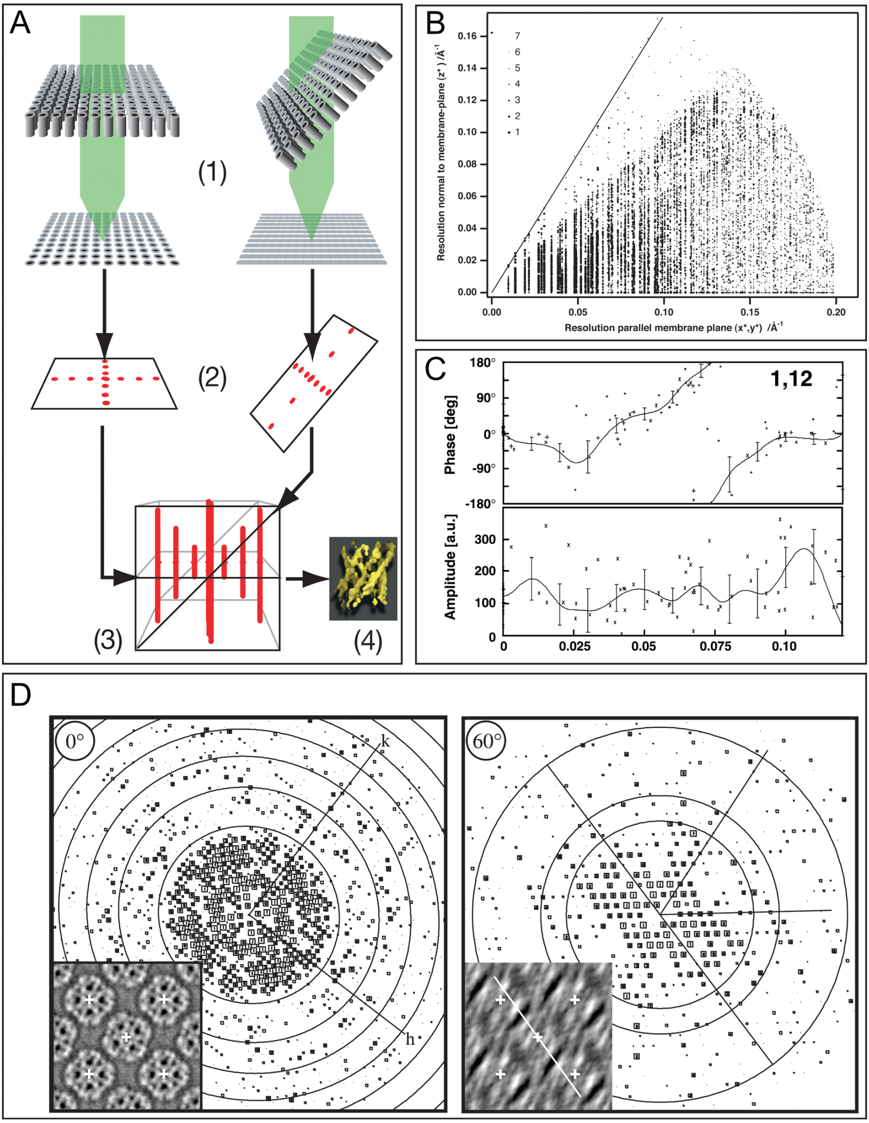


Fig. 23.7. 3D reconstruction of membrane proteins by 2D crystallography. **(A)** To obtain a 3D reconstruction from 2D crystals, projections are recorded at different tilt angles (step 1). The images are Fourier filtered and processed as described in Fig. 23.8 (step 2), and the Fourier transforms are combined in the 3D Fourier space according to the central section theorem (step 3). The discrete orders in the Fourier transform from the crystal are aligned in continuous lattice lines along *z** since the sample is not periodic in the

(iii) To extract the phase and amplitude information from the 2DFT, sharp diffraction peaks are required. Hence, all the parts of a 2D crystal that are disordered and would contribute to the background must be masked away or corrected by unbending. The latter is achieved by determining the position of all unit cells from the cross-correlation function with a first average unit cell projection obtained from the uncorrected 2D crystal. The displacement vector field between the fitted lattice and the actual unit cell positions is an ideal indicator of crystal quality (Fig. 23.8). The possibility of such corrections is the great advantage of recording an image of a 2D crystal rather than a diffraction pattern. Ultimately, it is possible to extract atomic scale resolution from small or fragmented crystals that do not exhibit sufficiently large highly ordered areas for electron diffraction. This procedure provides the averaged unit cell, or the projection map of the particular 2D crystal. On the other hand, electron diffraction of large highly ordered crystals deliver information to a resolution of 2 Å or beyond (see Fig. 23.4(C) and Ref. 6). The intensity of all peaks is measured by integration over the extent of the peak and subtraction of the local background. The signal-to-noise ratio obtained in this way is used to weight the contribution of the respective peak.

(iv) The data collected from many images or diffraction patterns needs to be merged to populate the 3DFT as shown in Fig. 23.7. To this end, the projection maps obtained in step (iii) need to be

z-direction. The lattice lines are regularly interpolated to sample the 3D Fourier space on a cubic raster. Back-transformation of the combined data finally leads to the representation of the 3D unit cell (step 4). **(B)** Azimuthal projection of the sampling in *z** direction. The different tilt angles can be distinguished. In this case, a maximal nominal tilt angle of 60° was applied, indicated with a black line revealing the missing cone. The lattice lines are visible, and an example is given in the panel. **(C)** Amplitude and phase of lattice line 1,12 revealing a z* resolution of $(7\ \text{Å})^{-1}$. The plotted curve indicates the interpolation of the lattice line. **(D)** Power spectra of an untilted and 60°-tilted 2D crystal. The inset shows the Fourier-filtered projection map from an unbent image. Perpendicular to the tilt axis (line in the 60° panel), the resolution is reduced as a result of support non-flatness and charging.

correctly centered. Errors in this step introduces errors in the phase Ψ(z*), which are usually identified during lattice line fitting (Fig. 23.7(C) and refinement. To obtain the amplitude $A(z^*)$, correct scaling of the data sets from single images or diffraction patterns is important. Scaling is optimized during refinement and lattice line fitting. Once the merging is achieved, lattice line data are fitted by a continuous function, and sampled on a raster to calculate the 3D potential map by inverse Fourier transformation. From this first 3D map that comprises the total data set, projection maps can be calculated along any projection direction for subsequent refinement runs (Fig. 23.8).

Such data processing is critical to extract all the information initially transferred by the electron microscope to film or to a CCD camera. It is a laborious process that contributes to the slowness of electron crystallography. Efforts are currently invested to improve the automation of data processing as well as the accuracy of critical algorithms involved.[32] Automation in data acquisition and processing will contribute to making electron crystallography a more widely used method.

23.5 3D Electron Microscopy of Protein Complexes

The 3D structure of large protein complexes that cannot be crystallized is assessed by recording 2D projections, and calculating the 3D structure by "weighted back-projection" methods.[33] Sample preparation, electron microscope performance and data acquisition, accuracy of projection angle determination, and the data refinement cycles dictate the resolution. All the steps should exhibit the same perfection as for electron crystallography. To eliminate the statistical noise, projections are selected, aligned, classified, and averaged. From averaged projections, the 3D map is calculated once the projection angles have been determined. A key prerequisite for this procedure is the sample homogeneity: complexes need to be all in the same specific conformational state; if this is not the case, the information from the different states will be merged by the back-projection step into a blurred

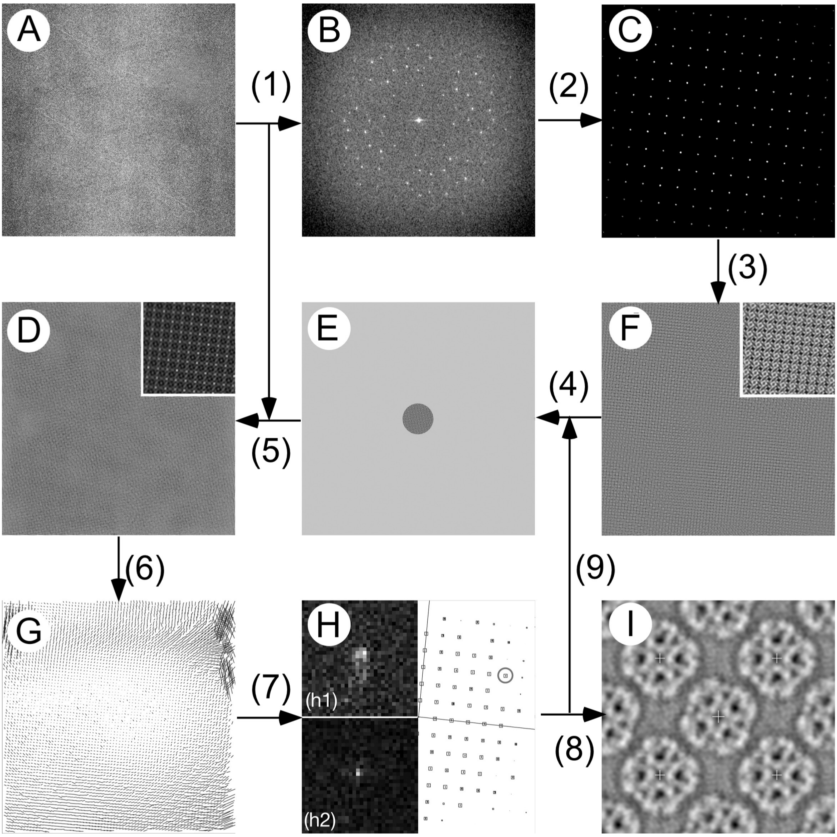


Fig. 23.8. Fourier peak-filtering and unbending of 2D crystals. The raw image **(A)** is Fourier transformed (step 1) and the crystal lattice is indexed in the power spectrum $|2\text{DFT}|^2$ of the raw image **(B)**. Note that in this case, two crystalline layers of the flattened crystalline vesicle have to be separated. For the Fourier peak-filtering, the diffraction peaks containing all the crystal information are transmitted with weight 1, while the signal outside the mask-area (containing the other crystal layer and noise) is set to 0 (step 2). **(C)** The image of the inverse Fourier transformation **(D)** reveals already the packing of the crystal (step 3). To unbend the 2D crystal, a reference **(E)** is selected from d (step 4) and a cross-correlation (step 5) with the raw image is calculated. The cross-correlation **(F)** reveals the positions of the unit cells. These can be compared to the fitted lattice (step 6) and difference vectors can be generated **(G)**. This information can be used to interpolate the raw image to unbend the crystal and to eliminate badly distorted regions (step 7). As a result, the spots of the power spectrum are focused. **(H)** In inset (h1), peak 5,3 (indicated with a circle) is depicted before unbending, and in (h2) after

3D map. Noise elimination requires processing of a large number of projections: it has recently been shown that typically some 10^4 projections will suffice to resolve a complex of about 10^6 Daltons to better than 10 Å.[34] Further, the orientation of the complexes should be random, i.e. all possible projection angles should be present. Structural preservation and random orientation are achieved best when the protein complexes are suspended in a thin solution layer that is vitrified by shock-freezing in liquid ethane.[7]

An experienced operator can efficiently select projections manually, even from an inhomogeneous preparation. However, such selection may induce a bias, and projections may be too noisy to allow discrimination by eye. In any case, manual selection is a labor-intensive bottleneck when datasets of a 10 to 100 000 projections need to be selected. Therefore, algorithm development of fully automated particle selection has been an important objective in the field. Approaches explored can broadly be classified into (i) template-matching, (ii) edge detection, (iii) intensity comparison, (iv) texture comparison, and (v) neural network-based methods. Template-matching uses the cross-correlation signal of a particle field with a reference set, which is calculated from an initial 3D model.[35] Feature-based algorithms exploit local features of a projection set that are not calculated from a model, but are derived by machine learning. A self-learning algorithm used for face recognition has been applied to particle selection,[36] and a self-learning neural network approach has been presented.[37,38] A recent evaluation of different approaches based on a common data set reveals that automated selection of asymmetric particle projections from cryo-EM images is an unresolved challenging problem.[39] The major problem is the balance between missing many true projections and picking many false positives from the background noise. If the false projections are not eliminated early in the calculation, they can severely degrade the result of the 3D map finally obtained.

unbending. Of such unbent crystal images, amplitudes and phases of the spots are read out and combined with the data of other crystals (step 8). The 3.7 Å map of GlpF (**I**) revealing the typical tetrameric structures of aquaporins demonstrates what is achievable. The map is used as a new reference for refining the data extraction (step 9).

Once the random projections are selected, they need to be aligned for classification. Since both translational and angular alignment is required, the problem arises that an angularly misaligned particle cannot be aligned translationally and vice versa. This dilemma is solved in various ways. Depending on the particle selection step, projections might already be well-centered, hence allowing the rotational alignment to be achieved directly. Another approach takes advantage of the autocorrelation function (ACF), which allows the angular alignment of particles even if they are not centered. Because the ACF often has weak features at large radii, the angular alignment can be improved on the unprocessed but centered projections in a second step. Further rotation-, translation-, and mirror-invariant functions can be derived from the input images for alignment or classification.[40] All alignment steps rely on the cross-correlation signal with a suitable reference, which in turn may introduce a bias. Reference-free alignment algorithms have thus been developed to overcome the propensity of reference-based algorithms to reinforce the reference motif in very noisy situations.[40,41]

Projections of randomly oriented particles in a thin vitrified ice layer need to be sorted into classes representing similar projection angles [Fig. 23.9(B)]. This is efficiently achieved by multivariate statistical analysis, where an image comprised of $n \times n$ (e.g. 4096) pixels, each having a value between 0 and 255, represents a vector in an $n \times n$-dimensional space, with axes extending from 0 to 255. Similar images will correspond to vectors that almost coincide, and clusters of image vectors will represent the members of a certain class. The ensemble of all images (comprising all clusters) will be distributed in a restricted volume of a particular shape in the $n \times n$ dimensional space. Multivariate statistics approaches determine the Eigenvectors and respective Eigenvalues describing this volume, which allow the dimensions of the image space to be drastically reduced.[42,43] An efficient search for delineating images clusters within the space of the most significant Eigenvectors is then provided, and class averages are calculated from projections belonging to specific clusters. Ideally, such class averages represent the averaged projection of a complex along a given projection direction.

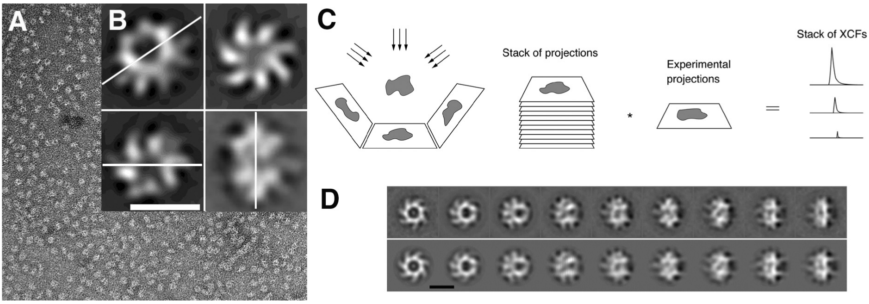


Fig. 23.9. Single-particle 3D reconstruction from random projections. **(A)** Images of negatively stained or vitrified complexes are projections from all possible directions, provided the complexes are oriented randomly. Vitrified layers of particle suspension often fulfill this prerequisite. Projections are classified by multivariate statistical methods and averaged. **(B)** To boot-strap the 3D reconstruction, a trained operator can identify the tilt axis and estimate the tilt angle from class averages. Alternatively, image pairs are taken at 0° and 45° to estimate the Euler angles from the projection pairs. A first, a 3D map is calculated by weighted back-projection. **(C)** From this map, a set of projections covering the entire range of Euler angles is calculated for use as references. Multi-reference refinement cycles then indicate the best match of calculated and measured projections, thereby improving the accuracy of the angular and translational alignment of the projections, as well as the Euler angle determination. **(D)** Refinement cycles are repeated until the class averages (top row) match the calculated projections (bottom row). Modified from Ref. 67.

Conceptually, the determination of the 3D potential map $\rho(x,y,z)$ of a biological molecule from a large number of 2-D electron microscopy projections of isolated (single) particles with random and unknown orientations may be considered as a nonlinear optimization problem, which seeks to determine the minimum of

$$\sum_{i=1}^{m} \left\| T(x_i, y_i) P(\phi_i, \theta_i, \psi_i) \rho(x, y, z) - b_i(x, y) \right\|^2 \tag{23.7}$$

where $P(\phi_i,\ \theta_i,\ \psi_i)$ is a line integral operator that projects $\rho(x,y,z)$ onto a 2-D plane after $\rho(x,y,z)$ is rotated by a set of unknown Euler angles ϕ_i, θ_i, and ψ_i. $T(x_i,\ y_i)$ is a translational operator, and $b_i(x,y)$ is the experimental projection map. To boot-strap the determination of the first 3D map, the random conical tilt approach can be most useful.[44] Here, projections of a particle field are acquired at 0° tilt and at 45° tilt. The Euler angles ϕ_i, θ_i, and ψ_i are determined from the projection pair of each particle, and the back-projection operation can be executed. Alternatively, projection angles may be estimated roughly by visual inspection of class averages [Fig. 23.9(B)], or quantitatively by the sinogram correlation function.[33,45] The resulting first map $\rho_0(x,y,z)$ provides the platform to refine the structure, because it allows multiple references to be calculated and projections of randomly oriented particles to be better aligned. To this end, all projections are cross-correlated with all references to determine the best match[33,46] [Fig. 23.9(C)]. Correlation coefficients allow the Euler angles of the respective projections to be refined, the projections to be centered more accurately, and atypical projections to be eliminated. Back-projection with refined parameters produces a new map, $\rho_1(x,y,z)$, whose resolution is assessed by the Fourier Shell Correlation (FSC) technique,[47] to appropriately low-pass filter the map for the next refinement cycle. The quality of both, the first 3D map, and the experimental data dictates the convergence of the refinement. Multi-reference alignment procedures are computationally intensive and key in refining the structure.[33,46]

One aspect to be addressed concerns sample heterogeneity. First, it is often not easy to produce a chemically pure sample, and the recognition and subsequent elimination of atypical projections is therefore important. More important is the possibility to study conformational heterogeneity, which is a great advantage compared to crystallographic methods, where proteins are necessarily in a single conformation. However, the problem to solve is to properly sort out all the projections into sets belonging to specific conformations. Considering the low signal-to-noise ratio of projections from shock-frozen preparations recorded at low dose, and the large number of projections to be processed, this task is enormous and still not completely tackled. Nevertheless, studies of ribosomal conformations have paved ways to approach this problem,[48–50] and new image processing methods have now emerged.[51]

23.6 Electron Tomography

Since electron micrographs represent two-dimension (2D) projections of the object's 3D potential distribution, features from different levels within the object are superimposed and cannot be separated. Therefore, tomographic techniques acquire projections of the object viewed from different directions and then merge them computationally to obtain a 3D reconstruction of the object volume. In the electron microscope, the specimen holder is rotated incrementally around an axis perpendicular to the electron beam, and images are taken at each position, to use them for calculating the 3D map by back-projection. About 100 projections need to be acquired, and the critical dose dictated by the beam damage is 1000–2000 electrons/nm^2. Therefore, the dose per projection is typically 10–20 electrons/nm^2. As projections of a single object are recorded by rotating it in small angular increments, the alignment of individual projections would ideally not be necessary. However, this would necessitate a perfect eucentric goniometer stage that does not displace the sample from the center to warrant alignment of projections and keep the object in focus. Since such accuracy is technically not possible, procedures have been developed to measure the displacement of the sample on an adjacent area and to correct its effects automatically before recording the next projection. Nevertheless, a final centering of all projections by cross-correlation is required before the weighted back-projection can be executed such as for single-particle reconstruction (Fig. 23.9). The number n of projections taken dictates the resolution d according to Equation 23.8:

$$d \approx \frac{\pi T}{n} \tag{23.8}$$

Accordingly, the resolution to be achieved from a slab of thickness $T = 200$ nm given a series of 100 projections, is 6 nm. Several examples document that 3 nm can be reached, mainly by recording more projections. At this resolution, it is possible to identify large complexes by pattern recognition, and thus, to perform visual proteomics of single cells.[5,52]

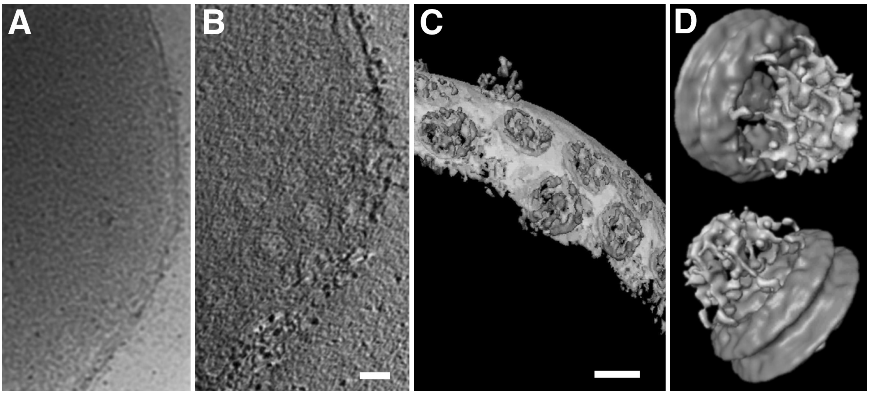


Fig. 23.10. Tomogram of a nucleus from *Dictyostelium discoideum* and 3D map of the native nuclear pore complex. **(A)** A single projection of a nucleus does not exhibit much detail. From many projections (about 100) taken at different angles, a tomogram is calculated. **(B)** A section through the tomogram reveals significant structural features; nuclear pore complexes can be discerned. **(C)** The resulting 3D map after denoising and surface rendering shows six complexes in different orientations. **(D)** Collecting the 3D maps of a 267 complexes, aligning and averaging them emphasizes the non-variable features of the native nuclear pore complex.[54] (By courtesy of Wolfgang Baumeister.)

As the apparent specimen thickness increases at higher tilt angles and the tilt range of cold stages is limited, the practical accessible tilt range is restricted to 70°. Therefore, data are missing from a wedge-shaped region of the 3DFT, resulting in a non-isotropic resolution in the reconstructed volume. Dual axis tilting provides a partial solution of this problem.[53] Multiple inelastic scattering, which degrades the image quality as result of chromatic aberration is another limiting factor for specimens that are significantly thicker than the mean free path for inelastic scattering ($\Lambda_{inel} \approx 350$ nm for 300 kV electrons). Energy filters operated in a zero-loss mode remove the blurring contributions of the inelastically scattered electrons and the sharpness of the image is improved. Since multiple elastically scattered electrons are scattered beyond the objective aperture and inelastic electrons are eliminated by energy filtering, the useful signal decreases according to the total mean free path Λ_{tot}, which is about 200 nm at 300 kV. As a rule of thumb, good quality tomograms can be obtained with samples that have a thickness smaller than 3 Λ_{tot}, but in practice, tomograms of amazing quality have been obtained with samples up to 1 μm.[54] Larger structures, however, must be sectioned for the tomographic analysis, preferably using cryo-sectioning.[10]

23.7 Future Outlook

Modern electron microscopy tools are now available to visualize the proteome of a single cell in its entirety. Although eukaryotic cells may be too large to be imaged as a whole, methods are being developed that allow large cells or even tissues to be vitrified by high-pressure freezing and sectioned at low temperature,[9] producing samples that are suitable for electron tomography.[10] The resolution of the latter technique has now reached a level, where protein complexes can be recognized by pattern-recognition algorithms,[55] opening an avenue for visual proteomics.[52] Recent results indicate that cells are crowded with protein complexes, many of which are likely to disassemble, and hence, are not accessible to standard biochemical analyses.[5] 3D electron microscopy appears to have become the tool of choice to visualize the complexity of cells in the context of future developments in systems biology.

More stable complexes, however, can be inspected by single-particle methods at far better resolution, allowing the structure and dynamics of molecular machines to be studied at a resolution that allows atomic structures to be fitted.[18,33,56–58] While this is certainly attractive for large complexes that resist all attempts to force them into highly ordered 3D crystals, the possibility to visualize the function-related conformational changes are even more important.[59–61] Current efforts to tackle the formidable task to sort out projections of conformational heterogeneity foster the hope that electron microscopy will soon be used routinely to visualize the dynamics of biological molecular machines.[51] The prospect to image such dynamic complexes at a resolution of around 10 Å in various conformations suggests that the fitting of atomic models to such 3D maps will allow the chemistry behind these conformational changes to be unraveled.

Tremendous progress in the study of membrane protein structure by electron crystallography demonstrates another direction of interest. First, membrane proteins crystallized by reconstitution in the presence of lipids are embedded in their native environment, and are directly accessible to functional assays.[21] Second, the surfaces of proteins in such crystals are available to ligands, making the analysis of corresponding complexes possible, likely without the need for co-crystallization. Third, 2D crystallization methods have improved, with resolutions beyond 2 Å demonstrated.[6] Although only a few groups have pushed the related methods of sample preparation methods, data acquisition, and image processing,[3] impressive recent results[6,30,62–65] promise that electron crystallography will become more routinely used in membrane protein structure research. Combining this progress with that in single-cell imaging and single-particle analyses, the goal of modeling and understanding a cell at the atomic scale seems to be getting closer.

Acknowledgement

This work has been supported by the Maurice E. Müller Foundation of Switzerland, by the Swiss National Foundation (SNF) within the framework of the National Center of Competence in Research for Structural Biology, the SNF grant 3100-059415 to AE, the NoE 3D-EM (EU project LSHG-CT-2004-502828), and the HT-3DEM (EU project LSHG-CT-2005-018811).

References

1. Dubochet J, Adrian M, Chang JJ, *et al.* (1988) Cryo-electron microscopy of vitrified specimens. *Q Rev Biophys* **21**: 129–228.
2. Al-Amoudi A, Chang JJ, Leforestier A, *et al.* (2004) Cryo-electron microscopy of vitreous sections. *EMBO J* **23**: 3583–3588.
3. Fujiyoshi Y. (1998) The structural study of membrane proteins by electron crystallography. *Adv Biophys* **35**: 25–80.
4. Koster AJ, Grimm R, Typke D, *et al.* (1997) Perspectives of molecular and cellular electron tomography. *J Struct Biol* **120**: 276–308.
5. Baumeister W. (2005) A voyage to the inner space of cells. *Protein Sci* **14**: 257–269.

6. Gonen T, Cheng Y, Sliz P, *et al.* (2005) Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. *Nature* **438**: 633–638.
7. Adrian M, Dubochet J, Lepault J, McDowall AW. (1984) Cryo-electron microscopy of viruses. *Nature* **308**: 32–36.
8. Al-Amoudi A, Dubochet J, Norlen L. (2005) Nanostructure of the epidermal extracellular space as observed by cryo-electron microscopy of vitreous sections of human skin. *J Invest Dermatol* **124**: 764–777.
9. Al-Amoudi A, Norlen LP, Dubochet J. (2004) Cryo-electron microscopy of vitreous sections of native biological cells and tissues. *J Struct Biol* **148**: 131–135.
10. Garvalov BK, Zuber B, Bouchet-Marquis C, *et al.* (2006) Luminal particles within cellular microtubules. *J Cell Biol* **174**: 759–765.
11. Malac M, Beleggia M, Egerton R, Zhu Y. (2007) Imaging of radiation-sensitive samples in transmission electron microscopes equipped with Zernike phase plates. *Ultramicroscopy* **107**: 40–49.
12. Gyobu N, Tani K, Hiroaki Y, *et al.* (2004) Improved specimen preparation for cryo-electron microscopy using a symmetric carbon sandwich technique. *J Struct Biol* **146**: 325–333.
13. Philippsen A, Engel HA, Engel A. (2007) The contrast-imaging function for tilted specimens. *Ultramicroscopy* **107**: 202–212.
14. Müller SA, Engel A. (2006) Biological scanning transmission electron microscopy: imaging and single molecule mass determination. *CHIMIA* **60**: 749–753.
15. Takamori S, Holt M, Stenius K, *et al.* (2006) Molecular anatomy of a trafficking organelle. *Cell* **127**: 831–846.
16. Dube P, Herzog F, Gieffers C, *et al.* (2005) Localization of the coactivator Cdh1 and the cullin subunit Apc2 in a cryo-electron microscopy model of vertebrate APC/C. *Mol Cell* **20**: 867–879.
17. Chami M, Guilvout I, Gregorini M, *et al.* (2005) Structural insights into the secretin PulD and its trypsin-resistant core. *J Biol Chem* **280**: 37732–37741.
18. Eifler N, Vetsch M, Gregorini M, *et al.* (2006) Cytotoxin ClyA from *Escherichia coli* assembles to a 13-meric pore independent of its redox-state. *EMBO J* **25**: 2652–2661.
19. Müller SA, Pozidis C, Stone R, *et al.* (2006) Double hexameric ring assembly of the type III protein translocase ATPase HrcN. *Mol Microbiol* **61**: 119–125.
20. Mueller CA, Broz P, Müller SA, *et al.* (2005) The V-antigen of *Yersinia* forms a distinct structure at the tip of injectisome needles. *Science* **310**: 674–676.
21. Walz T, Smith BL, Zeidel ML, Engel A, Agre P. (1994) Biologically active two-dimensional crystals of aquaporin CHIP. *J Biol Chem* **269**: 1583–1586.
22. Braun T, Engel A. (2005) Two-dimensional electron crystallography. *Nat Encyc Life Sci* A0003044, **in press**.
23. Mannella CA. (1984) Phospholipase-induced crystallization of channels in mitochondrial outer membranes. *Science* **224**: 165–166.

24. Unger VM, Kumar NM, Gilula NB, Yeager M. (1999) Expression, two-dimensional crystallization, and electron cryo-crystallography of recombinant gap junction membrane channels. *J Struct Biol* **128**: 98–105.
25. Jap BK, Zulauf M, Scheybani T, *et al.* (1992) 2D crystallization: from art to science. *Ultramicroscopy* **46**: 45–84.
26. Engel A, Hoenger A, Hefti A, *et al.* (1992) Assembly of 2-D membrane protein crystals — dynamics, crystal order, and fidelity of structure analysis by electron microscopy. *J Struct Biol* **109**: 219–234.
27. Levy D, Chami M, Rigaud JL. (2001) Two-dimensional crystallization of membrane proteins: the lipid layer strategy. *FEBS Lett* **504**: 187–193.
28. Rigaud JL, Mosser G, Lacapere JJ, *et al.* (1997) Bio-Beads: an efficient strategy for two-dimensional crystallization of membrane proteins. *J Struct Biol* **118**: 226–235.
29. Remigy HW, Caujolle-Bert D, Suda K, *et al.* (2003) Membrane protein reconstitution and crystallization by controlled dilution. *FEBS Lett* **555**: 160–169.
30. Hiroaki Y, Tani K, Kamegawa A, *et al.* (2006) Implications of the aquaporin-4 structure on array formation and cell adhesion. *J Mol Biol* **355**: 628–639.
31. Henderson R, Baldwin JM, Ceska TA, *et al.* (1990) Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J Mol Biol* **213**: 899–929.
32. Philippsen A, Schenk AD, Signorell GA, *et al.* (2007) Collaborative EM image processing with the IPLT image processing library and toolbox. *J Struct Biol* **157**: 28–37.
33. van Heel M, Gowen B, Matadeen R, *et al.* (2000) Single-particle electron cryo-microscopy: towards atomic resolution. *Q Rev Biophys* **33**: 307–369.
34. Stagg SM, Lander GC, Pulokas J, *et al.* (2006) Automated cryoEM data acquisition and analysis of 284742 particles of GroEL. *J Struct Biol* **155**: 470–481.
35. Rath BK, Frank J. (2004) Fast automatic particle picking from cryo-electron micrographs using a locally normalized cross-correlation function: a case study. *J Struct Biol* **145**: 84–90.
36. Mallick SP, Zhu Y, Kriegman D. (2004) Detecting particles in cryo-EM micrographs using learned features. *J Struct Biol* **145**: 52–62.
37. Ogura T, Sato C. (2004) Automatic particle pickup method using a neural network has high accuracy by applying an initial weight derived from eigenimages: a new reference free method for single-particle analysis. *J Struct Biol* **145**: 63–75.
38. Ogura T, Sato C. (2001) An automatic particle pickup method using a neural network applicable to low-contrast electron micrographs. *J Struct Biol* **136**: 227–238.
39. Zhu Y, Carragher B, Glaeser RM, *et al.* (2004) Automatic particle selection: results of a comparative study. *J Struct Biol* **145**: 3–14.

40. Schatz M, van Heel M. (1990) Invariant classification of molecular views in electron micrographs. *Ultramicroscopy* **32**: 255–264.
41. Penczek P, Radermacher M, Frank J. (1992) Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy* **40**: 33–53.
42. Frank J, van Heel M. (1982) Correspondence analysis of aligned images of biological particles. *J Mol Biol* **161**: 134–137.
43. van Heel M, Frank J. (1981) Use of multivariate statistics in analysing the images of biological macromolecules. *Ultramicroscopy* **6**: 187–194.
44. Radermacher M, Wagenknecht T, Verschoor A, Frank J. (1986) A new 3-D reconstruction scheme applied to the 50S ribosomal subunit of *E. coli*. *J Microsc* **141**: RP1–2.
45. Van Heel M. (1987) Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **21**: 111–123.
46. Yang C, Penczek PA, Leith A, *et al.* (2007) The parallelization of SPIDER on distributed-memory computers using MPI. *J Struct Biol* **157**: 240–249.
47. Saxton WO, Baumeister W. (1982) The correlation averaging of a regularly arranged bacterial cell envelope protein. *J Microsc* **127**: 127–138.
48. Diaconu M, Kothe U, Schlunzen F, *et al.* (2005) Structural basis for the function of the ribosomal L7/12 stalk in factor binding and GTPase activation. *Cell* **121**: 991–1004.
49. Penczek PA, Frank J, Spahn CM. (2006) A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. *J Struct Biol* **154**: 184–194.
50. Taylor DJ, Nilsson J, Merrill AR, *et al.* (2007) Structures of modified eEF2 80S ribosome complexes reveal the role of GTP hydrolysis in translocation. *EMBO J* **26**: 2421–2431.
51. Scheres SH, Gao H, Valle M, *et al.* (2007) Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nat Meth* **4**: 27–29.
52. Nickell S, Kofler C, Leis AP, Baumeister W. (2006) A visual approach to proteomics. *Nat Rev Mol Cell Biol* **7**: 225–230.
53. Nickell S, Hegerl R, Baumeister W, Rachel R. (2003) Pyrodictium cannulae enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *J Struct Biol* **141**: 34–42.
54. Beck M, Forster F, Ecke M, *et al.* (2004) Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. *Science* **306**: 1387–1390.
55. Frangakis AS, Bohm J, Forster F, *et al.* (2002) Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proc Natl Acad Sci USA* **99**: 14153–14158.
56. Gao H, Frank J. (2005) Molding atomic structures into intermediate-resolution cryo-EM density maps of ribosomal complexes using real-space refinement. *Structure* **13**: 401–406.

57. Mitra K, Schaffitzel C, Shaikh T, *et al.* (2005) Structure of the *E. coli* protein-conducting channel bound to a translating ribosome. *Nature* **438**: 318–324.
58. Nitsch M, Walz J, Typke D, *et al.* (1998) Group II chaperonin in an open conformation examined by electron tomography. *Nat Struct Biol* **5**: 855–857.
59. Golas MM, Sander B, Will CL, Luhrmann R, Stark H. (2005) Major conformational change in the complex SF3b upon integration into the spliceosomal U11/U12 di-snRNP as revealed by electron cryomicroscopy. *Mol Cell* **17**: 869–883.
60. Golas MM, Sander B, Will CL, Luhrmann R, Stark H. (2003) Molecular architecture of the multiprotein splicing factor SF3b. *Science* **300**: 980–984.
61. Stark H, Dube P, Luhrmann R, Kastner B. (2001) Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle. *Nature* **409**: 539–542.
62. Oshima A, Tani K, Hiroaki Y, Fujiyoshi Y, Sosinsky GE. (2007) Three-dimensional structure of a human connexin26 gap junction channel reveals a plug in the vestibule. *Proc Natl Acad Sci U S A* **104**: 10034–10039.
63. Kukulski W, Schenk AD, Johanson U, *et al.* (2005) The 5A structure of heterologously expressed plant aquaporin SoPIP2;1. *J Mol Biol* **350**: 611–616.
64. Schenk AD, Werten PJ, Scheuring S, *et al.* (2005) The 4.5 A structure of human AQP2. *J Mol Biol* **350**: 278–289.
65. Murata K, Mitsuoka K, Hirai T, *et al.* (2000) Structural determinants of water permeation through aquaporin-1. *Nature* **407**: 599–605.
66. Gregorini M, Wang J, Xie X, RA M, Engel A. (2007) Three-dimensional reconstruction of bovine brain V-ATPase by cryo-electron microscopy and single particle analysis. *J Struct Biol.*
67. Stasiak AZ, Larquet E, Stasiak A, *et al.* (2000) The human Rad52 protein exists as a heptameric ring. *Curr Biol* **10**: 337–340.

Chapter 24

New Frontiers in Characterizing Structure and Dynamics by NMR

M. Nilges*,†, P. Markwick†, T. Malliavin†, W. Rieping‡ and M. Habeck§

24.1 Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy has emerged as the method of choice for studying both the structure and dynamics of biological macromolecules in solution. Despite the maturity of the NMR method for structure determination, its application faces a number of challenges. The method is limited to systems of relatively small molecular mass, data collection times are long, data analysis remains a lengthy procedure, and it is difficult to evaluate the quality of the final structures. The last years have seen significant advances in experimental and analysis techniques to overcome or reduce some limitations.

The function of bio-macromolecules is determined by both their 3D structure and conformational dynamics. The molecules are

*Corresponding author.

†Unité de Bio-Informatique Structurale, Institut Pasteur, and CNRS URA 2185, 25-28 rue du docteur Roux, F-75015 Paris, France. Email: nilges@pasteur.fr

‡Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge, CB2 1GA, UK.

§Max-Planck-Institute for Developmental Biology, Spemannstr. 35 and Max-Planck-Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany.

inherently flexible systems displaying a broad range of dynamics on timescales from picoseconds to seconds. NMR is unique in its ability to obtain dynamic information on an atomic scale. The experimental information on structure and dynamics is intricately mixed. It is, however, difficult to unite both structural and dynamical information into one consistent model, and protocols for the determination of structure and dynamics are performed independently.

This chapter deals with the challenges posed by the interpretation of NMR data on structure and dynamics. We will first relate the standard structure calculation methods to Bayesian probability theory. We will then briefly describe the advantages of a fully Bayesian treatment of structure calculation. Then, we will illustrate the advantages of using Bayesian reasoning at least partly in standard structure calculations. The final part will be devoted to the interpretation of experimental data on dynamics.

24.2 Determination of Structure

The principal experimental information for structure elucidation by NMR comes from short inter-atomic distances obtained from nuclear Overhauser effects (NOEs) between proton spins.[1] Long-range distance information can be acquired from paramagnetic relaxation.[2] These distances are supplemented by information on torsion angles from through-bond scalar couplings and chemical shifts. Residual Dipolar Couplings (RDCs) provide powerful additional orientational restraints,[3,4] which, in contrast to the other experimental information, do not characterize local relative positions, but the orientation of inter-atomic vectors with respect to a reference frame. The chemical shift depends on the local structure,[5] and, together with data base searches, it can be sufficient to determine the 3D structure.[6] Whereas structure determination from NMR in general relies on data from liquid samples, the feasibility of protein structure determination by solid state NMR has recently been demonstrated.[7] This has potential applications for molecules that are neither soluble nor form three-dimensional crystals easily, such as membrane proteins.

24.2.1 *The Hybrid Energy Function and Bayes's Rule*

Experimental data are rarely sufficient to determine the three-dimensional structure of a macro-molecule by themselves, but need to be complemented with prior physical information. Therefore, structure calculation is typically a search for conformations that simultaneously have a low physical energy $E_{phys}(X)$ and minimize a cost function, $E_{data}(X)$, quantifying the disagreement between a structural model X and the data. This approach was first implemented for X-ray crystal structure determination as the minimization of hybrid energy[8,9]

$$E_{hybrid}(X) = E_{phys}(X) + w_{data}E_{data}(X) \tag{24.1}$$

The weight E_{data} controls the contribution of the data relative to the physical energy (for example, a molecular dynamics force field adapted to the structure determination task). Jack and Levitt remarked in their paper that correct weighting of the data "is something of a problem."[8] Its value can be critical: If it is too large, the contribution of the force field might be too small (over fitting of the data); if the weight is too small, the data may contribute too little to define the structure (under fitting of the data).

Calculation methods to minimize the hybrid energy function have been amply reviewed[9,10] and will not be our concern here. Structure calculation is an example of the fitting of parameters (principally, the atomic coordinates) to experimental data. "*To be genuinely useful, a fitting procedure should provide* (i) *parameters,* (ii) *error estimates of the parameters, and* (iii) *a statistical measure of a goodness of fit.*"[11] Whereas the standard procedures applied to NMR structure calculation provide the "parameters," they do not supply meaningful error estimates, and due to the usual procedure to convert the experimental data into loose bounds,[1] they do not provide a statistically meaningful measure of goodness of fit.

Minimizing the hybrid energy function is motivated by maximizing the posterior probability of a structure when prior information and experimental data are available. If we restrict the analysis to the molecular co-ordinates X, Bayes's theorem[12] provides the posterior probability distribution for the unknown coordinates:

$$p(X \mid D, I) \propto \pi(X \mid I) L(D \mid X, I) \tag{24.2}$$

The posterior $p(X|D, I)$ factorizes into two natural components (Fig. 24.1):

1. The prior $\pi(X|I)$ describes knowledge about general properties of biomolecular structures. At temperature T, the Boltzmann distribution $\pi(X) \propto \exp(-\beta E_{phys}(X))$ is the least biasing prior distribution.[12] β is $(k_B T)^{-1}$, with the Boltzmann constant k_B.

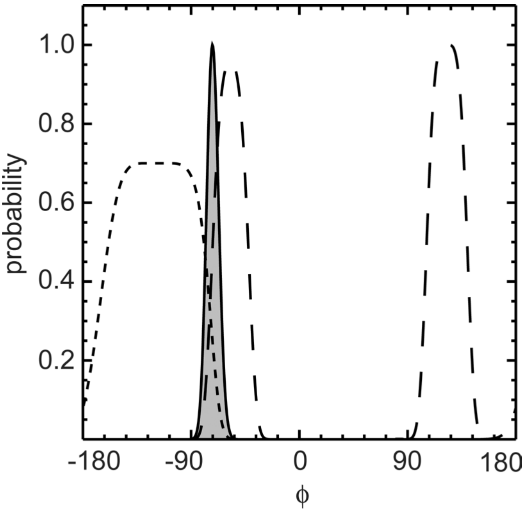


Fig. 24.1. Prior knowledge is incorporated in a natural way using the laws of probability theory. In the illustrated case, the prior knowledge (dotted line) is the probability to observe a particular torsion angle before any data is measured (for example, we know that the protein backbone torsion angle, ϕ, is in most cases negative). The likelihood (dashed line) adds the knowledge obtained from the data: In our case, there are two peaks in the likelihood. The posterior probability (solid line) is obtained by multiplication of the prior probability and the likelihood, and represents the total knowledge we have about the conformation.

2. In the second term on the right-hand side, $L(D|X, I)$, the likelihood is the probability of the data D, given the molecular structure X. For its evaluation, we need a theory allowing us to calculate the data from a structure, and an error model accounting for deviations between calculated and measured data.

Bayes's rule combines the two components (the prior and the likelihood; quantities we can calculate) to derive the probability of a particular structure X (the quantity we are interested in).

For the error model, one usually assumes that the experimental data are distributed around the average value in the same way they are distributed around the value predicted from the theory, i.e. that the model does not introduce a systematic bias. This does not imply that the average experimental value and the predicted value are identical, only that the distributions are similar.

For example, the data may follow the distribution:

$$L(D \mid X, I) \propto \exp\left\{-\frac{1}{2\sigma^2}\chi^2(X)\right\} \tag{24.3}$$

where the function $\chi^2(X)$ quantitates the average discrepancy between the experimental measurements y_i and the data predicted from the structure X by the theory, $y_i(X)$. For a Gaussian distribution, this is:

$$\chi^2(X) = \sum_{i=1}^{n}[y_i - y_i(X)]^2 \tag{24.4}$$

σ is the mean deviation of the measurements from the theoretical value. This is an important parameter we generally cannot measure but need to introduce for the modeling.

If we take the negative logarithm of both sides of Equation (24.2), we obtain an equation of the form of Equation (24.1):

$$-\log[p(X \mid D, I)] = -\log[\pi(X \mid I)] - \log[L(D \mid X, I)] + const \tag{24.5}$$

We can identify the hybrid energy function $E_{hybrid}(X)$ with $-\beta\log[p(X|D, I)] - const$, and the pseudo energy $E_{data}(X)$ with the negative logarithm of the likelihood, $-\log[L(D|X, I)]$. If we assume a known and constant σ, we obtain with the likelihood in Equation 24.3:

$$E_{hybrid}(X) = \beta E_{phys}(X) + \frac{1}{2\sigma^2}\chi^2(X) \tag{24.6}$$

where the factor β defines the energy scale; it is 1 if we measure the energies in units of k_BT.

As an additional conclusion, we gain from this analysis that, if we know σ, it determines the weight w_{data}. It should ideally be set to reflect the quality (consistency) of the data: the larger σ is, the lower the weight. For now, we assumed that we know σ (and all parameters necessary to formulate the likelihood $L(D|X, I)$). However, this is not necessary in a truly probabilistic analysis. The assumption of the prior knowledge of σ is unrealistic, since σ varies from experiment to experiment, and is, for NMR, dominated by discrepancies between theory and experiment and not by experimental noise.

24.2.2 *Obtaining Coordinates and Their Precision*

In structure determination by NMR, one usually tries to obtain a measure of coordinate uncertainty by repeated, independent minimizations of Equation (24.1). Estimating uncertainties in coordinates in this way has been a pre-occupation in NMR structure determination since its beginning.[1,13] The results differ from calculation to calculation with identical data, since all usually employed minimization approaches contain a random element, and since they are unable to locate "the" global minimum of the rugged energy surface but get invariably trapped in a local minimum. The resulting structure ensemble can be characterized by its average structure and its distribution around the average — in the simplest case, the RMS deviation.

This distribution of structures is influenced by the data quality. Most structure calculations employ lower and upper bounds with error tolerances that should be set according to σ: the wider the bounds, the larger the difference between individual structures.[13] If the weight, w_{data}, depends on the standard deviation σ of the data, the influence of the data is reduced for low quality data, and the results of repeated structure calculations show a larger variation.

There are two fundamental problems with this approach. First, optimization algorithms are neither guaranteed to find all important regions of the distribution nor to reproduce the correct populations of the different regions. The "sampling" provided by optimization methods, starting from randomly varying initial points, will mostly depend on algorithmic properties. Second, many of the parameters that are necessary for calculating structures (such as the weight w_{data} in Equation (24.1) need to be fixed before the calculation, and the influence of their value and variation on the co-ordinate precision cannot be assessed.

In order to rigorously address the problem of obtaining unbiased co-ordinate precision with full dependency on all unknowns, we developed the Inferential Structure Determination approach (ISD).[14] This abandons the idea of minimizing a hybrid energy or maximizing the probability. Rather, it aims at evaluating a probability P_i for "all" possible structures X_i. Generally, a continuum of P_i values is distributed over conformational space (Cartesian coordinates, dihedral angles), and P_i is a density $p(X|D, I)$; the integral $\int_R dXp(X|D, I)$ evaluates the probability that region R of conformational space contains the true structure.

Once the model to describe the data (i.e. the likelihood function) has been chosen, the rules of probability theory, Equation (24.2) or (24.7), uniquely determine the posterior distribution. The appropriate statistics for modeling distances and NOEs are discussed further below. No additional assumptions need to be made.

24.2.3 *Treatment of Additional Parameters*

The full power of the Bayesian treatment of the problem becomes apparent if there are additional unknown, auxiliary parameters (called "nuisance parameters"). It is basically always necessary to introduce such auxiliary parameters in order to describe the problem adequately.

For example, the parameters A, B, C of the Karplus relationship are unknown for the particular protein that one is investigating. Also, the data quality σ is an unknown parameter, as is the calibration factor γ for NOE volumes.

In Bayesian theory, these additional parameters are called "nuisance parameters." In ISD, all additional unknown parameters of the error model and the theory are estimated along with the structure. They are treated in the same way as the co-ordinates. For the unknown σ, for example, we replace X with (X, σ) in Equation (24.2), and the full posterior becomes

$$p(X,\sigma \mid D,I) \propto \pi(X \mid I)\pi(\sigma \mid I)L(D \mid X,\sigma,I). \tag{24.7}$$

Here, we assume *a priori* independence of X and the nuisance parameters — the prior for the coordinates does not depend on the values of the σ and *vice versa* — and we introduce the additional prior $\pi(\sigma|I)$ (Jeffreys prior[15]), expressing our ignorance on this parameter. Other nuisance parameters (the calibration factor γ, Karplus parameters, tensor parameters, etc.) are treated in exactly the same way and simply lead to additional terms in the equation.

The posterior density for the coordinates by themselves is formally obtained by integration over nuisance parameters (also called marginalization[12]). In fact, in order to account for our ignorance regarding the nuisance parameter σ, we have to replace $L(D|X, I)$ in Equation (24.2) with a *weighted average* over the likelihood conditioned on all possible values of σ. This is in marked difference to standard structure determination by minimization, where the value of any unknown parameter needs to be fixed before the structure calculation, and therefore, only one single value is used. In contrast, the result of a structure calculation by inference directly contains the influence of the uncertainty in the additional parameters.

24.2.4 *Sampling the Posterior Probability Distribution*

For a single — or very few — unknowns (co-ordinates and other parameters), one could calculate the probability of every conformation for example by a grid search. For the large number of unknowns typical for the structure determination of a macromolecule, this is unfeasible and the space of possible conformations has to be explored by a suitable sampling algorithm. ISD[14] is therefore based on Monte-Carlo sampling to explore the probability distribution over conformational space. Monte Carlo sampling is not used as a means to find the maximum of a probability but to evaluate the integrals over parameters that appear in the use of Bayes's rule, Equation (24.2) or (24.7).

A good sampling algorithm will produce samples with the correct probability. That is, the probability can be directly calculated from the number of times a particular region is visited. The replica-exchange Monte Carlo scheme for simulating the posterior densities[14,16] satisfies this criterion. In contrast to an optimization algorithm, it is designed to visit all regions of high probability, and not to locate efficiently one of the maxima.

The result of a Bayesian structure calculation is a large ensemble of structures sampled at many different values of the nuisance parameters, allowing for statistically meaningful, objective error bars for atom positions and nuisance parameters (Fig. 24.2). The variance of the structures automatically contains the influence of the nuisance parameters on the structures. Only a Bayesian treatment provides estimates for all unknown parameters, error estimates of the parameters, and a statistical measure of a goodness of fit.

Since the method has no free parameters that need to be fixed before the calculation, user intervention is not necessary, and structure determination becomes more objective.

24.2.5 *Data Statistics and Restraint Potentials*

The likelihood function $L(D|X, I)$ (or the related potential $E_{data}(X)$) needs to be known for any structure calculation. It has an important influence on the resulting distribution of structures. The role of $L(D|X, I)$ or $E_{data}(X)$ is to introduce our knowledge about expected deviations between measured and calculated data, and to evaluate the importance of these deviations. For certain data types, a Gaussian distribution is a good approximation, e.g. for scalar or residual dipolar

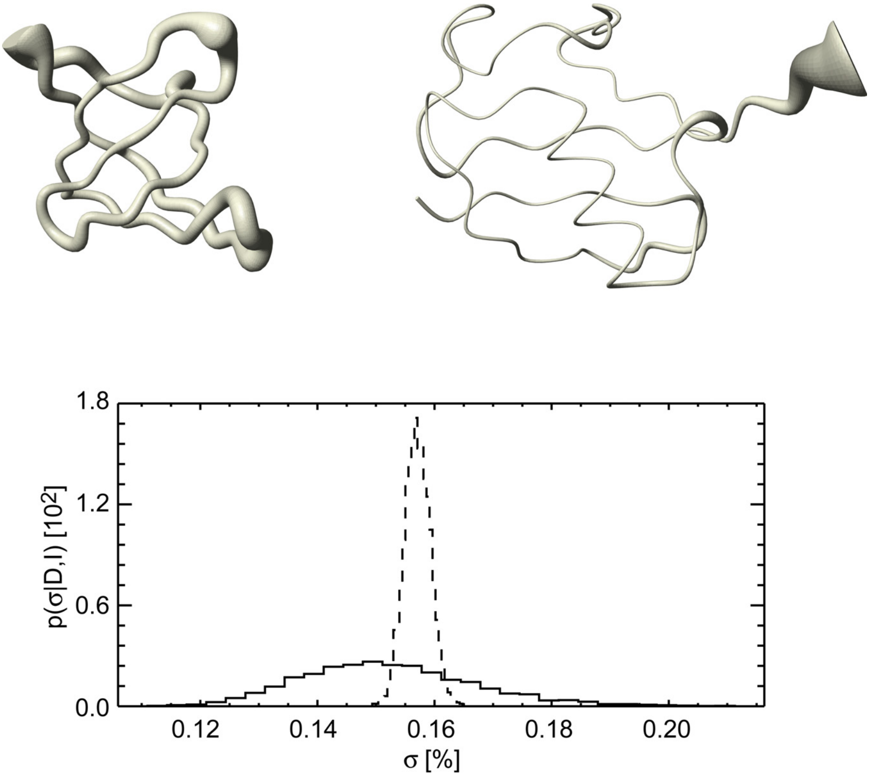


Fig. 24.2. Result of a structure calculation with ISD for two test cases, ubiquitin and an SH3 domain. The data set for ubiquitin consisted of 1444 non-redundant distances taken from the restraint file, PDB code 1d3z; the data set for SH3 was for a perdeuterated sample and contained 150 distances between exchangeable protons. Top: Ensembles of most probable structures for SH3 (left) and ubiquitin (right). The width of the "sausage" is proportional to the RMSD around the average structure. The distribution of structures contains the uncertainty due to the unknown auxiliary parameters. Bottom: Distribution for the nuisance parameter σ (error of the log-normal model) obtained from Monte Carlo samples for ubiquitin (dashed line) and SH3 (solid line). The reason for the difference in the width of the distributions is the large difference in the number of data points.

couplings. In contrast, NOEs and derived distances have too many large deviations to be well-represented by a Gaussian.

The distribution of errors of NOE-derived distances is *a priori* unknown. If we knew the error distribution $g(d, d_0)$ in the distances

d around the "true" distance d_0, we could construct a restraint potential by taking the negative logarithm of the distribution. Assuming that the individual distance measurements are statistically independent, we obtain as potential E_i^{NOE} for a single restraint i:

$$E_i^{NOE} \propto -\log[g(d_{\exp}^i, d^i(X))] \tag{24.8}$$

where $d^i(X)$ is the distance calculated in the structure X, and $d_{\exp}^i$ is the measured distance.

An appropriate error distribution can be derived from fundamental properties of NOEs and derived distances, by analyzing the expected deviation of a measurement from the ideal value. NOE intensities and derived distances are inherently positive. The simplest theory to convert NOE intensities into distances, the isolated spin pair approximation (ISPA), introduces a calibration factor γ: $I_{calc} = \gamma d^{-6}$. Changing the units does not affect the information content of the data. Hence, the distribution $g(I_{obs}, I_{calc})$ of the deviations between observed and calculated intensities must be invariant under scaling, i.e. $g(I_{obs}, I_{calc}) = \alpha g(\alpha I_{obs}, \alpha I_{calc})$, which follows from the transformation rule of probability densities. A distribution that shows this scale invariance is the lognormal distribution[17]:

$$g(I_{obs}, I_{calc}) = \frac{1}{\sqrt{2\pi\sigma^2}\, I_{obs}} \exp\left[-\frac{1}{2\sigma^2}\log^2\left(\frac{I_{obs}}{I_{calc}}\right)\right] \tag{24.9}$$

This distribution is restricted to the positive axis and is asymmetric around its median I_{calc}. Measurements are incorporated without bias in the sense that the probability of over- or underestimating the true intensity is both half. This is not the case for error distributions defined on the entire axis, such as a Gaussian, which assign a non-vanishing probability to unobservable negative intensities. The parameter σ quantifies the relative deviation of the observed from the calculated intensity, provided that their difference is sufficiently small. Experimental NOE data follow this distribution quite well,[17] indicating that the validity of the assumption that the shapes of the distributions around

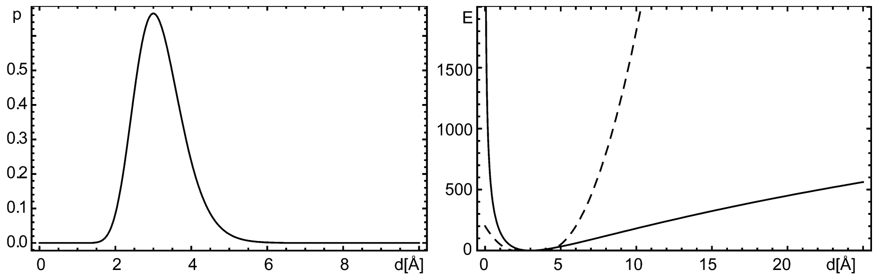


Fig. 24.3. Left: example of a log-normal distribution, with $\sigma = 0.2$ and an average value of 3.0 Å. Right, solid line: potential derived from the lognormal distribution, with a weight factor corresponding to Equation (24.12); dashed line, standard potential with lower and upper bounds at ±1 Å of the target value of 3.0 Å, and a weight factor typically used in structure calculations.

the mean and around the value calculated by the theory are indeed similar, even for the simple ISPA approximation.

Figure 24.3 shows an example of the log-normal distribution and the derived potential for a target distance of 3 Å and σ of 0.5. The distribution is asymmetric and long-tailed; both properties are much better accounted for by the lognormal distribution than by a Gaussian distribution, which would significantly underestimate the probability of large deviations. The lognormal model has other favorable properties. Unlike a probability distribution corresponding to a flat-bottom potential, it has a unique maximum. Hence, measurements are not weighted equally between bounds but are always penalized depending on the degree of disagreement with the structure. Furthermore, the lognormal distribution is invariant under power law transformations. If we raise the intensity to a power, the transformed intensity still follows a lognormal law with transformed median and error parameters.

The negative logarithm of the distribution in Equation (24.9) is the corresponding restraint potential:

$$E_i^{NOE} \propto -\log[g(I_{obs}, I_{calc})] = \frac{1}{2\sigma^2}\log^2\left(\frac{I_{obs}}{I_{calc}}\right) \tag{24.10}$$

which is harmonic in the difference of the logarithm of calculated and experimental intensities. Note that this "log-harmonic" potential has only one single parameter, the weight depending on σ.

24.2.6 *Data Quality and the Weight on $E_{data}(X)$*

As already mentioned above, the weight plays a fundamental role in calculating structures from experimental data. Within ISD, the weight is estimated along with all other unknown parameters (see above). In a standard structure calculation by minimization, the experimental data are weighted empirically: w_{data} is set *ad hoc* and held constant during structure calculation.

An unbiased empirical method to determine the optimal weight in NMR is complete cross-validation, see Ref. 18 for a recent application. Probability theory gives us another possibility to weight experimental data in an objective way, as in the ISD approach[14] described above.

Probabilistic analysis can also be used to derive an optimal weight for a minimization approach[19]: By taking the negative logarithm of Equation (24.7), the full posterior probability, including the nuisance parameter σ, becomes:

$$E_{jo\,\mathrm{int}}(X,\sigma) = \frac{1}{2\sigma^2}\chi^2(X) + \beta E_{phys}(X) + \log\left[\frac{Z(\sigma)}{\pi(\sigma)}\right] \qquad (24.11)$$

where, for the lognormal model, $\chi^2(X) = \sum_{i=1}^{n} \log^2[y_i / y_i(X)]$.

The last term on the right-hand side of the joint hybrid energy, Equation (24.11), is not included in the standard target function E_{hybrid}, Equation (24.1). Both $Z(\sigma)$, which originates in the normalization of $L(D|X, \sigma, I)$, and $\pi(\sigma)$ are absent in usual optimization-based approaches. It is the ratio of these two terms that allows us to determine the error.

Naively, one might think that including the weight directly into a restraint energy would favor large values for σ with the corresponding weight approaching 0, since this would automatically minimize the restraint energy. However, in the joint target function E_{joint} Equation (24.11), two contributions counterbalance each other:[19] χ^2/σ^2 decreases when σ increases, thus preferring large values for the error when E_{hybrid} is minimized with respect to the error. In contrast, the term $\log[Z(\sigma)/\pi(\sigma)]$ is monotonically increasing with σ.[19] The ratio of the two terms shows a finite minimum, which can be used to calculate the error, and correspondingly, the optimal weight.

Minimization of the resulting joint hybrid energy $E_{joint}(X,\sigma)$ yields the most probable structure X_{max} and the most probable error σ_{max}. In case of the log-normal model, Equation (24.9), we obtain $\sigma_{max} = \sqrt{\chi^2(X_{max})/(n+1)}$. Further analysis yields for the average weight

$$\langle w_{data} \rangle = \frac{n}{\chi^2(X)} \tag{24.12}$$

as an estimate. The average weight quantifies, in good approximation, how well the structure fits the data, independent of the size of the data set. The precision of the estimate, i.e. the width of the weight distribution, in contrast, decreases with the square root of the number of data points,[19] see Fig. 24.2.

To apply this estimate in the context of structure determination by minimization, we can iteratively update the current weight. The obtained weight is a conservative estimate since it is always smaller than $\chi^2(X_{max})/(n+1)$, the most probable weight derived from the most probable structure.

24.3 Probing Structural Dynamics by NMR

NMR is an ideal tool for probing dynamics occurring across a broad hierarchy of time-scales (Fig. 24.4).

Difficulties arise in the specific interpretation and quantification of the dynamic processes being observed. Raw experimental NMR data allows the identification of the dynamically active regions in the system over a given time-scale. However, this information is encoded in a complex manner and does not directly provide specific information about the molecular motions. To this end, experimental NMR

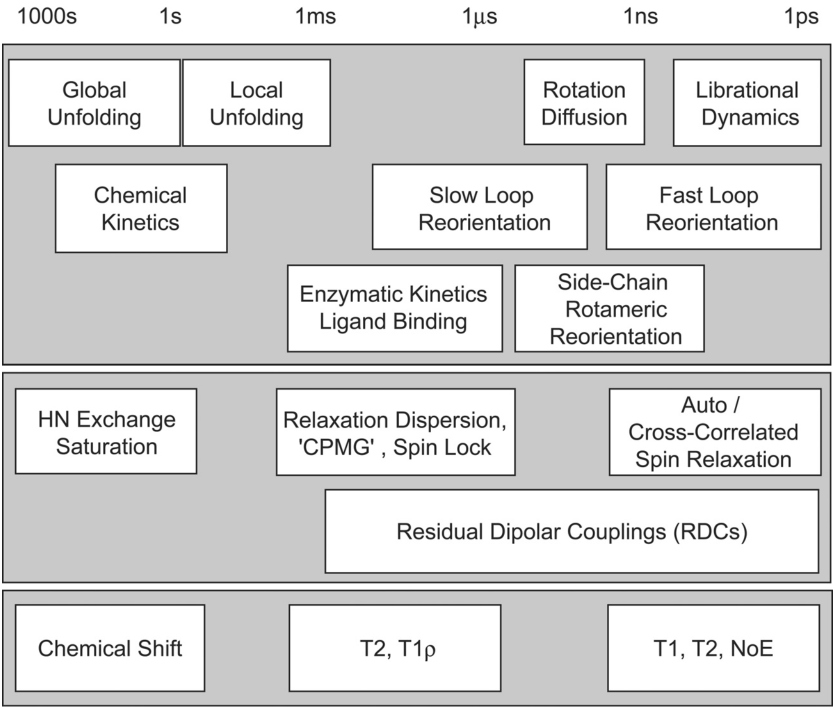


Fig. 24.4. Structural dynamics by NMR: Upper panel: type of dynamic process occurring in the bio-molecular system. Central Panel: Type of NMR experiments to probe dynamics across a particular time-scale. Lower Panel: Associated NMR observables that are probed. The time-scale is given at the top of the figure. For example: Enzymatic kinetics and ligand binding generally occur on the micro- to milli-second timescale, and can be probed using relaxation dispersion measurements and RDCs. The NMR parameter probed in the relaxation dispersion experiment is either T2 (for CPMG experiments) or T1ρ (for spin lock experiments).

data is complemented by geometric models, and increasingly, by molecular dynamics (MD) simulation to characterize at an atomistic level local dynamic processes and complex structural transitions.

The close connection between NMR experiment and computer simulation has a long history, in particular between molecular dynamics simulations and NMR experiments. Simulations are necessary to interpret the data, and NMR experiments serve to improve force fields.

24.3.1 *Experimental Approaches*

Spin relaxation measurements provide precise information about local dynamics on pico- to nano-second time-scales. The study of fast time-scale dynamics in proteins remains a rapidly developing and exciting field employing an increasing variety of experimental and theoretical methods. The importance of fast time-scale dynamics is often under-estimated: fast time-scale motions act to stabilize the protein in its folded state, and their presence is a necessary prerequisite for slower time-scale dynamics involving large-scale collective motions.

Historically, fast peptide plane motions have regularly been characterized using ^{15}N spin relaxation. In order to provide a more complete description of fast time-scale dynamics, numerous experiments have been developed to characterize dynamics for vectors other than the N–H vector in both the backbone and side-chains.[20] Cross-correlated relaxation (CCR), which arises from the interference of two relaxation mechanisms such as the chemical shift anisotropy (CSA), and dipole-dipole interaction has emerged as a powerful tool to study local anisotropic dynamics.

Many biologically important processes, such as enzyme catalysis, signal transduction, ligand binding, and allosteric regulation occur on the micro- to milli-second time-scale. Despite their obvious importance, the study of these slow motions remains a challenge to both experimentalists and theoreticians alike. The study of dynamics at these longer time-scales are centered mostly on relaxation dispersion and RDC measurements.

The characterization of motion by relaxation dispersion involves measuring the excess transverse relaxation rate caused by the exchange of nuclei between different conformations or sites with different characteristic chemical shifts. Recent methodological advances in experimental techniques have extended both the time-scale of observable dynamic processes[21] and the sensitivity[22] of the experiments to exchange processes[23,24] and ligand binding.[25] Relaxation dispersion measurements provide information concerning the location of dynamically active sites in the molecule, the exchange rates between the different conformational states, and their relative free energies (and thus their populations). However, unfortunately these experiments do not provide any direct structural information about the different conformational states, and a structural model of the observed dynamic processes is difficult to extract, making it necessary to combine this information with other experimental data[26] or MD simulations.

In the presence of a suitable alignment medium, RDCs report an average over all orientations of the magnetic dipolar interaction tensor up to a time-scale defined by the inverse of the alignment-induced coupling, making them sensitive to dynamic processes up to several milli-seconds. The ability of RDCs to probe dynamics on extended time-scales was recognized early.[3] In comparison to relaxation dispersion data, RDCs provide a detailed quantitative view of the time- and ensemble averaged protein structure and the amplitude and direction of slow time-scale motions.

24.3.2 *Interpreting Experimental Measures of Dynamics*

In analogy to structural determination, the interpretation of dynamics from NMR data needs complementary information from theory such as motional models or MD simulations. Interpretation of spin relaxation data probing pico- to nano-second dynamics traditionally employs the "model-free"[27] approach, in which the local internal motions are characterized using two parameters (an order parameter defining the spatial restriction of the motion, and a relaxation time) without making reference to a specific motional model. On the other hand, numerous explicit analytical models have been developed to describe fast time-scale local dynamic fluctuations.[28–30] One of the most popular anisotropic models is the 3D-Gaussian Axial Fluctuation (GAF) model[31] based on the observation of peptide plane motions extracted from a MD trajectory. Alternative approaches to interpreting spin relaxation make use of the strong relationship between structure and local dynamics[32,33] to rapidly predict ^{15}N order parameters from a known structure.

The study of fast local dynamic fluctuations using MD simulation is now routine. Experimentally determined auto-relaxation order parameters are regularly used to gauge the accuracy of MD simulations,[34] and more recently, MD simulations have been employed to study cross-correlated relaxation rates.[35] Continued research in the area of force-field development[36] has resulted in a marked improvement in the prediction of order parameters.[37] The inclusion of polarization and quantum effects of the atomic nuclei in the next generation of force fields will no doubt lead to further improvement. However, discrepancies between experimental and simulated order parameters may not be solely due to inadequacies in the force fields, but to incomplete conformational sampling.[38]

Despite considerable advances in our interpretation of spin relaxation data, many issues remain unresolved in the analysis of local dynamics: e.g. different models of molecular motion are equally capable of reproducing the experimental results. Many long-held assumptions concerning the local molecular geometry of the peptide plane, and in particular, the position of the amino-proton are being revisited. Also, the generally accepted idea that fast internal motion and molecular tumbling can be treated independently has been brought into question.[39]

The situation is even more complicated for slow time-scale motions deduced from RDC measurements. Even today, there are conflicting views concerning the sensitivity of RDCs to slow time-scale motions and the ability to separate the contributions to RDCs arising from structural and dynamic properties of the system. Thus, several studies on proto-typical systems[40,41] have concluded that a single copy representation of the molecule is in general sufficient to explain the data, and only a small subset of residues exhibit large amplitude fluctuations on slower time-scales.

In contrast to this, simultaneous structure-dynamics determination approaches have suggested the presence of significant slow time-scale molecular motions. Independent studies performed on Ubiquitin using model-free approaches[42–44] showed an effective homogeneous distribution of long time-scale dynamics across the molecule. A 3D-GAF based RDC analysis of the protein GB3 suggested a heterogeneous distribution of highly anisotropic long time-scale dynamics.[45,46] In part, the discrepancy between different analyses can be ascribed to the very small number of systems studied in detail to date, and no general trends can be expected as yet. However, considering the fact that the two proteins studied in most detail (GB3 and ubiquitin) show a similar fold, it is surprising that the observed distribution of slow motions appear to be so different.

Despite the continual increase in both available computational power and the efficiency of contemporary algorithms, the simulation of slow motions in proteins involving stochastic transitions over large energy barriers on the rugged and highly structured potential energy surface remains a challenging and active field of research. Considerable progress has been made in the development of new methods to sample the conformational space of proteins more efficiently, such as in conformational flooding,[47] accelerated MD,[48] and others reviewed recently.[49] "Biased potential" MD simulations have successfully identified large-scale slow collective motions in proteins.[50,51] A 0.2μs "brute-force" MD simulation of ubiquitin showed considerable dynamics occurring on time-scales beyond those probed by spin-relaxation measurements,[52] and very recently, accelerated MD simulations of the GB3 domain reliably reproduced RDC-based order parameters.[38] In light of these early successes, the study of long time-scale dynamics using a combination of MD simulation and experimental NMR holds great promise for the future.

24.3.3 *Simultaneous Calculation of Structure and Dynamics*

The most severe approximation to structure determination is the general assumption that the experimental data can be represented by a single structure, neglecting effects of internal dynamics. The ISD approach deals with statistical uncertainties in a rigorous manner, maintaining, however, the single copy model. Any ensembles generated by ISD or repeated minimization cannot represent true dynamics, but only the lack of information. It is therefore not meaningful to try to optimize the precision of the ensemble to some expected (or measured) dynamical property.

Several attempts to go beyond the single-structure approximation have been reported, using both molecular modeling approaches and MD simulation. The “Flexible Meccano” approach provides a simultaneous structure-dynamic model for RDC data: the model exploits the dependence of the measured RDCs on the orientation of the peptide plane and includes a single parameter describing the anisotropic motion about each of the 3D-GAF axes.[46]

MD simulations have been employed to refine ensembles of structures,[53] trajectories,[54] or ensembles of trajectories against the available experimental data. The latter variant[55] includes both structural and dynamic data in the fitting process, by using a potential derived from observed order parameters in the ensemble calculation. In these approaches, not only the MD force field is used to complement the absence of information on structural features, but also the motion generated by the simulation is used as a model to explain dynamical features. Several difficulties are associated with these approaches: Merely generating an ensemble of structures does not include the relative free energy weighting of each member of the ensemble; also, adding a pseudo-potential for the experimental data in a MD simulation perturbs the dynamics in a non-predictable manner, making a detailed analysis of the resulting trajectories difficult; and finally, the force fields are themselves subject to continual improvement and are capable of reproducing the NMR data to varying degrees of accuracy.

A serious consideration with these approaches is the danger of over-fitting the data, by introducing more degrees of freedom than necessary. Cross-validation can be used to try to determine the ideal number of conformers in an ensemble refinement.[56] However, recent experiences in x-ray crystal refinement[57] make it doubtful that cross-validation is sufficient as a criterion.

24.4 Future Outlook

The fundamental challenge to NMR remains to combine and reconcile all the available information, both structural and dynamic, into a complete, and therefore, intrinsically more accurate representation of the conformational space sampled by bio-molecular systems, with the aim of resolving the relationship between structure, dynamics, and function.

One of the most appealing aspects of NMR is that it is not limited to states that are well-structured, and an exciting new application of NMR-based experiments has emerged in the field of natively unstructured proteins. Fully or partially natively unstructured proteins make up a substantial part of protein sequences coded in eukaryotic genomes[58] and they play a key role in some of the most important biological processes and degenerative pathology. It is possible to measure small but finite RDCs from natively unstructured or unfolded proteins.[59] The interpretation of these RDCs is rather complex, since a single structure is certainly no longer appropriate, rather a large ensemble of interchanging structures are required to fully describe the conformational behavior of such systems. Such ensembles can be generated by empirical database random sampling[60,61] or alternatively by extended free energy sampling MD approaches.[62]

The application of NMR methods can be extended to the characterization of molecular interactions and dynamics in very large molecular assemblies[63] such as the proteasome.[64] Importantly, NMR allows the study of transient interactions and of low-affinity complexes.[65,66] Even weakly populated "excited" states of proteins can be detected.[25]

With the increased speed and reliability of X-ray crystal structure determination, the true power of NMR will thus lie in its applicability to a wide range of problems in structural biology, and its complementarity to other experimental techniques. Thus, the dynamics and interactions of a structure solved by X-ray crystallography can be characterized by NMR spectroscopy, making use of the best of the two worlds: the speed and the accuracy of X-ray crystallography, and the detailed study of dynamics and interaction that NMR can offer.

References

1. Wüthrich K. (1986) *NMR of Proteins and Nucleic Acids.* John Wiley, New York.
2. Iwahara J, Clore GM. (2006) Detecting transient intermediates in macromolecular binding by paramagnetic NMR. *Nature* **440**: 1227–1230.

3. Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH. (1997) NMR evidence for slow collective motions in cyanometmyoglobin. *Nat Struct Biol* **4**: 292–297.
4. Tjandra N, Bax A. (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278**: 1111–1114.
5. Wishart DS, Case DA. (2001) Use of chemical shifts in macromolecular structure determination. *Meth Enzymol* **338**: 3–34.
6. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. (2007) Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci USA* **104**: 9615–9620.
7. Castellani F, van Rossum B, Diehl A, *et al.* (2002) Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature* **420**: 98–102.
8. Jack A, Levitt M. (1978) Refinement of large structures by simultaneous minimization of energy and *R* factor. *Acta Cryst A* **34**: 931–935.
9. Brunger AT, Nilges M. (1993) Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR spectroscopy. *Quart Rev BioPhys* **26**: 49–125.
10. Güntert P. (2004) Automated NMR structure calculation with CYANA. *Meth Mol Biol* **278**: 353–378.
11. Press W, Flannery B, Teukolsky A, Vetterling W. (1986) *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge.
12. Jaynes ET. (2003) *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge UK.
13. Havel TF, Wüthrich K. (1985) An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. *J Mol Biol* **182**: 281–294.
14. Rieping W, Habeck M, Nilges M. (2005) Inferential structure determination. *Science* **309**: 303–306.
15. Jeffreys H. (1946) An invariant form for the prior probability in estimation problems. *Proc Roy Soc* **A186**: 453–461.
16. Habeck M, Rieping W, Nilges M. (2005) Replica exchange Monte Carlo scheme for Bayesian data analysis. *Phys Rev Lett* **94(1)**: 018105.
17. Rieping W, Habeck M, Nilges M. (2005) Modeling errors in NOE data with a lognormal distribution improves the quality of NMR structures. *J Am Chem Soc* **127**: 16026–16027.
18. Nilges M, Habeck M, O'Donoghue S, Rieping W. (2006) Error distribution derived distance potentials. *Proteins* **64**: 652–664.
19. Habeck M, Rieping W, Nilges M. (2006) Weighting of experimental evidence in macromolecular structure determination. *Proc Natl Acad Sci USA* **103**: 1756–1761.

20. Bruschweiler R. (2003) New approaches to the dynamic interpretation and prediction of NMR relaxation data from proteins. *Curr Opin Struct Biol* **13**: 175–183.
21. Loria JP, Rance M, Palmer AG. (1999) A relaxation-compensated carr-purcell-meilboom-gill sequence for characterizing chemical exchange by NMR spectroscopy. *J Am Chem Soc* **121**: 2331–2332.
22. Pervushin K, Riek R, Wider G, Wüthrich K. (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Nat Acad Sci USA* **94**: 12366–12371.
23. Eisenmesser EZ, Bosco DA, Akke M, Kern D. (2002) Enzyme dynamics during catalysis. *Science* **295**: 1520–1523.
24. Feher VA, Cavanagh J. (1999) Millisecond-timescale motions contribute to the function of the bacterial response regulator protein SpoOF. *Nature* **400**: 289–293.
25. Mulder FAA, Mittermaier AA, Hon B, *et al.* (2001) Studying excited states of proteins by NMR spectroscopy. *Nat Struct Biol* **8**: 932–935.
26. Wang L, Pang Y, Holder T, *et al.* (2001) Functional dynamics in the active site of the ribonuclease binase. *Proc Nat Acad Sci USA* **98**: 7684–7689.
27. Lipari G, Szabo A. (1982) Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *J Am Chem Soc* **104**: 4546–4559.
28. Woessner DE. (1965) Nuclear magnetic dipole-dipole relaxation in molecules with internal motion. *J Chem Phys* **42**: 1855–1859.
29. Daragan VA, Mayo KH. (1997) Motional model analyses of protein and peptide dynamics using 13C and 15N NMR relaxation. *Prog NMR Spectrosc* **31**: 63–105.
30. Chang SL, Tiandra N. (2001) Molecular dynamics and NMR spin relaxation in proteins. *J Am Chem Soc* **123**: 11484–11485.
31. Bremi T, Bruschweiler R. (1997) Locally anisotropic internal polypeptide backbone dynamics by NMR relaxation. *J Am Chem Soc* **119**: 6672–6673.
32. Zhang F, Brüschweiler R. (2002) Contact model for the prediction of NMR N-H order parameters in globular proteins. *J Am Chem Soc* **124**: 12654–12655.
33. Abergel D, Bodenhausen G. (2005) Predicting internal protein dynamics from structures using coupled networks of hindered rotators. *J Chem Phys* **123**: 204901.
34. Case DA. (2002) Molecular dynamics and NMR spin relaxation in proteins. *Acc Chem Res* **35**: 325–331.
35. Markwick PRL, Sprangers R, Sattler M. (2005) Local structure and anisotropic backbone dynamics from cross-correlated NMR relaxation in proteins. *Ang Chem Int Ed* **44**: 3232–3237.
36. Mackerell Jr AD. (2004) Empirical force fields for biological macromolecules: Overview and issues. *J Comput Chem* **25**: 1584–1604.

37. Hornak V, Abel R, Okur A, *et al.* (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**: 712–725.
38. Markwick PR, Bouvignies G, Blackledge M. (2007) Exploring multiple timescale motions in protein GB3 using accelerated molecular dynamics and NMR spectroscopy. *J Am Chem Soc* **129**: 4724–4730.
39. Tugarinov V, Liang Z, Shapiro YE, Freed JH, Meirovitch E. (2001) A structural mode-coupling approach to 15N NMR relaxation in proteins. *J Am Chem Soc* **123**: 3055–3063.
40. Ulmer TA, Ramirez BE, Delaglio F, Bax A. (2003) Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *J Am Chem Soc* **125**: 9179–9191.
41. Clore GM, Schwieters CD. (2004) How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? *J Am Chem Soc* **126**: 2923–2938.
42. Peti W, Meiler J, Bruschweiler R, Griesinger C. (2002) Model-free analysis of protein backbone motion from residual dipolar couplings. *J Am Chem Soc* **124**: 5822–5833.
43. Lakomek NA, Carlomagno T, Becker S, Griesinger C. (2006) A thorough dynamic interpretation of residual dipolar couplings in ubiquitin. *J Biomol NMR* **34**: 101–115.
44. Briggman KB, Tolman JR. (2003) *De novo* determination of bond orientations and order parameters from residual dipolar couplings with high accuracy. *J Am Chem Soc* **125**: 10164–10165.
45. Bouvignies G, Bernado P, Meier S, *et al.* (2005) Identification of slow correlated motions in proteins using residual dipolar couplings and hydrogen-bond scaler couplings. *Proc Nat Acad Sci USA* **102**: 13885–13890.
46. Bouvignies G, Markwick PRL, Bruschweiler R, Blackledge M. (2006) Simultaneous determination of protein backbone structure and dynamics from residual dipolar couplings. *J Am Chem Soc* **128**: 15100–15101.
47. Grubmüller H. (1995) Predicting slow structural transistions in macromolecular systems: Conformational flooding. *Phys Rev* **E52**: 2893–2907.
48. Hamelberg D, Mongan J, McCammon JA. (2004) Fast peptidyl cis-trans isomerization within the flexible Gly-rich flaps of HIV-1 protease. *J Chem Phys* **120**: 11919–11929.
49. Elber R. (2005) Long-timescale simulation methods. *Curr Opin Struct Biol* **15**: 151–156.
50. Schulze BG, Grubmüller H, Evanseck JD. (2000) Functional significance of hierarchical tiers in carbonmonoxy myoglobin: conformational substates and transitions studied by conformational flooding simulations. *J Am Chem Soc* **122**: 8700–8711.

51. Hamelberg D, McCammon JA. (2005) Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J Am Chem Soc* **127**: 13778–13779.
52. Nederveen AJ, Bonvin AMJJ. (2005) NMR relaxation and internal dynamics of ubiquitin from a 0.2 μs MD simulation. *J Chem Theory Comput* **1**: 363–374.
53. Kim Y, Prestegard JH. (1989) A dynamic model for the structure of acyl carrier protein in solution. *Biochemistry* **28**: 8792–8797.
54. Torda AE, Scheek RM, van Gunsteren WF. (1990) Time-averaged nuclear Overhauser effect distance restraints applied to tendamistat. *J Mol Biol* **214**: 223–235.
55. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M. (2005) Simultaneous determination of protein structure and dynamics. *Nature* **433**: 128–132.
56. Bonvin AM, Brünger AT. (1995) Conformational variability of solution nuclear magnetic resonance structures. *J Mol Biol* **250**: 80–93.
57. Chang G, Roth CB, Reyes CL, *et al.* (2006) Retraction. *Science* **314**: 1875.
58. Dyson HJ, Wright PE. (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**: 197–208.
59. Shortle D, Ackerman MS. (2001) Persistence of native-like topology in a denatured protein in 8 m urea. *Science* **293**: 487–489.
60. Bernardo P, Blanchard L, Timmins P, *et al.* (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc Nat Acad Sci USA* **102**: 17002–17007.
61. Jha AK, Colubri A, Freed K, Sosnick T. (2005) Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc Nat Acad Sci USA* **102**: 13099–13104.
62. Mukrasch MD, Markwick PRL, Biernat J, *et al.* (2007) Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. *J Am Chem Soc* **129**: 5235–5243.
63. Horst R, Wider G, Fiaux J, *et al.* (2006) Proton-proton Overhauser NMR spectroscopy with polypeptide chains in large structures. *Proc Natl Acad Sci USA* **103**: 15445–15450.
64. Sprangers R, Kay LE. (2007) Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature* **445**: 618–622.
65. Jr DCW, Cai M, Suh JY, Peterkofsky A, Clore GM. (2005) Solution NMR structure of the 48-kda IIAMannose-HPr complex of the *Escherichia coli* mannose phosphotransferase system. *J Biol Chem* **280**: 20775–20784.
66. Tang C, Iwahara J, Clore GM. (2006) Visualization of transient encounter complexes in protein-protein association. *Nature* **444**: 383–386.

This page intentionally left blank

Section V

Selected Topics

This page intentionally left blank

Chapter 25

Docking for Neglected Diseases as Community Efforts

M. Podvinec*,†, T. Schwede† and M. C. Peitsch‡

25.1 Introduction

In the previous chapters, the importance of structure-based computational approaches to the development of drugs, from the selection of first hits to the prediction of pharmacological and ADME/Tox properties of candidate compounds, was discussed. Here, we shall discuss how the current state of the art in computational simulation can be combined with the recent large increase in available computational power driven by new resources, such as grid computing or community computing. Moreover, we shall discuss whether this combination can serve as a viable model for the development of drug candidates against diseases of special public interest, for instance, neglected tropical diseases.

Early in the drug discovery process, and once a suitable drug target has been identified for a given disease, a crucial step will be to identify small molecule ligands that bind to this target (so-called hits) and alter its activity (i.e. inhibition or activation). This is performed by screening large collections of compounds in dedicated assays. The

*Corresponding author.
†Swiss Institute of Bioinformatics and Biozentrum der Universität Basel, Klingelbergstrasse 50, CH-4056 Basel, Switzerland. E-mail: michael.podvinec@unibas.ch
‡Novartis Institutes of BioMedical Research, Postfach, 4002 Basel, Switzerland.

resulting hits can, if confirmed by one or more other relevant biological assays, become the starting point of a compound optimization process which aims at identifying analogues with a maximized therapeutic window, i.e. high efficacy for low toxicity. While compound screening is predominantly an experimental approach, structure-based computational approaches (so-called virtual screening) provide an alternative and complementary way to identify such hits. If the three-dimensional structure of the protein target is not available, computational approaches can exploit knowledge about the structure of known active compounds. With a reasonable number of pharmacologically characterized compounds, the study of the relationship between their structure and the respective activities can help divulge structural features shared by the active compounds. Understanding the ensemble of steric and electronic features necessary to ensure optimal interaction with the biological target, allows pharmacophore searches to be performed (see Chapter 18). These searches evaluate how individual compounds match the ensemble of relevant features and thereby allow scientists to sift through large collections of small molecular structures to pick and choose potential hits pending validation in further laboratory assays.

In cases where high-quality experimental three-dimensional structures or models of the target protein alone or in complex with a ligand are available, molecular docking approaches (see Chapter 17) can be used to simulate the non-bonded chemical interactions between the target protein and individual compounds stored in large libraries. Notably, this approach is of particular interest when none or only very few active compounds are available for a protein target. In principle, the intra- and intermolecular interactions of a protein-ligand complex can be simulated using molecular dynamics of the molecules in solution. In practice, however, the processes and molecular rearrangements observed in small molecule binding occur on time scales which currently preclude their accurate simulation as the available computing power is too low. Therefore, several docking algorithms have been designed that implement a number of simplifying assumptions and optimization strategies to enable docking with an acceptable amount of computational effort.

A commonly voiced skepticism regarding the validity of docking in drug discovery considers that the simplification of the physics of ligand binding, often embodied in empirical scoring functions, and the truncations of the conformational search space are problematic and limit the accuracy of binding predictions obtained by this approach. This criticism is fundamentally correct; however, the extent of the introduced errors is debatable. Recent publications from both the academia and industry convincingly demonstrate that careful application of structure-based virtual screening in combination with follow-up experimental verification can indeed lead to the discovery of new compounds active against diverse clinically relevant drug targets, e.g. NF-κB,[1] the nuclear receptor PPARγ,[2] or the CK2 protein kinase,[3] and are able to complement assay-based high-throughput screening approaches.[4] Further examples of successful structure-based virtual screening have recently been summarized in Ref. 5.

The computational intensity of published docking algorithms varies significantly. At one end of the spectrum, fast, rigid docking algorithms such as FRED can process dozens of compounds per second on a single modern CPU.[6] More typical, however, are processing times of one to several minutes per compound, e.g. using the AutoDock or Glide XP algorithms.[7,8] On the far end of the spectrum are techniques that estimate binding free energies based on molecular dynamics that require several CPU days to complete, rendering them currently still too complex to be used on large compound sets.[9,10]

25.2 Grid Computing

Traditionally, high-performance computing is done on large, monolithic multiprocessor machines with shared memory. These computers allow for fine-grained parallelization of tasks, as large sets of data can be exchanged efficiently between processes and processors. The drawbacks are the high initial and maintenance cost of such specialized systems, and the fact that these resources quickly become coveted assets. Over the last decade, cluster systems built from inexpensive commodity hardware have become popular for scientific and engineering applications wherever parallelization can be realized on a coarser level.

Inter-node process communication is commonly achieved through the message-passing interface (MPI) standard.[11] Compute clusters managed by a batch-queue local resource management system (LRMS) are the standard infrastructure for most computational problems in drug development today.

In recent years, grid computing has emerged as a new trend in high-performance computing,[12,13] a form of parallel computing well suited to tackle *embarrassingly parallel* problems. Here, a big problem can simply be subdivided into a large number of smaller problems that are data-independent of each other. At its core, grid computing promotes the unification of geographically and organizationally diverse computing resources, storage elements and even experimental instrumentation into a single-sign-on, decentralized entity that provides computing power on demand. In such an environment, compute elements can be anything from massively multiparallel computers to desktop PCs that process data during idle times of their CPU. While compute clusters are built as homogeneously as possible, grids are inherently heterogeneous in terms of hard- and software, and interfaces have to be defined to allow these disparate resources to communicate. Practically, this requires a software infrastructure (often referred to as grid middleware) that creates the necessary framework for authenticated, secure exchange of data, resource brokering to match job requirements and site availabilities, and monitoring and accounting of resource and job states, to allow computation on heterogeneous computational resources without the end user needing to know the precise details of the system configuration where his job is being executed.

A number of feature-rich middleware frameworks are currently in use and constantly being enhanced, such as NorduGrid ARC/KnowARC,[14] gLite,[15] Condor,[16] Globus,[17] UNICORE,[18] Univa UD Cluster Express/MetaProcessor,[19] or the Berkeley Open Infrastructure for Network Computing (BOINC),[20] to name just a few examples. Interestingly, while unification of resource access is at the heart of grid computing, the development of unified interfaces between the middlewares themselves has only recently begun, e.g. by scheduling grid jobs from gLite or Globus and Condor, and by making ARC and gLite interoperable.

On the global scene, there are a number of public PC grids that attract the attention of volunteers to contribute their computing capacity to worthy causes. The first volunteer grid to gain mainstream media attention was SETI@home from UC Berkeley,[21,22] where participating computers analyze radio astronomy data in an ongoing search for signals of extraterrestrial intelligent origin. The same group later developed the open-source BOINC infrastructure for volunteer computation, which currently supports more than 20 large-scale computing projects.[20,23] Distributed computing projects have shown their ability to deal with a range of computationally complex problems in mathematics, cryptography, epidemiology, climate prediction or accelerator design.[24–27] PC Grids have also been used on a number of biomedical research problems, and one of these projects, the folding@home project run by the group of V. Pande[28] has recently announced its "virtual supercomputer" to constitute the largest known computer to date, having reached a performance peak of 1.2 petaflops.[29] From these numbers, it becomes clear that computationally intense projects that have the scientific and moral appeal to capture the imagination of the public can tap into very significant distributed computing resources, as long as the problem can be adapted to a grid computing model. On a more modest scale, desktop PC grids have been successfully deployed in a number of academic institutions and companies, ours including, to better utilize already existing compute infrastructure.

25.3 Grid Computing in Biomedical Research

Modern biology has become a science of information, analysis and prediction, and computing is firmly established as an essential component of biological research. It is therefore no surprise that computational biology is, after high-energy physics, among the most avid adopters of grid-based approaches for data management and processing.[30] Consequently, projects interested in answering chemical or biomedical questions using grid computing have emerged as virtual organizations within many grid projects, such as the Biomed virtual organization in EGEE or the NDGF BioGrid project, or as independent trans-institutional organizations, such as BIRN, the NIH biomedical informatics research network, the caBIG initiative of the US National Cancer Institute, the myGRID project in the UK, or the HealthGrid initiative.[31] Along with a handful of other life sciences applications, virtual compound screening (high-throughput docking) is a prime example of an application well-adapted to grid computing. Indeed, docking a molecule into a binding site is a discrete operation that can be executed for all molecules in a library in parallel.

The advantages of using grid computing in molecular docking campaigns are evident. Grid-based computations promise to provide ample computational resources, allowing the execution of even more elaborate screening campaigns where increasingly large compound libraries are, e.g. docked into receptor structure ensembles, or where more rigorous but time-consuming procedures for the sampling of the conformational space and for the scoring of resulting poses can be used, leading to a comprehensive and more accurate sampling and evaluation of the conformational, enantiomeric and tautomeric states of each ligand, including the consideration of multiple protonation states. Moreover, a number of protocols for ligand docking have been published that are able to take induced fit of the receptor or ligand polarization in the environment of the receptor into account. Both phenomena are known to be crucial in some cases of ligand binding, but their treatment is prohibitive in time, unless computational resources are considerably increased. This barrier can be lifted by accessing the power of distributed computers through grid computing.

To efficiently perform large-scale docking on a grid infrastructure, a number of non-trivial challenges specific to the grid computing domain need to be met. In contrast to computation on Linux clusters, which are common today in academic and industrial molecular modeling groups, the virtual cluster formed through grid computing is dynamic and heterogeneous. Remote compute centers may occasionally only provide backfill capacity, and may only make their resources available as long as there are no queued computations with higher priority (e.g. from in-house research groups or time-critical). In our experience, alongside high-performance clusters, individual desktop PCs can provide significant resources to a mixed grid computing approach able to federate the two types of computers.

In most cases, grid computing environments are therefore heterogeneous in terms of the type of CPU, operating system, local resource management system, and network access policies. This heterogeneity is a major challenge to successful grid computation. Most importantly, software in the field of computational biology and chemoinformatics has often been developed on a single computational platform without much investment to achieve portability. We and others experienced that many software packages are not numerically stable when executed on different hardware or software platforms,[32] (F. Grey, cited in Ref. 33). Such differences may lead to contradictory scientific conclusions depending on the platform where a calculation was performed. There is, therefore, a clear need for validation of numerical stability of algorithms used on grid architectures and, in some cases, additional porting effort.

Efficient distribution, updating, collection and management of the large data sets generated by large-scale docking campaigns is in itself a daunting task, and one that many standard applications and grid middleware stacks are not currently fully equipped to deal with. Moreover, in contrast to traditional batch queues, a docking pipeline needs to be able to flexibly deal with the large number of failed and erroneous work units caused by the dynamic and heterogeneous nature of the resource. For the time being, an end user is often left to create his own work around these problems.

One of the aims of the SwissBioGrid initiative was to further explore the requirements for a grid middleware to support computational life sciences. Some of the issues mentioned previously were identified in the course of this project, as well as partially addressed.[32] One outcome was the development of ProtoGRID, a simple framework for the execution of grid jobs on the SwissBioGrid computing resources. This middleware allows the submission of computational jobs to compute clusters managed by a diverse set of LRMS (SGE, PBSpro, Torque, Platform LSF), as well as to PC-Grids managed by GridMP. Three features were central to the design of ProtoGRID. 1) Users must be able to use pre-deployed software, but not deploy their own binaries for security and cross-platform stability issues; 2) The grid middleware must handle input and output data for grid jobs using a data proxy that intelligently caches reusable data; 3) The system must be flexible in accommodating local site policies and configuration settings. The software consists of two Linux daemons running centrally, the Grid Node Manager and Grid Data Manager, that deal with resource lookup and brokering, as well as data distribution. For each local resource, a QueueWrapper daemon is responsible for publishing current status information to the resource broker, as well as polling the scheduler and data manager for jobs and data sets. This system was successfully used to distribute docking tasks among four geographically and institutionally compute resources. Some performance metrics can be found in Ref. 32.

25.4 Grid-based Computation to Discover Drug Candidates Against Targets of Public Interest

As mentioned above, grid computing is an attractive platform and well suited to process high-throughput docking of chemical compounds into protein 3D-structures. More importantly, docking efforts against targets of public interest have excellent chances of gaining not only access to large transnational compute grids, but also of recruiting many altruistic volunteers donating their workstations idle time to PC-grids. Some volunteers, voices can be found in Ref. 34.

Not surprisingly, a number of drug discovery projects have started to make use of such grid resources and public volunteers, aiming at three types of disease targets:

(1) Targets with a peak in public awareness: In the wake of the terrorist attacks in the US in 2001, projects targeting smallpox and anthrax were launched.[35] Other targets belonging to this class are SARS[36] or the avian flu neuraminidase.[37] These targets, receiving intense media attention, can muster large computing resources in a short time. The narrow time window and therefore the rush in setting up such campaigns, however, often lead to a constellation

where large amounts of data are crunched in a short time with an unclear plan of how these results will be transferred to subsequent *in vitro* and *in vivo* assays.

(2) Diseases that affect large parts of the population and/or present unmet clinical needs. Examples for this class are cancer or AIDS.[38,39] Efforts targeting unmet clinical needs have good chances of obtaining support from funding agencies and resource providers, and can rely on a large number of sympathetic volunteers.

(3) Targets belonging to the class of *neglected diseases*: These are afflictions that mainly affect the developing countries, and therefore hold little commercial interest. Hence, they rarely attract large investments in drug discovery and development. In recent years, however, this class of diseases has attracted increasing attention from NGOs, academic and industrial players, and has led to the formation of private-public partnerships.[40] As a consequence, a number of drug discovery efforts against this class of targets have been launched. They rely on grid-based virtual screening of public compound libraries to provide a starting point of a medicine development program against diseases, such as malaria or dengue.[32,41]

We believe that a strong case can be made in favor of the last of these classes and in particular for infectious diseases where private-public partnerships are an essential strategy to fill the drug discovery pipelines.[42] While the second class of projects, targeting large unmet clinical needs, certainly constitutes a valid and worthy cause, these targets are at the same time of high commercial interest and thus are being actively pursued by pharmaceutical companies. In stark contrast, neglected diseases clearly are not primary discovery targets in commercial research. The public-private partnership model, however, can provide a novel approach to drug discovery, alleviating at least some of the high costs of drug discovery and development through collaboration with academia and non-governmental organizations.

We foresee a strong emphasis on computational structure-based methods in such projects, which provides a dual opportunity: firstly, computational approaches to drug discovery can be used to save substantial parts of the high costs spent in a small-molecule screening and optimization campaigns. Secondly, by combining the current best thinking in computational drug discovery and virtually unlimited computing resources, we can explore the extent to which computational methods can speed up and facilitate the selection of lead compounds. Such a private-public partnership would have access to large computational resources that allow the docking of large publicly available libraries of (purchasable) chemical compounds, e.g. from the ZINC database.[43] Results from these docking studies could then be refined by more time-consuming molecular dynamics-based methods like the MM-GBSA (Molecular Mechanics-Generalized Born Surface Area) or LIECE (Linear Interaction Energy with Continuum Electrostatics) approaches.[44,45] The hits identified by these approaches would then be transferred to academic or industrial laboratories for experimental validation. Such a setup can generate the momentum needed to jump-start a drug discovery program by a pharmaceutical industry partner with the required know-how in drug discovery and development.

25.5 Public-Private Partnerships: A Model for Drug Discovery Against Neglected Diseases

In the following, we outline a project that probes the feasibility of public-private partnerships in finding and developing drug candidates against neglected diseases. For the private partners, such an arrangement is of interest, as they can build on academic expertise in target identification and validation, while jumpstarting the drug discovery process with a list of selected compound hits. Conversely, collaborating with an industry partner is of benefit to academic drug discovery efforts, which lack the necessary drug development know-how. As compounds progress beyond the initial hit stage, academia needs to rely on the rich drug development experience of the industrial setting.

The project described below targets dengue fever as an exemplary neglected disease, and was conceived in 2004 to demonstrate the validity of the public-private partnership concept. We started the project by seeking a strong partnership for the computational and experimental parts of the project: The Swiss Institute of Bioinformatics and the University of Basel act as academic partners and provide structural bioinformatics expertise. The Novartis Institute for Tropical Diseases (NITD) in Singapore is a drug discovery research institute dedicated to finding new drugs for the treatment of tropical diseases, and contributes the experimental follow-up. On the computational side, Schrödinger, Llc provides essential computational chemistry tools and specific scientific expertise. As the project as described here has extraordinary demands in terms of computing and data storage, we co-founded the SwissBioGrid initiative to establish grid computing resources to the Swiss biomedical research community.[32]

The target disease, dengue fever, is a viral disease that is transmitted by mosquitoes, predominantly by *Aedes aegyptii*. Fifty to 100 million cases of dengue fever are estimated to occur annually, with the numbers on the rise, due to increasing urbanization (the *Aedes* mosquito thrives well in urban areas) and the failure to effectively control the spread of the mosquito vector. For 40% of the human population, dengue is a daily fact of life, as they live in areas where the virus is endemic. Initial disease symptoms are flu-like, comprising fever, headache and severe myalgia (break-bone fever). More severe cases can progress into dengue hemorrhagic fever and dengue shock syndrome with considerable lethality. The current treatment is non-specific and symptomatic with a regimen of analgesics and fluid replacement.

The dengue virus belongs to the *flaviviridae* genus of enveloped viruses and exists in four distinct serotypes. It possesses a single-stranded positive RNA genome, which is translated into a single polyprotein during the viral life cycle. Subsequently, the polyprotein is cleaved by cellular and viral proteases into 10 mature proteins. Three of the proteins have a structural role (C, prM, and E). In addition, seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) are formed. Of many of these non-structural proteins, the exact role or roles in the viral life cycle are not yet fully understood. Due to the compact viral genome and repertoire of proteins, there is a conspicuous accumulation of multifunctional proteins.[46,47]

Among the 10 proteins processed from the viral polyprotein, three can be considered of special interest as targets for drug development:

1. Glycoprotein E, along with the viral M protein, composes the viral surface. Glycoprotein E plays an important role in opening the viral envelope upon entry into the cell. The structure of the E protein has been determined in several oligomerization states.[48,49]
2. Non-structural protein 3 (NS3), a multifunctional protein which exhibits protease, helicase, nucleoside 5′-triphosphatase and 5′-terminal RNA triphosphatase activities. Of these, at least two functions have been demonstrated to be essential for the virus. The proteolytic activity of NS3 maps to the N-terminal region of this protein. It is involved in cleaving the viral polyprotein into the mature protein forms. At the C-terminal part of NS3, the RNA helicase domain is located. In the flaviviridae family, this function is essential: viruses with impaired helicase activity have been shown to be unable to replicate. This enzyme is capable of unwinding duplex RNA structures by disrupting the inter-strand hydrogen bonds. This activity is associated with NTP hydrolysis. Recently, the structure of the whole NS3 protein in complex with 18 residues of the NS2B cofactor has been determined.[50]
3. The non-structural protein 5 (NS5), likewise, demonstrates several activities: at its N-terminus, a S-adenosyl L-methionine-dependent RNA methyltransferase is found. At the C-terminus, motifs characteristic of RNA-dependent RNA-polymerases are found. The methyltransferase is involved in the posttranscriptional capping process of RNA. This enzyme catalyzes the methylation of the cap-guanoside at either the 2′-oxygen of the sugar moiety or the 7-nitrogen of the guanine. For host RNA, these reactions occur in the nucleus. Since viral RNA is produced in the cytoplasm of the infected cells, many viruses provide their own capping machinery while relying on the host for subsequent translation of mature mRNAs. As the cellular and viral capping

apparatus differs significantly in fungi, metazoans, and humans,[51] the viral methyltransferase is an enzyme that can potentially be specifically targeted. The structure of the dengue NS5 methyltransferase has been solved alone or in complex with nucleoside analogues and S-adenosyl L-homocysteine (AdoHcy).[52,53]

To select the most promising target to focus our initial efforts on, we have analyzed the prevalence of mutations by considering the published sequences of clinical dengue isolates of all the serotypes. In this study, unsurprisingly, the amino acid sequences of NS3 and NS5 appeared to be more conserved than that of the envelope protein. After further structural examinations, we chose the NS5 methyltransferase as our target of first choice. Here, competitive inhibition is feasible at two binding sites: the RNA-cap binding site, and the site of the cofactor, S-adenosyl L-methionine.

We next composed a library of commercially available compounds, predominantly from the ZINC database,[43] totaling close to 6 million individual compounds. We considered purchasable compounds only, as these stand a reasonable chance of being available in sufficient amounts for subsequent *in vitro* validation. Next, all the compounds were docked into each of the methyltransferase binding sites. No prior selection according to drug-likeness or similarity criteria was performed. This is in contrast to the widespread practice of selecting a "focused library" before docking. We opted for this approach for two reasons: only a few inhibitors of the dengue methyltransferase have been described, and focusing a library on similarity to these may erroneously restrict our search space. Moreover, using a grid computing approach, we can consider computational resources as non-limiting.

Docking calculations were performed using the Glide 4.5 algorithm, following a three-stage protocol, ranging from a fast initial "high-throughput" screen to the final calculation using the extra-precision Glide XP protocol. Figure 25.1 compares the best docked pose of the cognate ligands AdoHcy (a) and ribavirin triphosphate (RTP) (b) as compared to the experimentally determined structure.

After each screening round, compounds ranked poorly were discarded, resulting in a progressively reduced selection of compounds.

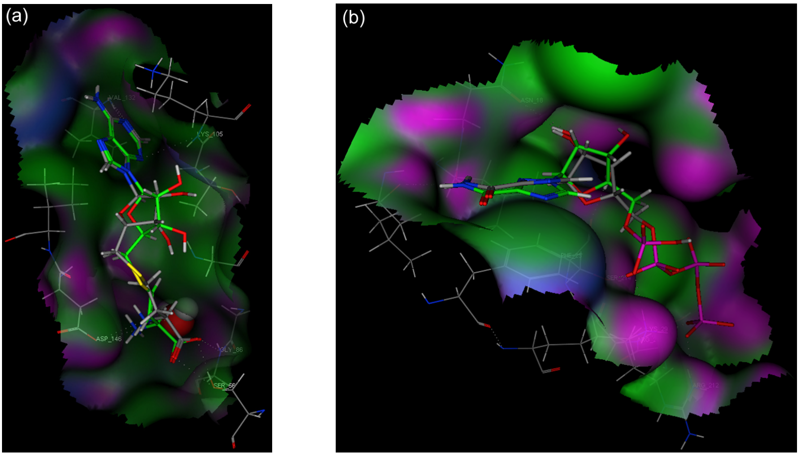


Fig. 25.1. Experimentally determined binding mode (green sticks) from PDB:1R6A and re-docked conformation (orange sticks) for NS5 ligands. **(a)** AdoHcy is bound in an elongated binding pocket. Reproduction of the adenine moiety pose is near-perfect, and hydrogen bonds to the amino acid group of the molecule are consistent between experimental and docked structure. **(b)** For docking, the position of Lys22 was changed to another rotamer to slightly open the site near the imidazole group of RTP. We expect that this caused the slightly altered placement of the ligand above Phe25. Notably, RTP was found among the top compounds selected for testing in our docking campaign.

The majority of the docking calculations were performed on a PC grid, underlining the usefulness of such a resource for parallel docking approaches. In total, more than 7.5 years of CPU time were spent in the docking calculations, which were completed within 72 days, including filtering and analysis steps between screening runs, which were not fully automated. For each binding site, 4000 top-scoring compounds were then subjected to post-processing, where a more rigorous XP docking protocol, starting from a set of alternative conformers, was executed. Finally, three scoring function variants were examined: the original Glide XP score (GScore), a variant of GScore accounting for ligand strain[54] and binding energy as estimated by the MM-GBSA approach.[44] By adding up individual ranks, compounds were given a consensus rank. Finally, molecules top-ranked in the consensus score or in each of the constituent scores were collected into a list for visual inspection. A "tasting panel" of five experts then visually inspected these short lists. From their recommendations, 200 molecules were finally selected for *in vitro* validation.

In less than three months, we were thus able to reduce an initial compound list of 6 million molecules to 200 plausible hits that are commercially available. Currently, these compounds are being assayed for their ability to inhibit the function of the viral methyltransferase. In these assays, the transfer of a labeled methyl group from the cofactor SAM to an RNA cap substrate is measured in the presence of potential inhibitor compounds.[53] These measurements will lead to IC_{50} (concentration at half-maximal inhibition) values. Viral replication assays, where the increase of viral RNA or protein in suitable mammalian cells is followed in the presence of suspected inhibitor compounds, lead to the determination of EC_{50} (concentration at half-maximal effect) values. While this work is currently ongoing, we expect to find low-micromolar inhibitors of the methyltransferase. We recently performed a smaller study, starting with 127 000 compounds of the National Cancer Institute Developmental Therapeutics Library. Here, two active compounds with low-micromolar IC_{50} were identified among 36 tested high-ranking compounds.

25.6 Discussion

We believe that the scientific focus of community-based public-private partnerships is slightly different from virtual screening campaigns carried out in academic or industrial lead finding programs. In a first approximation, the available computer power can be considered non-limiting, as evidenced by some of the larger international volunteer computing efforts mentioned above. While porting applications to support a grid approach is not a trivial and short-time undertaking, the rewards in terms of accessible, relatively cheap computing power are substantial. Therefore, such a scenario can become a fertile ground on which to explore how much computational chemistry and biology can achieve in drug discovery, the aim being to use exhaustive screening and further refinement protocols to reduce costly laboratory assays to the necessary minimum.

In our opinion, the most crucial element in an endeavor as presented here is the inclusion of an experimentalist group as collaborators even before the start of the large-scale docking. While many docking campaigns report impressive enrichment, simple virtual screening does not go far enough. There is a number of high-profile, extremely large screening projects that have not produced any results beyond the lists of scored compounds that may or may not be made publicly available. Evidently, the plan of waiting for an experimental group to pick up these results and continue work is failing. One reason may be the unwillingness of a laboratory to invest substantial amounts of time and material costs into an abstract artifact, such as a list of compound identifiers. With close scientific interactions and involvement already established between the computational team and the experimentalists prior to the screening phase, we have been able to foster interdisciplinary trust and collaboration necessary for such a project. In the final analysis, finding compounds active in a laboratory assay is the touchstone by which to measure the success of such a campaign.

If scientifically sound, a public-private collaborative project has the potential to rapidly generate a shortlist of interesting compounds and thus kick start the drug discovery program. We must appreciate this work in the appropriate context. Developing a drug is a lengthy and expensive endeavor. In the year 2000, an average of US$802 million were spent per successfully developed drug, and the process took 12 years on the average.[55] Knowing that significant compound attrition occurs even at the late-stage drug development, and comparing this large cost to the savings generated by a successful high-throughput docking campaign, which may amount to US$5–10 million, one may wonder whether this effort is at all enough to make a difference. We firmly believe the contribution to be much more significant than its direct monetary value may suggest. Indeed, beyond monetary savings, public-private partnerships are interesting in terms of bidirectional technology and knowledge transfer, and if well done and managed, harness the enthusiasm of the public sector to drive projects which may not otherwise be resourced. Moreover, the successful selection of hits from a virtual screen can make the crucial difference in getting a drug development program underway. Another important factor may be the concept of "piggy-backing,"[42] where a drug may be developed for two indications, one of which having the potential to recover at least some of the development costs.

Clearly, there are various ways of starting a drug discovery program, including high-throughput docking, high-throughput experimental screening, selection of focused libraries, concentrating on natural compound libraries, and many of these have been suggested as the starting point to generate new leads against neglected diseases. We and others have decided to place a strong emphasis on computational approaches, as these can be carried out with more easily accessible resources than the facilities and budget necessary to perform high-throughput screening.

The general consensus is that computational structural biology approaches should be measured by how well they reproduce experimental evidence, and in the case of docking, how well they reproduce the hit list obtained from a high-throughput screening campaign. Interestingly, though, the most potent inhibitor of CK2 kinase was discovered by docking and not experimental means.[3] The cumulative experience of many subsequent docking and screening campaign shows that by and large these approaches are complementary and yield hit lists which do not completely overlap. The combined hit lists are, therefore, a better starting point for lead selection and often provide scaffolds coming from both docking and screening campaigns. Given the recent successes in hit and lead identification in the industry, docking has become a routine process[3] which is applied whenever enough structural information about the protein target is available.

25.7 Future Outlook

As a consequence of the case study described above, we believe that there is great potential in establishing an *in silico* drug discovery platform focused on neglected diseases. The core role of this platform would be to manage all the computational aspects of such projects, through the creation of an *in silico* pipeline implementing standardized ways to move targets and compounds through the docking process, to manage and to analyze the data. Everything else would be kept modular — target disease, participating institutions and companies, and computational approaches (e.g. selection of docking algorithms and their parameters, compound libraries, compound selection flow chart, etc.). The platform would be governed by a body ensuring the scientific soundness of the proposals and controlling the quality of the produced data.

While this chapter focused on the automated docking of large compound libraries into a viral target enzyme, the computational pipeline used in this context could be extended to include further elements of drug discovery amenable to *in silico* approaches. Indeed the *in silico* drug discovery pipeline[56,57] could start with the homology-based 3D-structure modeling of the protein targets,[58] continue with the virtual screening of compounds for hit finding, employ targeted docking to support lead optimization, and producing ADME and toxicology predictions and alerts. Many of the individual elements of this pipeline already exist, but need to be integrated into a coherent pipeline that flexibly provides this computational toolbox to drug discovery projects.

Neglected diseases are an ideal topic for such ventures — on the one hand, this is a just cause for which many companies and individuals are willing to donate some of their efforts, resources or time; on the other, there is a dire need for new leads and new treatment modalities to tackle this class of diseases.[42] Beyond the altruistic motivation, many of the neglected diseases are caused by pathogens that increasingly threaten to invade the industrialized world (some have already begun to do so). The Asian tiger mosquito (*Aedes albopictus*) has in recent years become endemic in the southern US, in eastern Canada, and in 12 European countries, where it caused an outbreak of Chikungunya fever in Italy in 2007.[59] *Aedes albopictus* is equally able to transmit the dengue virus. At the same time, international air travel and tourism lead to imported cases of dengue fever every year (93 cases of confirmed or probable dengue fever were reported in the Vienna central hospital between 1990 and 2005[60]), creating a situation where the occurrence of local cases of dengue fever becomes a simple matter of probabilities. It would be deceptive to rely on perceived advances in civilization and technology to ward off this danger. Highly developed cities, such as Singapore and Madrid, have not been able to contain mosquito populations despite significant efforts. While these ongoing developments will most likely change our level of interest in these diseases, it might be prudent to start developing drugs now.

Acknowledgements

The authors would like to acknowledge their project partners for their important contributions to the dengue docking project, and for many invaluable discussions around the topic: Peter S. Shenkin, Paul Sanschagrin, and Jörg Weiser from Schrödinger, Llc., Subhash Vasudevan, Siew Pheng Lim and Sebastian Sonntag from the Novartis Institute for Tropical Diseases. We would also like to thank all the members of the SwissBioGrid initiative, and finally, everyone whose desktop computer contributed to the production of results for the dengue project.

References

1. Leban J, Baierl M, Mies J, *et al.* (2007) A novel class of potent NF-kappaB signaling inhibitors. *Bioorg Med Chem Lett* **17**(21): 5858–5862.
2. Scarsi M, Podvinec M, Roth A, *et al.* (2007) Sulfonylureas and glinides exhibit peroxisome proliferator-activated receptor gamma activity: a combined virtual screening and biological assay approach. *Mol Pharmacol* **71**(2): 398–406.
3. Vangrevelinghe E, Zimmermann K, Schoepfer J, *et al.* (2003) Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J Med Chem* **46**(13): 2656–2662.
4. Doman TN, McGovern SL, Witherbee BJ, *et al.* (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J Med Chem* **45**(11): 2213–2221.
5. Cavasotto CN, Orry AJ. (2007) Ligand docking and structure-based virtual screening in drug discovery. *Curr Top Med Chem* **7**(10): 1006–1014.
6. McGann MR, Almond HR, Nicholls A, *et al.* (2003) Gaussian docking functions. *Biopolymers* **68**(1): 76–90.
7. Morris GM, Goodsell DS, Halliday RS, *et al.* (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**(14): 1639–1662.

8. Friesner RA, Murphy RB, Repasky MP, *et al.* (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem* **49**(21): 6177–6196.
9. Grosdidier A, Zoete V, Michielin O. (2007) EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins* **67**(4): 1010–1025.
10. Thorsteinsdottir HB, Schwede T, Zoete V, Meuwly M. (2006) How inaccuracies in protein structure models affect estimates of protein-ligand interactions: computational analysis of HIV-1 protease inhibitor binding. *Protein Struct Funct Bioinform* **65**(2): 407–423.
11. **Message Passing Interface Forum** [http://www.mpi-forum.org]
12. Foster I, Kesselman C, Nick JM, Tuecke S. (2002) Grid services for distributed system integration. *Computer* **35**(6): 37–46.
13. Foster I, Kesselman C, Tuecke S. (2001) The anatomy of the grid: enabling scalable virtual organizations. *Int J High Perform Comput Appl* **15**(3): 200–222.
14. Eerola P, Ekelöf T, Ellert M, *et al.* (2006) Roadmap for the ARC grid middleware. In J Dongarra, K Madsen, J Wasniewski (eds.), *Applied Parallel Computing State of the Art in Scientific Computing*. Springer.
15. Laure E, Fisher SM, Frohner A, *et al.* (2006) Programming the grid with gLite. *Comput Meth Sci Tech* **12**(1): 33–45.
16. Frey J, Tannenbaum T, Foster I, *et al.* (2002) Condor-G: a computation management agent for multi-institutional grids. *Cluster Comput* **5**(3): 237–246.
17. Foster I, Kesselman C. (1997) Globus: a metacomputing infrastructure toolkit. *Int J Supercomp Appl High Perform Comput* **11**(2): 115–128.
18. Streit A, Erwin D, Lippert T, *et al.* (2005) UNICORE — from project results to production grids. In L Grandinetti (ed.), *Grid Computing: The New Frontiers of High Performance Processing*, pp. 357–376. Elsevier.
19. **Univa UD** [http://www.univaud.com]
20. **Berkeley Open Infrastructure for Network Computing (BOINC)** [http://boinc.berkeley.edu]
21. Anderson DP, Cobb J, Korpela E, *et al.* (2002) SETI@home — an experiment in public-resource computing. *Commun Acm* **45**(11): 56–61.
22. **SETI@HOME** [http://setiathome.berkeley.edu]
23. Anderson DP. (2004): **BOINC: a system for public-resource computing and storage.** In: *5th IEEE/ACM International Workshop on Grid Computing*. Pittsburgh, USA.
24. **distributed.net** [http://www.distributed.net/]
25. **malariacontrol.net** [http://www.malariacontrol.net/]
26. **climateprediction.net** [http://www.climateprediction.net/index.php]
27. **lhc@home** [http://lhcathome.cern.ch/lhcathome/]

28. Zagrovic B, Snow CD, Shirts MR, Pande VS. (2002) Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J Mol Biol* **323**(5): 927–937.
29. **Crossing the petaflop barrier** [http://folding.typepad.com/news/2007/09/crossing-the-pe.html]
30. Foster I. (2005) Service-oriented science. *Science* **308**(5723): 814–817.
31. Buetow KH. (2005) Cyberinfrastructure: empowering a "third way" in biomedical research. *Science* **308**(5723): 821–824.
32. Podvinec M, Maffioletti S, Kunszt P, *et al.* (2006) The SwissBioGrid project: objectives, preliminary results and lessons learned. In *2nd IEEE International Conference on e-Science and Grid Computing (e-Science 2006)*. IEEE Computer Society Press, Amsterdam, The Netherlands.
33. Bohannon J. (2005) Distributed computing. Grassroots supercomputing. *Science* **308**(5723): 810.
34. Bohannon J. (2005) Distributed computing. Grid sport: competitive crunching. *Science* **308**(5723): 812.
35. Richards WG, Grant GH, Harrison KN. (2004) Combating bioterrorism with personal computers. *J Mole Graph Model* **22**(6): 473–478.
36. Zhang W, Du X, Ma F, *et al.* (2006) (DDGrid: harness the full power of supercomputing systems. In *Grid and Cooperative Computing Workshops: GCCW '06 Fifth International Conference on*.
37. Lee HC, Salzemann J, Jacq N, *et al.* (2006) Grid-enabled high-throughput *in silico* screening against influenza a neuraminidase. *IEEE Trans Nanobiosci* **5**(4): 288–295.
38. Richards WG. (2002) Virtual screening using grid computing: the screensaver project. *Nat Rev Drug Discov* **1**(7): 551–555.
39. Chang MW, Lindstrom W, Olson AJ, Belew RK. (2007) Analysis of HIV wild-type and mutant structures via *in silico* docking against diverse ligand libraries. *J Chem Inform Model* **47**(3): 1258–1262.
40. Croft SL. (2005) Public-private partnership: from there to here. *Trans Roy Soc Trop Med Hyg* **99**: S9–S14.
41. Kasam V, Zimmermann M, Maass A, *et al.* (2007) Design of new plasmepsin inhibitors: a virtual high-throughput screening approach on the EGEE grid. *J Chem Inform Model* **47**(5): 1818–1828.
42. Nwaka S, Hudson A. (2006) Innovative lead discovery strategies for tropical diseases. *Nature Rev Drug Discor* **5**(11): 941–955.
43. Irwin JJ, Shoichet BK. (2005) ZINC — a free database of commercially available compounds for virtual screening. *J Chem Inform Model* **45**(1): 177–182.
44. Gohlke H, Kiel C, Case DA. (2003) Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RaIGDS complexes. *J Mol Biol* **330**(4): 891–913.

45. Huang DZ, Luthi U, Kolb P, *et al.* (2005) Discovery of cell-permeable non-peptide inhibitors of beta-secretase by high-throughput docking and continuum electrostatics calculations. *J Med Chem* **48**(16): 5108–5111.
46. Mukhopadhyay S, Kuhn RJ, Rossmann MG. (2005) A structural perspective of the Flavivirus life cycle. *Nat Rev Microbiol* **3**(1): 13–22.
47. Padmanabhan R, Mueller N, Reichert E, *et al.* (2006) Multiple enzyme activities of flavivirus proteins. *Novartis Found Symp* **277**: 74–84; discussion 84–76, 251–253.
48. Modis Y, Ogata S, Clements D, Harrison SC. (2004) Structure of the dengue virus envelope protein after membrane fusion. *Nature* **427**(6972): 313–319.
49. Zhang Y, Zhang W, Ogata S, *et al.* (2004) Conformational changes of the flavivirus E glycoprotein. *Structure* **12**(9): 1607–1618.
50. Luo D, Xu T, Hunke C, Gruber G, *et al.* (2007) Crystal structure of the NS3 protease-helicase from dengue virus. *J Virol.*
51. Shuman S. (2001) Structure, mechanism, and evolution of the mRNA capping apparatus. *Prog Nucl Acid Res Mol Biol* **66**: 1–40.
52. Egloff MP, Benarroch D, Selisko B, *et al.* (2002) An RNA cap (nucleoside-2′-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. *EMBO J* **21**(11): 2757–2768.
53. Benarroch D, Egloff MP, Mulard L, *et al.* (2004) A structural basis for the inhibition of the NS5 dengue virus mRNA 2′-O-methyltransferase domain by ribavirin 5′-triphosphate. *J Biol Chem* **279**(34): 35638–35643.
54. Sanschagrin P., personal communication (2007).
55. DiMasi JA, Hansen RW, Grabowski HG. (2003) The price of innovation: new estimates of drug development costs. *J Health Econ* **22**(2): 151–185.
56. Peitsch MC, Searls DB, Shapiro E, Ferguson N. (2006) Revolutionising medicine. In S Emmot, S Rison (eds.), *Towards 2020 Science*. Microsoft.
57. Lewis R, Ertl P, Jacoby E, *et al.* (2005) Computational chemistry at novartis. *Chimia* **59**(7–8): 545–549.
58. Arnold K, Bordoli L, Kopp J, Schwede T. (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**(2): 195–201.
59. Rezza G, Nicoletti L, Angelini R, *et al.* CHIKV study group. (2007) http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=ShowDetailView&Te. Infection with chikungunya virus in Italy: an outbreak in a temperature region. *Lancet 1* **370**(9620): 1840–1846.
60. Laferl H, Szell M, Bischof E, Wenisch C. (2006) Imported dengue fever in Austria 1990–2005. *Travel Med Infect Dis* **4**(6): 319–323.

Chapter 26

Protein Structure Databases

D. Dimitropoulos, M. John*,† E. Krissinel, R. Newman
and G. J. Swaminathan

26.1 Introduction

This chapter provides a description of protein structure databases with particular reference to the one constructed and maintained by the Molecular Structure Database (MSD) group at the European Bioinfomatics Institute (EBI). An in-depth understanding of databases and database terminology is not assumed on behalf of the reader. After a brief introduction, the chapter covers the curation process for newly submitted structures, the loading of data into a relational database, and the physical and logical architectures of the database. An account is provided of the problems, advantages, and disadvantages of storing protein structure data in a relational database. A brief description of current online services is given to demonstrate the uses to which protein structure data are being put.

The EBI (Table 26.1) was established in 1995 as a portal for biological databases covering a broad range of topics from nucleotide sequence to protein function. Since its inception it has hosted the EMBL nucleotide sequence database[1] and the trEMBL protein sequence database, now a part of the UniProtKB composed of

*Corresponding author.

†European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom. E-Mail: melford@ebi.ac.uk.

Table 26.1 URLs of Protein Structure Databank Services

No	URL	Description
1	www.ebi.ac.uk	European Bioinformatics Institute
2	www.ebi.ac.uk/msd	MSD group
3	www.wwpdb.org	worldwide Protein Data Bank
4	www.ebi.ac.uk/msd-srv/autodep4	AutoDep4 deposition system
5	www.ebi.ac.uk/msd-srv/emdep	EMDep protein structure depositions
6	www.ebi.ac.uk/msd-srv/emsearch	EMDB search tool
7	www.ebi.ac.uk/msd-srv/docs/dbdoc	MSDSD logical model
8	www.ebi.ac.uk/msd-srv/msdsite	MSDsite service
9	www.ebi.ac.uk/msd-srv/docs/sifts	SIFTS initiative
10	www.ebi.ac.uk/msd-srv/msdlite	MSDlite service
11	www.ebi.ac.uk/msd-srv/msdpro	MSDpro service
12	www.ebi.ac.uk/msd-srv/msdchem	MSDchem service
13	www.ebi.ac.uk/msd-srv/ssm	MSDfold service
14	www.jmol.org	Jmol open-source viewer
15	www.ebi.ac.uk/msd-srv/prot_int/pistart.html	MSDPISA service
16	www.ebi.ac.uk/msd-as/MSDvalidate	MSDanalysis service
17	www.ebi.ac.uk/msd-srv/MSDtemplate	MSDtemplate service
18	www.ebi.ac.uk/msd-srv/msdmine	MSDmine service
19	www.ebi.ac.uk/msd-srv/msdmotif	MSDMotif service

SWISS-PROT, trEMBL and PIR.[2] The MSD group[3,4] (Table 26.1) was set up in 1996 as a European initiative for the collection, organization, and distribution of macromolecular structure data. The MSD, Research Collaboratory for Structural Biology (RCSB), and PDB Japan (PDBj) came together in 2003 to form the worldwide Protein Data Bank (wwPDB) (Table 26.1)[5,6] in order to maintain and manage a single repository for macromolecular structures called the Protein Data Bank (PDB). In 2006, the wwPDB was further joined by the Biological Magnetic Resonance Data Bank (BMRB). Together, these four organizations serve as deposition, data processing, and distribution sites of the PDB archive. Furthermore, each wwPDB site provides its own view of the primary data, thus providing a variety of tools and resources to the global community. The four main areas that the MSD group has focused on are:

- Accepting and processing depositions to the Protein Data Bank (PDB).[7]
- Accepting and processing depositions to the Electron Microscopy Data Bank (EMDB).[8]
- Transforming the PDB flat file archive to a relational database system.[3]
- Developing services to search the data in the MSD database.[3,4]

The challenge of presenting the available information in an intuitive way to users from various backgrounds and expertise demands that the data are processed and structured in a meaningful and flexible way. Relational database technology offers both the flexibility and framework to achieve this goal. The MSD has applied these database technologies to the extremely complex processes of importing legacy data from the Protein Data Bank (PDB),[7] the creation of a submission system for new depositions to the PDB with automated annotation procedures, in addition to achieving data conformity and the integration of relevant information from other biological databases. Different query systems have been developed to allow access to the MSD structural database (MSDSD). The overall system has been designed from the outset to cope with the expected exponential growth in structure data through the structural genomics initiatives.[9] This chapter addresses the challenges faced in designing a robust relational database system for structural data, and introduces various deposition, search, analysis, and retrieval tools developed by the MSD.

26.2 PDB Data Deposition and Processing

Macromolecular structures are being determined at an ever-increasing rate, with improvements in protein expression, data collection, structure refinement, and computer technologies. Additionally, many worldwide structural genomics initiatives have now begun to produce a large number of structures for deposition with the PDB. The number

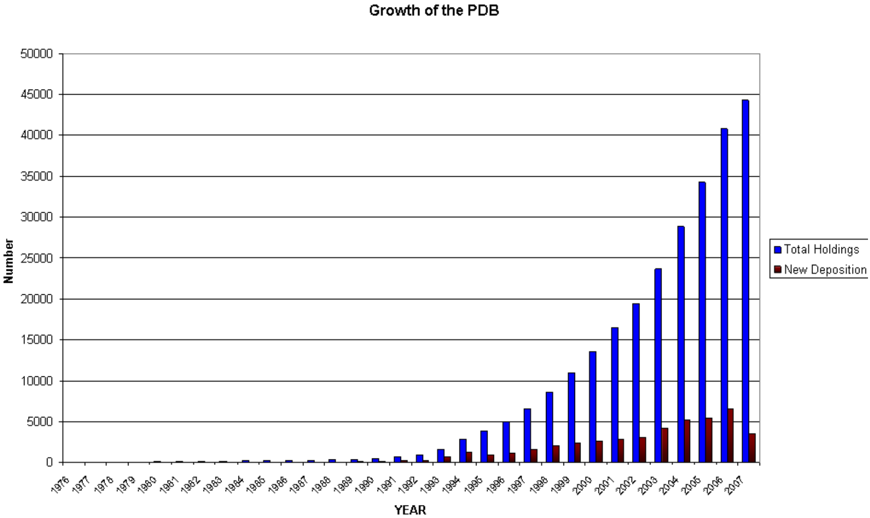


Fig. 26.1 Growth of the PDB from 1976 to June 2007.

of entries in the PDB has almost doubled from about 23 000 entries in 2003 to nearly 45 000 in 2007 (Fig. 26.1). This exponential trend is expected to continue for quite some time.

Macromolecular structures are deposited with the MSD using a deposition system called AutoDep4[10] (Table 26.1). Structures can also be deposited with the RCSB and PDBj using the ADIT system.[11] The MSD deposition tool AutoDep4 is designed to implement a consistent approach to the handling and curation of deposition data. The AutoDep system also allows value-added information (quaternary structure assessments, structure quality, etc.) to be returned in a safe and secure manner into the password-protected deposition session only accessible to the depositor, following annotation by the curation staff. Depositors may also choose to install a local copy of AutoDep4 in order to complete the deposition and validation in-house before uploading the deposition session to the public site, thereby reducing time and effort.

The annotation of the deposited PDB entry involves the use of a large set of in-house programs and third-party software, which help to automate many tasks in the post-deposition pipeline. Issues relating to the annotation of depositions are beyond the scope of this chapter and have been described elsewhere.[12,13]

The PDB operates a weekly release schedule and new entries are constantly released to the public archive every Wednesday from a central staging site. Once these structures are released, they are loaded into the MSDSD and made accessible via various MSD services as described below. At the depositor's request, macromolecular structures deposited with the PDB can, however, remain "on hold" for a maximum of one year from the date of deposition.

26.3 The Electron Microscopy Databank (EMDB)

Cryo-electron microscopy (cryoEM) is a high-resolution imaging technique that is particularly appropriate for the structural determination of large macromolecular assemblies, which are difficult to study by X-ray crystallography or NMR spectroscopy. For some biological molecules that form two-dimensional crystals, the application of cryoEM and image reconstruction can help elucidate structures at atomic resolution. In instances where crystals cannot be formed, atomic-resolution information can be obtained by combining high-resolution structures of individual components determined by X-ray crystallography or NMR with image-derived reconstructions at moderate resolution. This can provide unique and crucial information on the mechanisms of these complexes. Image reconstructions may be used to augment X-ray studies by providing initial models that facilitate phasing crystals of large macromolecular machines such as ribosomes and viruses.

CryoEM methods are rapidly improving, and with the adoption of standards, the results of these studies need to be made accessible to a much larger community. To this end, the EMDB[14] has been established in the MSD Group at the EBI and has been accepting depositions since June 2002. The requirement to deposit protein structures in the form of maps determined by Electron Microscopy led to the implementation of a web-based deposition system, EMDep,[8] a search tool, EMSearch, and Atlas pages for each entry.

26.3.1 *Data Deposition*

We have designed a tool (EMDep) that provides a direct means of entering and validating the data while also facilitating the release of the submitted data in order to meet public demands. The system is simple to use and takes advantage of previously submitted information wherever possible. EMDep is a flexible and portable system (Table 26.1) that allows for the acceptance and validation of data by an interactive depositor-driven operation.

26.3.2 *Annotation*

Depositors complete the submission process only after the structure data has passed validation and they have reviewed the completed EMD entry and its companion report file, and accepted the EMDB release policy. EMDep returns an acknowledgment letter with the assigned EMD ID code by e-mail. The final version of the structural description is supplied to the depositors for review and approval. In this way, EMDep maximizes the usefulness and timeliness of the structure data produced by research scientists independent of any work by the EBI staff, thereby enabling deposition centers to keep up with an ever-increasing flow of data.

26.3.3 *EMSearch*

An EMDB search tool, EMSearch (Table 26.1), gives access to brief details of each entry, including sample name, author names, resolution of map, map submission, release dates, and a link to the atlas pages for each entry.

26.4 The MSD Relational Database

This section provides a brief introduction to databases, outlining their advantages over file-based systems, and includes an account of MSDSD, a relational database. The primary advantages of storing protein structures in a database are flexibility and speed, which provide a platform upon which different services can be readily built and *ad hoc* queries executed.

26.4.1 *Storing Data in Files*

Where data is stored in files, its management and manipulation is *par force* programmed into the application that accesses it. The amount of programming required for multi-user systems as opposed to single-user systems is much greater, as some form of record locking needs to be developed to prevent different users from updating the same data at the same time. Applications that use files to store data generally lack flexibility, as code has to be written by programmers (as opposed to end users) if new reports are required from the system. Therefore, whilst the use of files to store data may be workable for small, single-user systems with a specific well-defined purpose, any variation from this towards greater complexity, flexibility, or data volume makes the use of a ready-made database engine a better option.

26.4.2 *Storing Data in Databases*

The use of databases to store data offers critical advantages over the use of files as they generally have simple reporting languages such as Structured Query Language (SQL) and come with built-in capability to handle multiple users and data integrity. Unlike file systems, databases include features that can be used to enhance performance, such as indexes, partitioning, and parallel processing, all of which provide scope for data growth. With advancements in storage and CPU technology, databases consisting of tens of terabytes of data, possessing tables with hundreds of millions of rows, provide acceptable performance on low to mid-range servers. However, commercial database software is expensive and requires specialist skills to build, tune, and maintain. Free database software is an option, but these are generally not as scalable and are not supported as well as commercial packages.

The initial databases of the 1960s organized data in hierarchical upside-down tree structures, which meant they were fast but lacked flexibility when it came to reporting. In 1970, Codd[15] introduced a new concept, the relational model, in which data was stored in relations (tables without order). In his model, Codd provided rules for integrity constraints and operators for the manipulation of data. These relational operators form the basis of SQL, which is standard across different database platforms. Today, the vast majority of databases are relational. A workable database system must provide data integrity, data security, availability, and performance.

26.4.2.1 *Data integrity*

In databases with high data integrity, the data is accurate, up-to-date, and structured in a manner that they can be routinely retrieved, updated, and manipulated using a standard data manipulation language (DML). Without data integrity, a database is not usable. The structuring of data is carried out via a normalization process where they are taken from a raw, unstructured form to a normalized (structured) form. There are five stages of normalization: first normal form to fifth normal form. Data in fifth normal is highly structured. In practice, third normal form is usually good enough for effective database design and development. Normalization aims to produce a logical data model in which data is organized in an efficient manner and duplication is minimized. The model consists of entities, primary keys, foreign keys, and links depicting the relationship between entities, which may be one-to-one, one-to-many or many-to-many. An entity is anything about which data can be stored, whilst a primary key is a unique value used to identify a specific occurrence of an entity (a record). Primary keys of one entity used in other entities (where they are not unique) are foreign keys. When building a database, the logical model is used as a base to produce a physical model in which entities become tables and some carefully selected data redundancies (duplication of data) may be introduced to enhance the database's performance.

Primary keys, foreign keys, and triggers are used to enforce data integrity. An attribute (column) of a table defined as a primary key can only contain unique values. An attribute defined as a foreign key links values in a child table to a single primary key of a parent table. The aim of this is that no value can be assigned to a child table that does not already exist in a parent table. Triggers are small programs associated with individual tables that can be set to execute automatically when data in the table is inserted, modified, or deleted. In this way, checks can be made at the time of change to enforce business rules.

26.4.2.2 *Data security*

Data security concerns the prevention of unauthorized access to data. At the highest level, passwords control the login access. At lower levels, control is exercised by granting specific privileges to select, insert, update, or delete data from tables. However, despite these controls, damage to data may still be caused by human error. One way to protect against this eventuality is by the routine use of backup and recovery schemes. A damaged table or whole database can be retrieved to a point in time before the damage took place if a database is run in archive log mode.

26.4.2.3 *Availability*

Databases must be online and available. To test for their availability, scripts can be executed at regular intervals that report via email when a particular database becomes unavailable. There are two ways to address such a situation. First, a standby database running on a different server can be brought online when the primary database fails. Second, one can operate two copies of a database online and set up a fail-over mechanism redirecting the traffic to the available copy.

26.4.3 *The MSDSD Production Line*

Structures are updated in the PDB once a week and are loaded into a deposition database. The loading procedure also checks data quality, syntax, and enforces strict constraints upon the data being loaded by comparison with reference data. At regular weekly intervals, data for new depositions is copied to a transformation database where it is transformed from a structure suitable for accepting primary data (i.e. many small tables with a large number of integrity constraints) to one suitable for searching (i.e. a small number of large tables with a limited number of integrity constraints).

After transformation, the new data is loaded into a test database against which online services are tested. If the test database passes rigorous quality assurance tests, it is put into production. Experience shows that the quickest and safest way to do this is to delete the production database and replace it with the test database. To avoid loss of service during updates and to provide greater availability during normal use, two copies of the production databases (located on different servers) are used. During updates, connections to the database fail-over to the copy that is still on line.

Approximately once a year, a full release is produced by loading and transforming the protein structures from archived text files. Currently, the whole process, which loads approximately 44 000 entries, is completed within 20 days. Such releases are produced to implement major changes in the data warehouse and add new services or new features to existing services. It also gives sites, which have academic licenses to use MSDSD in-house, the capability to carry out a fresh installation with up-to-date data without having to carry out a high number of incremental updates. Multiple concurrent sessions are utilized to minimize loading time, but loading performance tails off if more that two sessions per CPU are executed. This requires a server that has a high number of virtual or real CPUs (upwards of 16).

26.4.4 *Characteristics of MSDSD*

The MSDSD contains approximately 150 tables and 600 indexes occupying 300 GB of space. The advantage of using a high number of indexes is that they speed up the retrieval of data from tables. The disadvantage is that they slow down the loading of data as they are updated during this process. The largest table in MSDSD occupies 80 GB and contains 425 million rows. The indexes in the warehouse take up the same amount of space as tables. Very large tables and their indexes are partitioned. This results in a greater retrieval speed as only the relevant partition of a table or index is searched. Tables and indexes above 1 GB are stored in their own space to help reduce fragmentation, which if left unchecked has a dramatic impact on performance. Indexes and tables that are accessed at the same time are stored on different disks to reduce contention and improve speed.

26.4.5 *MSDSD Data Architecture*

In this section, a simplified account of the architecture and general structure of data in MSDSD is presented.

Data on the 3D structure of proteins is hierarchical in nature. Assemblies of biological molecules are made up of many chains, which are made up of many residues, which in turn are composed of many atoms. In the logical model of a relational database, one to many relationships are used to set up hierarchical links. Thus, the primary key of an assembly table at a high level is linked to an attribute in a chain table on the next level down via a foreign key so that a one-assembly to many-chain relationship is enforced.

In MSDSD, the topmost level is represented by assemblies, each of which is the complete collection of associated chains (macromolecules and associated small molecules, including solvent). This represents a level higher than tertiary structure. Data on assemblies is stored in a table named *assembly* that contains attributes such as assembly_type (monomeric, dimeric, etc.), num_chains (number of chains in assembly), and num_xchain_ss (inter-chain disulphides). A full account of this and other tables that make up the logical model of MSDSD is available online (Table 26.1). The assembly table currently contains information on 83 000 assemblies.

At the level below assemblies, we have chains that are of three types: polymer, non-polymer, and water. The water category includes only water and methanol, while ordered small molecules such as sulphate ions or acetone are categorized as individual non-polymer entities. Data on chains in MSDSD is stored in a table named *chain* that contains information such as chain_type (polymer C, non-polymer B, water W), num_residues (number of residues), and pdb_code (original code of the chain in the PDB file). The chain table currently contains information on 662 000 chains, giving an average of eight chains per assembly.

At the level below chains, we have residues, there being many residues per chain. Data is stored in a table named *residue* that contains information such as chain_pdb_code (code for chain to which residue belongs), chem_comp_code (standard extended molecular code of the amino acid or ligand), and residue_type (R:residue, B:bound molecule, W:water). The residue table currently contains information on 52 000 000 residues, giving an average of 79 residues per chain.

Information on the 3D coordinates of atoms is stored on the lowest level in a table named *atom_data* that currently contains 425 000 000 rows.

In order to provide users from different scientific disciplines access to data in the MSDSD relevant to their interests, the MSDSD is organized into sections referred to as data marts. Some of these have a central role in the database, whilst others are decoupled and may be used in a specialist manner. A few of these data marts are described below.

26.4.5.1 *Ligands*

Ligands are defined as small molecular entities that associate with proteins and either occur naturally (such as ATP) or not (such as drugs and inhibitors). The MSD group at EBI has built and actively maintains a catalogue of ligands that is used as reference data by other marts within MSDSD. The online service MSDchem provides access to this information that includes data on every small molecule in the PDB in the form of: atoms elements, standard nomenclature, connectivity, bond orders, aromaticity, and stereochemistry.

26.4.5.2 *Structure*

This is the core of the PDB data extended to provide coordinates of quaternary structures derived from deposited data. This section is organized in three different interrelated hierarchies that facilitate different points-of-view:

(i) The sequence point-of-view. The information in this hierarchy concerns the sequence and chemistry of the protein. Every macromolecule corresponds to a protein sequence, but it is possible to have more than one instance of this molecule in the PDB as an asymmetric unit (with slightly different overall conformations as these that were observed in the experiment). On the atomic level, constituent atoms model the abstract notion of a chemical atom that ignores alternative configurations or different NMR models.

(ii) The asymmetric unit point-of-view. The observed structure, as available in the PDB entry, describes only the asymmetric unit of a crystal. The contents of the asymmetric unit are also reused to create quaternary structures but are marked with a special non-symmetric-valid flag.

(iii) The assembly point-of-view. This corresponds to the actual quaternary assembly as derived from the deposited structure. This is usually the closest available model of the actual structure of the protein in solution and provides a complete understanding of inter-chain and ligand interactions often not represented in the PDB files. Some features of protein structures are apparent only in the quaternary assembly and could be easily missed by researchers examining the PDB file. All solvent and bound molecules are defined in separate chains and are associated with the protein chains they are closest to. During the process of assembly formation, bound molecules and waters may be replicated several times, as long as they have some form of interaction with the assembly.

26.4.5.3 *Secondary structure*

This section of the database stores detailed information about the secondary structure, including sheets and helices as well as more extended formations like bulges, hairpins, and motifs. Since secondary structure is not always available in PDB entries and/or its source or accuracy may be questionable, this information has been re-derived for all PDB entries with coordinates of the predicted quaternary structure using DOSS, a secondary structure prediction program based on DSSP[16] and ProMotif.[17]

This provides a consistent platform for comparisons and analysis of secondary structure. By using the predicted quaternary structure (the assembly) it is possible to identify secondary structure elements related to more than one chain in the assembly.

26.4.5.4 *Active sites*

This section of the database stores details of the binding sites of all small molecules found in complexes with macromolecules. Since this information is not always reliably represented in PDB files, site information has been calculated internally and forms the basis for the MSDsite service[18] (Table 26.1).

The binding sites of a macromolecule are determined based on the contacts it makes with a ligand. Contacts are defined based on the different types of bonds and interactions that take into account the distance and angles of the atoms, as well as other characteristics of the ligands and residues such as planes, etc.

26.4.5.5 *External cross-references/taxonomy*

Much effort has been devoted to providing complete and consistent cross-references with external databases like UniProtKB, SCOP, CATH, EC Enzyme, Gene ontology, Medline, and NCBI taxonomy databases. The cross-references are established not only on a residue level with UniProtKB, but also aggregated to facilitate data analysis on a higher level, and is described under the SIFTS initiative[19] (Table 26.1).

26.4.6 *MSDSD Distribution*

Apart from serving as the cornerstone for the MSD search systems and services, the MSDSD is also available for distribution in a number of different ways.

(i) Oracle replication*:* This is the only type of distribution for which we offer frequent (weekly) increments for users that wish to follow closely the PDB release cycle. This option requires an Oracle server license, database administration support, and adequate hardware infrastructure. Typically, a user of this replication will download and install the latest full release (full transformation) of MSDSD using the full installation instructions available from the MSD FTP site. Such full releases take place on a yearly basis, and this is the time of MSDSD reconciliation, since all PDB entries are refreshed and creeping inconsistencies are resolved.

Between full releases, the user may run the automatic synchronization script that will allow the download and inclusion of increments for the new PDB entries, released every week. Any corrections in reference data will not propagate back to the affected old entries in order to keep the increments manageable, so the only time that the full set of MSDSD relational constraints is guaranteed, is only immediately after a full release.

The MSDSD and the incremental updates are organized in sections (marts) so users are free to install and increment just the marts that they are interested in. There is also the option to specify which tables of a mart a user wishes to have installed, so users may in general replicate just a few individual tables.

(ii) Replication on MySQL: This distribution requires basic Linux administration support and is adequate for researchers and students with limited resources and technical support. Users download and install directly the MySQL data-files of the tables they are interested in from our FTP server following the MySQL installation instructions available from the MSD FTP site. The tables are available in compressed myIsam format without any pre-built indexes.

26.5 MSD Search and Analysis Services

The MSD group has worked to create many search and analysis services to cater to different categories of end-users. By definition, a search system that provides access to all aspects of biological data (structure, sequence, active site, published abstracts) must be accessible and understood by the novice scientist since biological data is highly diverse and complex. All search and analysis services offered by the MSD are either based on the data available inside the MSD relational database (MSDSD) or that are used to derive and populate information inside the database. Multiple search interfaces have been generated to try and provide access both for the expert user and novice scientist without making the data too complex to understand or require the user to be an expert database programmer. A few MSD search and analysis tools are described below.

26.5.1 *MSDlite*

MSDlite (Table 26.1) allows users to query MSDSD using a list of fields that include commonly used terms such as title, author, keywords, and general text searches. The capability to cross-reference with other data sources such as the NCBI taxonomy, Gene Ontology, Interpro, UniProtKB, and the Enzyme database through their respective IDs and accession numbers is also provided, as well the facility to select fields of output.

26.5.2 *MSDpro*

MSDpro (Table 26.1) is a more powerful search interface than MSDlite designed for experienced structural biologists. It includes a drag-and-drop query builder and the ability to save queries.

26.5.3 *MSDchem*

The wwPDB maintains a separate data resource: the ligand dictionary, which is the chemical reference database of all chemical entities in the Protein Data Bank. The MSD group has extended this ligand dictionary by utilizing chemoinformatics packages and incorporating additional annotation. This information has been loaded into a relational database, which is publicly available on the web through the MSDchem (Table 26.1) search system. The MSDchem search web service offers various options for searching the MSD ligand database and exporting data in order to serve different user requirements. Search options include:

- Searching for ligands using their unique PDB three-letter code, name, or formula range.
- Searching for ligands using outline or chemical sketches (using sub-graph searches) to identify molecule variants.
- Searching by fingerprint similarity to find ligands that have similar localized chemistry.

Users also have the option to use any combination of the above types of searches. On the result pages, they can examine, visualize, and export ligands or refer back to the relevant PDB entries that contain the ligand searched. Additionally, the MSDchem database is available for export in various formats, from a ready-to-use relational database, for *ad hoc* querying, to collections of commonly used chemical data files or chemical names and descriptor lists.

26.5.4 *MSDfold*

MSDfold (Table 26.1)[20] is a service for the rapid and accurate comparison/alignment of protein structures. The primary use of MSDfold is to be able to identify common structural motifs in a family of structures, often a starting point for most modeling and structure-based studies. MSDfold allows for the alignment of protein structures (identified by PDB/SCOP codes or uploaded as PDB/mmCIF files from a user's desktop), as well as for structural searches in the PDB and SCOP data sets. A user-defined data set may be uploaded as a tar archive with additional files specifying pairs of structures to be aligned. The calculations are distributed on automatically chosen number of parallel CPUs for faster performance.

On output, MSDfold returns a list of structural hits, ranged by one of available scores: Q-score, P-value, RMSD, alignment length, sequence identity, and size of common Secondary Structural Element (SSE) motifs. Each entry in the list of results allows for in-depth investigation on a residue level, where detail alignment of SSEs and residues are presented, together with matrices of best structure superposition, visualization, and download facilities. For visualization, either a client-installed Rasmol software[21] or server-supported Jmol applet (Table 26.1) may be used. MSDfold is also run on all entries inside the MSDSD in order to perform an all-against-all comparison of all structures inside the database. These results are in turn stored in the database and are accessible to MSD services.

26.5.5 *MSDPISA*

Physiological function of most proteins is closely associated with their ability to aggregate and is not independent from the context of macromolecular assembly. However, the final result of a crystallographic experiment provides coordinates for the crystallographic asymmetric unit (ASU), rather than the macromolecular assembly. Given the nature of the experiment, no link between the ASU and biological unit (macromolecular assembly) in PDB can be postulated. Assemblies may be formed by the contents of a few ASUs of the crystal, or a few complexes can be found in a single ASU, or a complex may be made by several parts of neighboring ASUs. As a result, one can infer on macromolecular complexes only if additional biological, structural, or functional information is provided. As a matter of fact, few PDB entries come with an experimentally verified oligomeric state.

MSD offers a tool, called MSDPISA (Protein Interfaces, Surfaces, and Assemblies) (Table 26.1)[22] that permit the reconstruction of macromolecular complexes from crystal data and analyses their properties. MSDPISA is implemented as a web-server, which includes a searchable database of pre-calculated results for all PDB structures determined by X-ray crystallography. It also allows for the upload of PDB and mmCIF[23] coordinate files for interactive processing. The calculations are distributed over a variable number of CPU nodes (1.2 GHz Pentium-4), depending on the estimated task complexity. Typically, the calculation results are returned in less than 30 seconds, while the most difficult cases may take up to 20 minutes. The server also provides a detail description of interfaces, structures and their assemblies, visualization and database search tools. For macromolecular interfaces, PISA[22] calculates the buried surface area, solvation energy effect, hydrogen, covalent and disulfide bonds, salt bridges, and statistical significance of interfaces.

MSDPISA also offers a number of database search options. For example, one may request a list of all macromolecular interfaces in PDB that are structurally equivalent to a selected interface in analyzed structure. Other options include querying PDB on particular assembly size, symmetry number, space group, range of surface/buried areas and dissociation energies, particular keywords, presence of specified ligands (small molecules), assembly type (homo-/hetero-), presence of salt bridges and disulfide bonds or any arbitrary combination of these factors.

26.5.6 *MSDanalysis*

MSDanalysis (Table 26.1) is an interface that allows statistical analysis of the data contained in the MSDSD. It allows users to perform statistical analysis of molecule and residue-based information and to select subsets of data based on correlations and multiple filters. Users can then download data selected by these filters and correlations. Using this service, a user can get answers to questions on a database-wide scale. For example, MSDanalysis could be used to correlate quaternary structure predictions with crystallographic symmetry; analyze the effect of data resolution on data quality, etc. Finally, there is a database browser interface (via SQL) that is part of MSDanalysis, where a user can issue their own SQL queries directly to the MSDSD. MSDanalysis also includes a structure validation tool that can be used to analyze structure quality within the MSDSD based on geometric criteria. The user may also choose to upload a structure in PDB format for similar validation analysis. This service is expected to be of primary interest to structural biologists.

26.5.7 *MSDtemplate*

MSDtemplate (Table 26.1) is a service based on residue groups, such as active sites, ligand environment, or any set of amino acids in particular spatial juxtaposition. The MSD has a library of templates generated by mining the PDB and looking for statistically significant collections of residues. These templates include all the known active sites and metal binding sites. The service is based on the matching of atoms in a template with those of a protein using weighted fuzzy superposition. A number of controls are available to adjust the superposition process and to fine-tune the analysis. MSDtemplate also includes search engines to identify templates in protein structures and the necessary tools to visualize these. Further analysis tools and a new and more extensive library of mined motifs are planned to be released soon.

26.5.8 *MSDmine*

MSDmine (Table 26.1) is a comprehensive mart-based data mining system that allows users to perform complex searches using any of the 90 tables available in MSDSD organized into eight different marts.

26.5.9 *MSDmotif*

MSDmotif (Table 26.1)[18] is a tool to provide insight into the PDB with respect to motifs in protein 3D structures, protein sequences, small bound molecules, ligand 3D environment, protein-protein and nucleic-acid interactions. It provides sequence and 3D structure annotations with PROSITE motifs,[24] secondary structure elements, 3D small motifs, binding and catalytic sites available in ePDB XML, eFamily XML, and DAS formats.[25]

26.6 The Future Outlook

The MSD group has worked towards the development of a fully integrated system for macromolecular structure data by the design, implementation, and use of a relational database. The MSD group continues to work closely with its partners, to further enhance the quality and consistency of the data in the database. MSD services designed on top of this database provide various access points to the data contained within, thereby catering to the scientific community consisting not only of experienced structural biologists, but also to chemists, molecular biologists, and other occasional users interested in macromolecular structures. In many ways, the MSDSD project defines a database standard for storing, managing, and distributing macromolecular structure data.

With the predicted explosion of structures determined by experimental methods such as X-ray diffraction and cryoEM, it has become even more imperative to have robust technologies in place that are not reliant on flat-file archives for data storage and analysis. The MSDSD has proved itself to be robust, fast, and capable of handling large amounts of data. The planned merger of data from the EMDB into the MSDSD in the near future is a step in the direction of creating a single resource for all experimentally determined structures in a relational database format. The MSDSD has been successfully distributed to external sites, and as part of the distribution package, a comprehensive application programming interface (API) has been developed, which allows users to extract data from the database using custom programs.

The MSD search systems and the underlying relational database continue to improve, with new features and capabilities being added to many services, moving us ever closer to our ultimate goal of becoming a comprehensive, integrated resource for the research community.

References

1. Hamm GH, Cameron GN. (1986) The EMBL data library. *Nucl Acids Res* **14**: 5–9.
2. Apweiler R, Bairoch A, Wu CH, *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucl Acids Res* **32**: D115–119.
3. Boutselakis H, Dimitropoulos D, Fillon J, *et al.* (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucl Acids Res* **31**: 458–462.
4. Golovin A, Oldfield TJ, Tate JG, *et al.* (2004) E-MSD: an integrated data resource for bioinformatics. *Nucl Acids Res* **32**: 211–216.

5. Berman HM, Henrick K, Nakamura H. (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* **10**: 980.
6. Berman HM, Henrick K, Nakamura H, Markley JL. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucl Acids Res* **35**: D301–303.
7. Berman HM, Battistuz T, Bhat TN, *et al.* (2002) The Protein Data Bank. *Acta Crystallogr D* **58**: 899–907.
8. Henrick, K, Newman R, Tagari M, Chagoyen M. (2003) EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information. *J Struct Biol* **144**: 228–237.
9. Service RF. (2000) Structural genomics offers high-speed look at proteins. *Science* **287**: 1954–1956.
10. Tagari M, Tate J, Swaminathan GJ, *et al.* (2006) E-MSD: improving data deposition and structure quality. *Nucl Acids Res* **34**: D287–290.
11. Westbrook J, Feng Z, Burkhardt K, Berman HM. (2003) *Meth Enzymol* **374**: 370–385.
12. Swaminathan GJ, Tate J, Newman R, *et al.* (2004) Issues in the annotation of protein structures. In AM Lesk (ed.), *Database Annotation in Molecular Biology*, John Wiley & Sons, New York.
13. Dutta S, Burkhardt K, Swaminathan GJ, *et al.* (2007) Data deposition and annotation at the worldwide Protein Data Bank. **In press**.
14. Tagari M, Newman R, Chagoyen M, Carazo JM, Henrick K. (2002) New electron microscopy database and deposition system. *Trends Biochem Sci* **27**: 589.
15. Codd EF. (1970) *A Relational Model of Data for Large Shared Data Banks. Commun ACM* **13**: 377–387.
16. Kabsch W, Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**: 2577–2637.
17. Hutchinson EG, Thornton JM. (1996) PROMOTIF — a program to identify and analyze structural motifs in proteins. *Protein Sci* **5**: 212–220.
18. Golovin A, Dimitropoulos D, Oldfield T, Rachedi A, Henrick K. (2005) MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins* **58**: 190–199.
19. Velankar S, McNeil P, Mittard-Runte V, *et al.* (2005) E-MSD: an integrated data resource for bioinformatics. *Nucl Acids Res* **33**: D262–265.
20. Krissinel E, Henrick K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three-dimensions. *Acta Crystallogr D* **60**: 2256–2268.
21. Sayle RA, Milner-White EJ. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem Sci* **20**: 374.
22. Krissinel E, Henrick K. (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol*, **in Press.**

23. Bourne PE, Berman HM, McMahon B, Watenpaugh J, Westbrook, Fitzgerald PMD. (1997) The Macromolecular Crystallographic Information File (mmCIF). *Meth Enzymol* **277**: 571–590.
24. Hoffmann K, Bucher P, Falquet L, Bairoch A. (1999) The PROSITE database, its status in 1999. *Nucl Acids Res* **27**: 215–219.
25. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L. (2001). The distributed annotation system. *BMC Bioinform* **2**: 7.

This page intentionally left blank

Chapter 27

Molecular Graphics in Structural Biology

A. M. Lesk*,†, H. J. Bernstein‡ and F. C. Bernstein§

27.1 Introduction

We can observe macroscopic biological systems directly with our senses. We can take photographs, draw sketches, and preserve the actual objects. A microscope with stains and fixatives gives us similar capabilities in dealing with smaller objects on the scale of cells and nuclei. Electron micrographs allow direct observation, almost, but not quite, to atomic resolution. For the final step into molecular biology at atomic resolution in the realm of structural biology, we need to create physical molecular models and hand- or computer-drawn renderings of idealized and abstracted molecular models. Biological macromolecules are inaccessible to our direct sensory observation because they are small. They are challenging to represent because they are complex.

Over the last half century, starting with the seminal work of Levinthal and Katz,[1] computer-based interactive molecular graphics

*Corresponding author.

†Department of Biochemistry and Molecular Biology and the Huck Institute for Genomics, Proteomics, and Bioinformatics. The Pennsylvania State University. University Park, PA 16802, USA. Email: aml2@mrc-lmb.cam.ac.uk.

‡Department of Mathematics and Computer Science. Dowling College, Oakdale, NY 11769, USA.

§Bernstein + Sons. Bellport, NY 11713, USA.

has become as essential a component of the toolkit of contemporary molecular biology as the centrifuge. It is our gateway to the data, most obviously of three-dimensional structures of nucleic acids, proteins, and other biological molecules, but also all other types of data that need to, and can, be presented pictorially.

Molecular graphics provides a notation for the understanding of chemistry. The development of that notation predates the development of the hardware we now use for the purpose. With the growth in number and variety of structures, and the need for comparative analysis of homologs — including the essential tool of superposition — physical models are inadequate. Nevertheless, like nostalgia buffs, we simulate wireframe, wooden, and plastic model-building kits on computer screens. But computers lend speed and accuracy absolutely essential to the progress of structural biology. Tasks that took months of patient effort by hand half a century ago now take seconds on a computer.

Computer graphics is now a mature science. Several decades of experience have produced effective and easy-to-use software. Driven by the entertainment industry, hardware has become readily available and inexpensive. The rendering of molecules has moved from specialized graphics systems to ordinary personal computers.

Traditional molecular models focused either on the skeletal pattern of bonds or on atoms. That is, they focus either on the connectivity of the structure or on the overall shape and packing. One of the most famous scientific photographs of the last century, showing Watson and Crick and their model of the double helix of DNA (http://www.chemheritage.org/classroom/chemach/pharmaceuticals/watson-crick.html), shows the skeletal representation, emphasizing bonds. The making and breaking of bonds, especially hydrogen bonds, are essential to life. Emphasizing the atoms, Corey, Pauling and Koltun designed space-filling models with atoms represented as segments of spheres, with connectors serving as chemical bonds.[2] Simulations of such bond-oriented wireframe models and atom-electron-density-oriented space-filling models are very much the norm in modern computer-based molecular graphics.

To help in understanding the role of computer-based molecular graphics in modern structural biology, we first review the evolution of graphical notations used in chemistry and biology, working on ever finer scales from gross morphology down to atomic resolution models.

27.2 History

Structural biology has its roots in both biology and chemistry, two sciences that have long depended on visual comprehension.

27.2.1 *Biology: Taxonomy by Morphology to Molecular Biology*

In general, different species of flora and fauna display significant differences in size, shape, and other structural features. Until the rise of molecular biology, the best available way to classify organisms into species was to invert this observation and create a taxonomy based on morphology, distinguishing species on the basis of morphological features and inferring relationships from similarities. The concept was brought to full flowering by Linnaeus for plants[3] building on the earlier work of Ray on both plants and animals.[4] This visual morphological perspective enabled the seminal works of Darwin[5] and Mendel[6] and the beginnings of modern evolutionary biology and genetics, which found a rigorous foundation in molecular structural biology from the work of Franklin and Gosling,[7] Watson and Crick,[8] and Kendrew and Perutz.[9] This visual approach to biology persists to this day in efforts to infer macromolecular function from structure and to design drugs by a visual structural understanding of active sites based on lock-and-key models.

It is now well known that species differ not only in the structure of the body, but also at the level of the structures of homologous proteins. What is not so widely known is that the first clues to the evolutionary divergence of protein structure are a century old.

In what is arguably the most premature scientific result of all time, in 1909 Reichert and Brown[10] published a study of crystals of hemoglobin isolated from different species of animals. Hemoglobin crystallography, three years before the discovery of X-ray diffraction, was limited to measuring the angles between crystal faces. Stenö's law (1669)[11] states that the interfacial angles of all crystals of a substance are the same, independent of the size and macroscopic shape of the crystal. Therefore, these angles characterize the substance. Reichert and Brown found that the patterns of divergence of these angles correlated with the evolutionary tree of the species. They even found differences between crystals of deoxy- and oxy-hemoglobin.

We can now interpret and appreciate these observations. The formation of crystals implies that the molecules can take up a definite structure, with the ability to pack themselves into regular arrays. The differences in interfacial angles imply that crystals of hemoglobins from different species have different structures. The correlation of the divergence patterns of the crystals and the species implies that evolution is shaping molecules as well as bodies, in parallel processes. The differences between crystals of deoxy- and oxy-hemoglobin imply that the protein undergoes a conformational change upon binding oxygen.

Fifty years later, Perutz announced the solution of the X-ray crystal structure of hemoglobin.

27.2.2 *Chemistry: Atoms to Bonds to Quantum Mechanics*

During the course of the nineteenth century, chemistry made a transition from the one dimension of formulas that indicated only a compound's atomic composition to two-dimensional bonding diagrams, and then to three-dimensional structures such as the tetrahedral carbon atom introduced by Paternò,[12] van't Hoff,[13] and Le Bel.[14] van't Hoff distributed molecular models with his original publication, as "supplementary material."

The impetus towards three dimensions had three main sources: crystallography, the relation of structure to optical activity, and the rationalization of isomers. Why does propyl alcohol have only two isomers? Why do monosubstituted benzenes have only one?

This transition in chemistry to three dimensions began with the late-eighteenth century observation by Haüy[15] that crystals could be regarded as formed from repeating identical microscopic units. Pasteur's separation, in 1847, of racemic tartaric acid by manual selection of crystals of different shape, and his demonstration that the mirror-image crystals had opposite optical activity, linked crystallography with spectroscopy.[16] In 1860, Pasteur published the observation that the mold *Penicillium glaucum* preferentially metabolizes one enantiomorph of tartaric acid.[17] This brought the idea of three-dimensional molecular structure into biology, to be developed strongly by Fischer in his studies of sugar structures and of enzyme-substrate interactions and specificities. Fischer's important, intrinsically three-dimensional "lock-and-key" hypothesis of 1894[18] remains an essential conceptual basis of modern drug design.

At the turn of the twentieth century, although chemists had good insight into the three-dimensional architecture of molecules, the structure of the atom itself remained a mystery. Indeed, even the idea that matter consisted of particles was not universally accepted until Perrin's work on Brownian motion.[19] The discovery of quantum mechanics and explicit formulas for atomic orbitals implied the importance of portraying their symmetry properties to rationalize the data of atomic spectroscopy and of chemical bonding. Pictures of atomic orbitals appear in an early article by White in 1931.[20]

The theory of atomic structure had immense impact on chemistry as well as on physics. Pauling's theory of hybridization rationalized the relationship between the symmetries of the molecular framework and the distribution of the electrons. In general, the picture of molecular orbitals as Linear Combination of Atomic Orbitals (LCAO), although an approximation, has been extremely valuable for chemists, as has the Hartree-Fock approximation, the most precise possible description of the electronic structure of atoms and molecules that retains the orbital picture. Several albums of atomic and molecular orbitals have been published on paper; many programs now have facilities for generating them.

27.2.3 *Choice of Representations*

The provision, by X-ray diffraction and NMR spectroscopy, of experimental structure determinations of proteins and nucleic acids in atomic detail creates the challenge of representing the results in an intelligible way. The problem is not that computers are not powerful enough to draw complete and detailed representations of large proteins, it is that we are swamped by the complexity thereby revealed.

Should we focus on the bonds as in Watson and Crick's DNA model or on a molecular surface created as in the Corey, Pauling, Kolton space-filling models? Computers are able to draw both types of representations.

For macromolecular structures, however, concentrating on the bonds and drawing the entire molecule atom-by-atom as a wireframe model provides an overwhelming amount of detail. Space-filling models are slightly less complex, because the interior structure is obscured by the van der Waals surface, but they are still too complex. Feldmann at NIH developed a computer system for drawing shaded spheres in color. He and Bing published a set of slides, "Teaching Aids for Macromolecular Structure (TAMS),"[21] which contained a set of stereo pairs on 35mm slides and a simple viewer. These showed both the strengths and weaknesses of the technique. Richards presented "cheese-wire" slices through proteins, showing atoms contoured at their van der Waals radii.[22] These showed the dense packing of protein interiors. It became clear that an arsenal of different representations for different purposes was needed to illustrate and analyze protein structures.

27.2.4 *Molecular Graphics by Artists*

Intelligible representations were required, and these first appeared in the works of artists and illustrators.

27.2.4.1 *Scientific renderings: Irving Geis*

Irving Geis was an artist creating illustrations for Scientific American when the structure of myoglobin was determined. His hand-drawn illustration of the myoglobin structure was the first of many that he drew. These illustrations set a standard of graphical notation for the field. His book with Dickerson[23] was a classic, introducing a generation of scientists to protein structure. Geis' pictures show molecules in their entirety: all bonds, including, in a myoglobin picture,[24] hydrogen atoms and α-helical hydrogen bonds, using strong depth cueing to help the eye sort out the structural relationships.

27.2.4.2 *Museum artists: Salvador Dali, Ben Shahn*

Two "museum artists" whose work includes pictures with molecular themes are Salvador Dali and Ben Shahn. Dali called one of his periods "The Molecular Dali." Shahn's work includes the Lute and Molecule series (an example of which includes a Patterson function), and a fairly representational drawing of an α-helix.

27.2.4.3 *Simplified representations or cartoons*

Traditional representations are not adequate for illustrating and analyzing complex structures. What is necessary is a simplified representation that still retains important features of the molecule: a schematic diagram or cartoon. This representation was developed by Rossmann[25] and by Furugren.[26] Helices can be depicted as cylinders, and strands of β-sheet by thick arrows. Alternatively, McLachlan's wide ribbon tracing the backbone can be drawn with different widths emphasizing the regions of secondary structure and de-emphasizing the loops (e.g. the illustrations in Ref. 27). Following the lead of these innovators, many other people have drawn such pictures by hand, with those of Richardson being the most widely known.[28]

Drawings by illustrators have certain disadvantages, including the difficulties of making stereo diagrams, and more generally, in changing the orientation. For these reasons, Lesk and Hardman wrote a computer program to generate schematic diagrams of proteins,[29] and Lesk and Lesk extended this program to nucleic acids and protein-nucleic acid complexes.[30] It produced two alternative formats of output: line drawings and shaded-surface drawings. When the program was first written, interactive real-time computer displays could display only vector graphics. Shaded-surface drawings, or raster graphics, were limited to static presentations. At that time, one of us [AML] gave a talk entitled, "The line will decline when the raster gets faster," and that has in fact happened.

Since that work, a large number of programs have been developed and distributed. Improvements in the power of hardware and systems software have made it quite easy to write them. Now, computer-generated pictures of large biological molecular structures are part of the scientist's everyday toolkit. Figure 27.1 shows an example of a

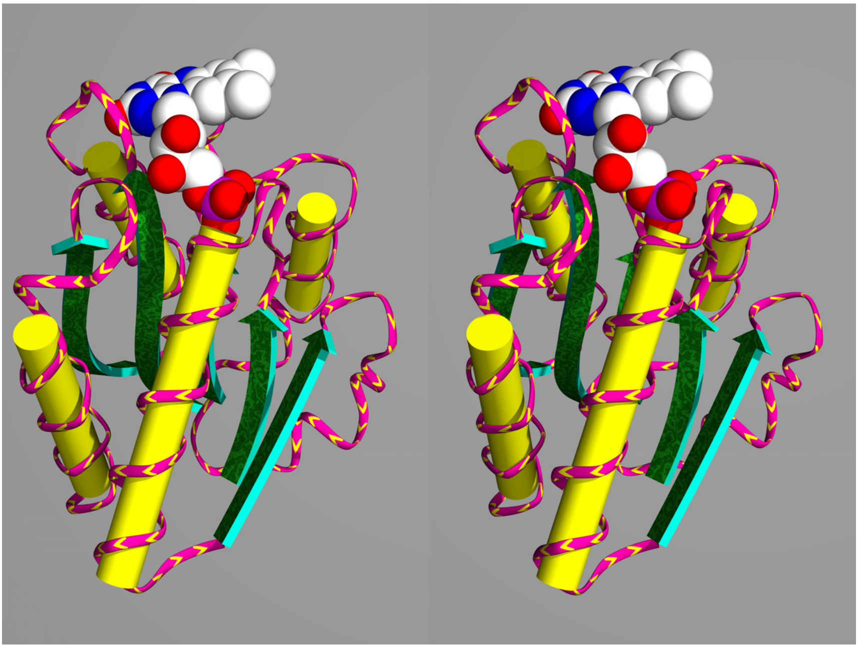


Fig. 27.1 Cross-eyed stereo view of schematic diagram of 5nll *Clostridium beijerinckii* flavodoxin.[79] Use of a general-purpose renderer allows the versatile representation of main-chain trace, schematic representation of helices as cylinders, and more detailed CPK representation of the ligand, flavin mononucleotide. Chevrons show the direction of the chain.

molecule rendered using shapes and surface textures to distinguish and highlight important features for publication. Many molecular graphics systems can support interactive real-time rotation of such images.

27.2.5 *History of Computer-based Tools*

The first interactive computer graphics system was created at MIT approximately 50 years ago. The pioneer in application to chemistry and molecular biology was Levinthal, who created the first interactive molecular graphics software at MIT and then established a group at Columbia University to pursue the development and applications of molecular modeling software. The technique rapidly proved its value and spread widely throughout the protein modeling community. First Adage, then Evans & Sutherland, and, later, Silicon Graphics, provided the most common hardware. With the development of inexpensive hardware, high-performance graphics is standard equipment in contemporary PCs.

27.2.6 *Contours and Surfaces*

Molecular graphics can produce pictures to represent different degrees of detail, using different abstractions and models. Individual semi-classical atoms and quantum-mechanical orbitals can be represented as textured surfaces, as if the viewer were looking at a topographic map, or as contour lines, as if the viewer were looking at level lines in such a map. Such plots help us in understanding essential characteristics of molecules: symmetry, size, shape, and orientation.

Fundamentally, every chemical structure — from the hydrogen atom to the ribosome and beyond — is an assembly of nuclei and electrons. Naturally, the larger the system, the more difficult for both the determination and the presentation of the finer details.

For large molecules such as proteins, it is neither possible nor in most cases desirable to work with a description more detailed than the overall layout of the structure in terms of the positions of the atoms, and some indication of their individual excursions. In addition, the primary experimental data in X-ray crystallography are not the result of an instantaneous snapshot of the true electron density, but an integrated average over time and over many uncorrelated samples of what may not be a homogeneous sample. The effective scattering from electron density is very different from both a quantum-mechanical model and from the hard-sphere atom model. To an approximation, one may think of the image of an atom, as reflected in X-ray diffraction data, as the internal electron distribution in the atom convoluted with the variation in atomic position, from thermal motion and/or disorder.

We treat the integrated diffraction intensities as if the X-rays were scattered from an averaged electron density. Whether in the form of a Patterson function, from unphased data (peaks in the Patterson function correspond to interatomic vectors), or a Fourier map of phased data, estimates of values of electron density are very helpful and usually critical in solving structures by X-ray crystallography.[31]

The averaged electron density is a function, ρ, of three spatial dimensions within a unit cell. If that density were sharply bounded, the boundary of its volume could be thought of as the hard-sphere model molecular surface. The density is not confined to a finite volume of space, but we can approximate a surface by looking at the boundary of a volume containing some substantial fraction of the integrated density, estimated by finding the boundary of the volume within which values of ρ are greater than some pre-determined value. Fortunately, when working with an ensemble average of densities at the normal resolutions for macromolecular work, electron densities can be well-approximated as smoothly varying sums of Gaussian distributions, centered at or near the atoms. Visualizing that information is a problem similar to that of understanding pressure, temperature, and humidity in the atmosphere, or the elevations of mountains. In the late eighteenth century, the problem of finding level lines was called "leveling."[32] Now, we speak of "contouring" and finding "isosurfaces."

27.2.6.1 *Applying the techniques of topographic mapping and weather mapping*

When we draw information on paper or on a screen we only have two spatial dimensions (the height and width of the paper or screen) on which to present data about electron densities, charges, and other properties of a molecule. The information we need to convey is indexed in three spatial dimensions, and perhaps, the fourth dimension of time. We can either cut slices (sections) through the density to get down to two dimensions and then find level lines of density to contour, or we can find a single isosurface on which the density is constant and do a two-dimensional rendering of that three-dimensional object. The case of dealing with a slice of density is much the same as for dealing with terrain elevation maps and surface pressure maps. For such maps, points on the surface of the earth that have the same elevation or the same barometric pressure are connected to form level lines. The case of dealing with an isosurface is much the same as finding the surface of a given constant pressure altitude along which to fly an airplane.

Two-dimensional slices give us more detail about the density, such as internal voids and cavities. Isosurfaces give us a more complete sense of the three-dimensional structure of the molecule.

Let's consider the case of a slice of density. We could just list the array of numbers giving the density, but adding contour lines connecting points of the same values is a more effective presentation, as shown in Fig. 27.2.

An isosurface can be presented as an opaque surface, but for model-fitting it can be more useful to use a transparent surface. Until computers became capable of representing a very wide range of colors and intensities, the most effective technique was to present the surface as a mesh of lines (an isomesh, see Fig. 27.3), or as a sparse set of dots, through which a fitted skeletal model is visible. An isomesh can be created by computing contours in sets of planes with normals in three independent directions (parallel to cell edges or to *x*, *y*, and *z*),[33] or by breaking up the surface into triangles or other polygons (a tessellation). A usable dot surface is harder to create on a computer. An artist can use dots to create an impression of a three-dimensional surface. Raphael, in his 1504 painting, "Madonna and Child Enthroned with Saints" (now in the Metropolitan Museum of Art, New York), presented the 3D shape of the Virgin's gown by the dot-surface technique: yellow dots on an opaque black surface.

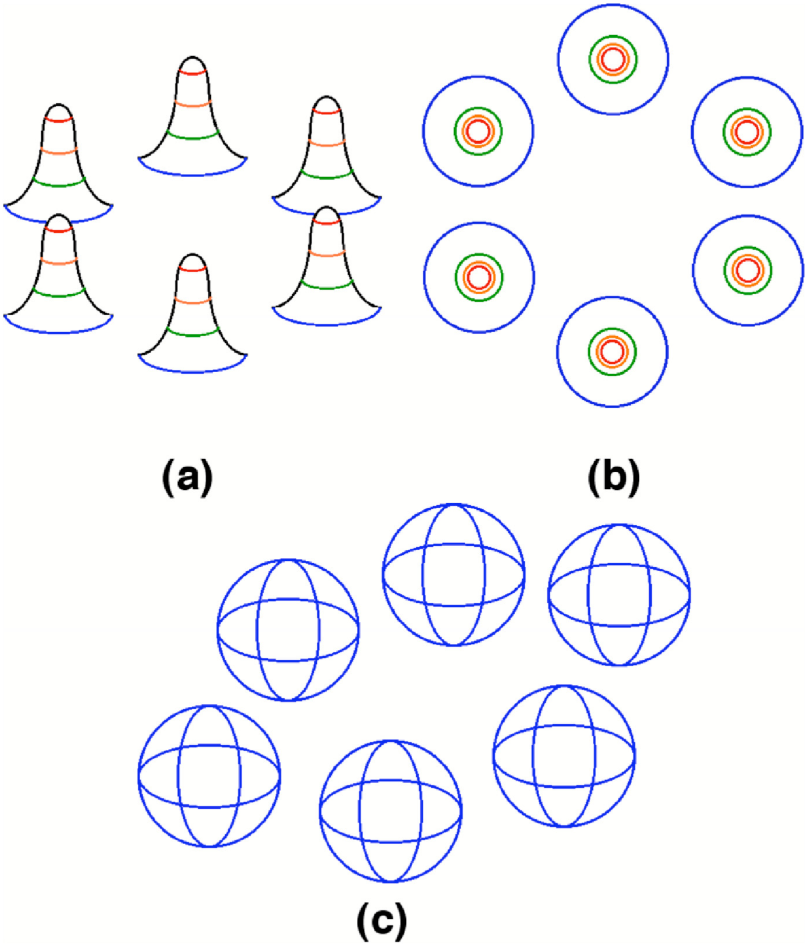


Fig. 27.2 **(a)** Gaussians for densities in a two-dimensional slice through a six-membered ring with level lines only down to the lowest atomic core densities (blue), but not down to the density in the bonds. **(b)** Contours for the same Gaussians. **(c)** Isomesh at the lowest level of the contours (blue).

Sparse dots by themselves create an impression of a transparent surface, but it is surprisingly difficult to use a computer to create a true three-dimensional dot surface that does not contain distracting surface pattern artifacts and textures in at least some orientations. Andrews created one of the first successful pattern-free dot surface algorithms

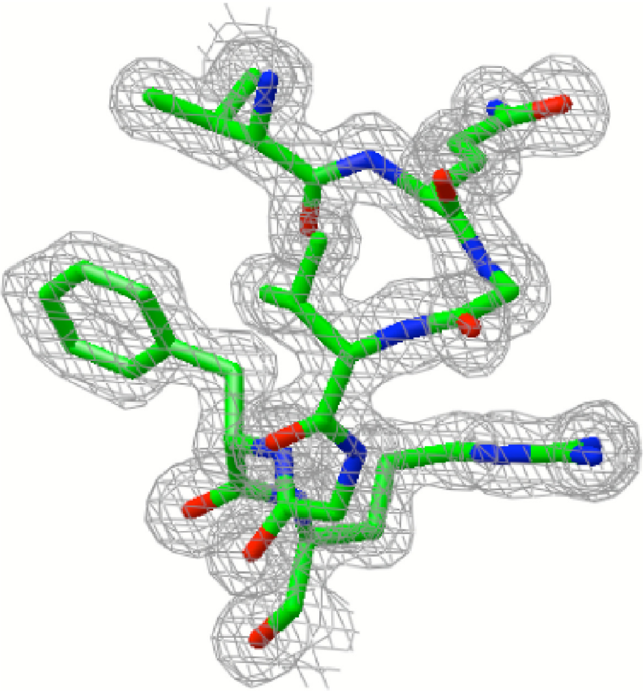


Fig. 27.3 CCP4mg display of portions of a $2F_o$–F_c map as an isomesh and 1w2i.pdb.[80]

for spherical atoms surfaces in the mid 1970s. Connolly developed a workable program to represent more general molecular surfaces by distributing dots on them in the early 1980s.[34] Many molecular graphics programs now have the capability to render three-dimensional dot surfaces.

27.2.6.2 *Patterson maps*

In X-ray crystallography, the electron density is not available until structure factors are phased. Unphased structure factors, however, can be used to compute the Patterson function,[35] producing the averaged electron density convolved with its enantiomorph, a function that can be plotted, with peaks at the vector difference between peaks in the actual averaged electron density. (In some space groups, certain sections of the Patterson map show the actual image of the molecule in projection.)

27.2.6.3 *Gaussian atoms and Kendrew models*

Classical small-molecule crystallography is based on the time-averaged electron density of individual atoms modeled as normally-distributed clouds of electrons. In the isotropic case, that allows the density contributed to each atom to be modeled as a Gaussian.

$$\rho(X,Y,Z)=\frac{N_e}{\sigma^3 2\pi\sqrt{2\pi}}\exp\left\{-\frac{(X-x)^2+(Y-y)^2+(Z-z)^2}{2\sigma^2}\right\}$$

While unphysical, such approximations simplify the problem for small molecules sufficiently to allow the available data to over-determine this limited set of parameters. If enough is known of the chemistry to set N_e, the number of electrons associated with the atom, this model needs only four additional parameters for each atom, the coordinates *x, y, z* of the center and the spatial standard deviation, σ, of the Gaussian, ideally a combined measure of the distribution of the intrinsic electron density of the atom and thermal vibrations. The anisotropic case introduces more parameters. If enough data are available, Gaussian atom models are physically plausible and good enough approximations for the interpretation of experimental measurements. Properly rendered isosurfaces of such models can be used to gain an understanding of some properties of a molecule such as ligand affinities, but other models can help to gain further understanding of these and other properties, such as evolutionary relationships. CPK models, wire models, and Lee-Richards surface models can help, especially when dealing with the interactions among multiple molecules. Experimental work with macromolecules usually does not provide sufficient data to determine even this limited set parameters for a macromolecule. In such cases, coarser models, such as Kendrew models that model groups of atoms as rigid bodies, are needed.

27.2.6.4 *CPK models*

A simple isosurface of density lacks information about the chemistry of a molecule. It is difficult to distinguish element types and to understand bonding patterns. Especially for small molecules, it can help to model a molecular surface emphasizing the distinct identities of individual atoms by representing each one as a hard sphere and coloring by element type. This was done with space-filling Corey-Pauling-Koltun (CPK) models, creating an implicit surface from the visible exterior of intersecting van der Waals spheres, coloring carbon atoms as grey, oxygen as red, nitrogen as blue, etc. Bonds are inferred from the overlap of the spheres. Figure 27.4 shows an example of a CPK model rendering of a portion of Protein Data Bank (PDB)[36,37] entry 4ins.[38]

27.2.6.5 *Wireframe models*

A CPK model emphasizes the atoms of a molecule. In a wireframe model, the bonds are emphasized by treating each atom as a point

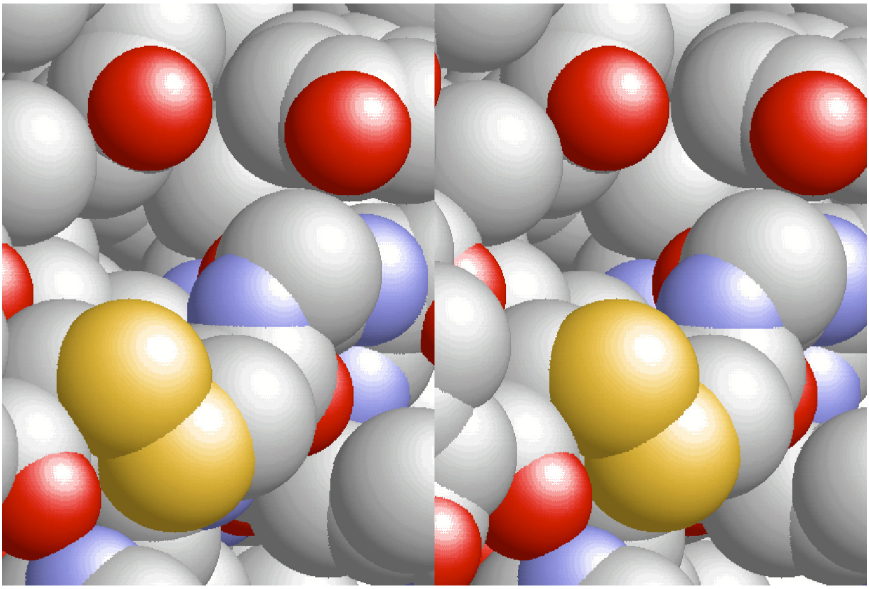


Fig. 27.4 RasMol cross-eyed stereo rendering of a CPK colored space-filling model of a portion of 4ins.

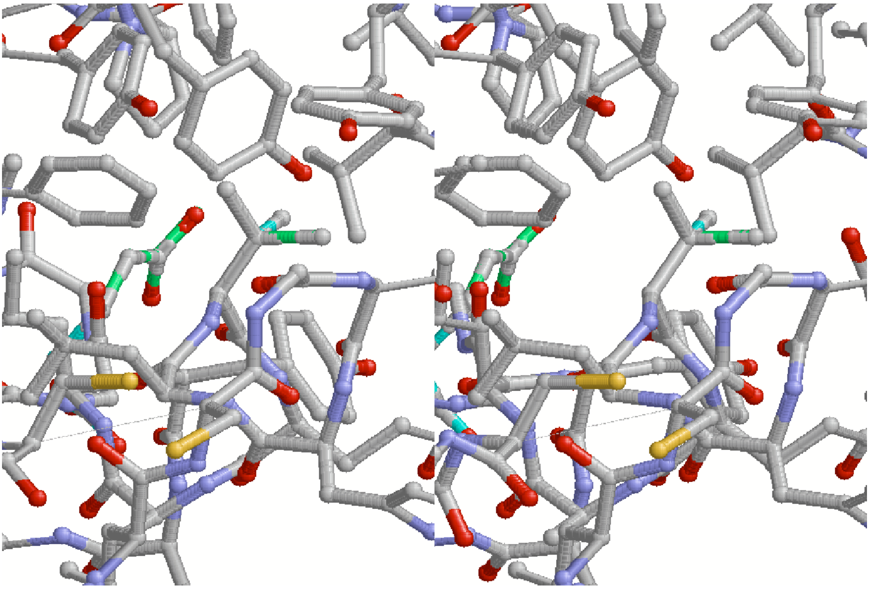


Fig. 27.5 RasMol cross-eyed stereo rendering of a wireframe model of a portion of 4ins.

and modeling the bonds as wires between them. The two approaches can be combined by representing the atoms as small balls and the bonds as sticks connecting those balls, creating a ball-and-stick model.[39] Johns Glass "Student Molecular Models" were used in the 1950s to build ball-and-stick models from colored wooden ball atoms with pre-drilled holes for coiled spring bonds. Kennard and Doré patented a way to construct complex wireframe models without the need for small balls to join the wires in 1966.[40] Figure 27.5 shows an example of a wireframe rendering of a portion of PDB entry 4ins.

27.2.6.6 *Model building and fitting to density*

X-ray crystallographers measure structure factors, the Fourier coefficients of the electron density in the unit cell. For noncentrosymmetric unit cells, the structure factors are complex numbers, with a magnitude (or absolute value) and a phase. The "phase problem" arises whenever it is possible to measure only the magnitudes of the Fourier coefficients. Solutions of the phase problem — experimental or theoretical — permit calculation of at least an approximate electron density. Solution of the structure requires interpretation of the electron density in terms of a molecular structure. Traditionally, this has required the expertise of a chemist, and the application of this expertise required a way to display the electron density and build a molecular model from it.

For small molecules, for which atomic resolution data are available, the calculated electron density contains well-defined peaks. In the first half of the twentieth century, this permitted the classic "needle and thread" approach: Crystallographers would lay out x and y coordinates on a sheet of graph paper pasted onto a styrofoam base. They would then insert knitting needles into the styrofoam: for each peak at coordinates (x, y, z) they would insert a needle into the styrofoam at (x, y) to a depth of $L - z$ where L is the length of the needle. Then, the crystallographer would tie threads between peaks nearby in space and thereby create a model of the molecular structure.

In the early days of protein crystallography, (and even now, depending on the disorder within the crystal) individual atoms and often even individual residues were not resolvable. This was the result in part of imprecision in measurement of structure factor magnitudes, but mostly of the relatively primitive power of phasing methods. It was not feasible for a computer program to interpret such electron density maps at atomic resolution. Instead, protein crystallographers became experts in pattern recognition and spent many hours building and rebuilding their models of the structures, trying to fit those models to their data.

27.2.6.7 *Kendrew models*

To build the structure of myoglobin, the first protein structure solved by X-ray crystallography, a set of vertical rods traversed molecular space, at a scale of 5 cm/Å. The electron density was represented by colored clips and the molecule built out of custom-built brass wire components.[41] Because of that pioneering work, standard, rigid body models of amino acids used to assemble protein models are called Kendrew models.

27.2.6.8 *Stacks of contour maps, isosurfaces*

Model building can be done as a theoretical exercise, but if one has crystallographic data, the model needs to be fit to the electron density inferred from the data. In the absence of phase information, one might think that the fitting would have to be done to a Patterson map, rather than to the density. However, a number of methods that yield approximate phases — isomorphous replacement, molecular replacement, multiwavelength anomalous dispersion, and direct methods — allow direct interactive fitting of a molecular model to the computed electron density.

Until the early 1960s, contour maps were generally computed by hand or by analogue computers.[42] During the 1950s, digital computer-driven plotters were developed,[43] e.g. Calcomp digital plotters that drew very precise, drafting quality images on paper, or most importantly, on clear plastic. In 1963, Dayhoff developed a program to contour electron density maps on clear plastic sheets that could then be stacked to produce a three-dimensional image of isosurfaces of density.[44]

Starting with Dayhoff, this approach involved contouring the electron density in successive planar slices of the unit cell, printing the contours on transparencies, and stacking the transparencies in a frame. The crystallographer would build a wire model to match the pattern in the electron density.

Richards contributed the ingenious idea of using a half-silvered mirror to allow optical superposition of a physical model and the electron density.[22] All protein crystallography laboratories had a "Richards box."

27.2.6.9 *Interactive graphics — Diamond, Katz/Levinthal, Jones*

The Richards box used computers to generate the contour and to provide the calculations upon which to base the building of the model, but the fitting itself was a manual process. Now, we interact with models and contours inside a computer, both manually and with software. Achieving that capability started with vector-based real-time renderings of molecular models on specialized computer hardware.

Katz and Levinthal pioneered the use of interactive graphics for structural biology in the 1960s.[45] By the mid 1970s, it was clear that interactive graphics would be more widely available for analysis.[46,47] By the late 1970s, interactive computer graphics was widely accepted and several groups created programs for the "electronic Richards box." Perhaps the first was Diamond's Bilder;[48] others were implemented at Washington University, St. Louis, and the University of North Carolina. Jones, working then in the laboratory of R. Huber in Munich, wrote a program FRODO[49] that became very widely used, going through several stages of development and implemented for various hardware configurations. Jones's program "O"[50] has superseded FRODO. The Richards boxes were put aside and the crystallographers spent long hours in dark rooms rebuilding their structures at computer graphics devices. This went on until the discovery that molecular dynamics refinement methods could shortcut the process.

27.2.6.10 *Surface models, rolling ball models, potentials*

Having the coordinates of protein structures in computers allowed more complex calculations to be applied to their analysis. Richards emphasized the importance of surface area and packing for understanding protein folding. However, a simple van der Waals surface of a protein is not sufficient to understand the problem. Atoms from the solvent need to be considered.

With Lee, Richards devised methods for measuring and displaying solvent-accessible surface area, and through Voronoi decomposition of the atomic distribution, the packing density inside proteins. Chothia began his work on proteins in that lab, showing that protein interiors were as densely packed as typical molecular crystals — despite the constraints of the chain connectivity. Chothia measured the burial of hydrophobic surfaces and derived the energy equivalent: 25 cal/Å^2.[51]

Lee and Richards defined the solvent-accessible surface as the locus of the center of a 1.4 Å-radius probe sphere (representing a water molecule) rolling around the protein structure, the surface of the sphere remaining in contact with the van der Waals surface (the surface of a CPK model) of the protein. The solvent-accessible surface is equivalent to a CPK model in which the van der Waals radius of each atom has been increased by the radius of the probe sphere.

A related concept is that of the solvent-excluded surface, which is the envelope of the macromolecule as touched by the rolling probe. This surface is identical to the van der Waals surface for portions of surface atoms of the macromolecule on which the probe can roll freely, but moves out into the solvent when the probe touches more than one atom, thereby creating an impression of a smoothed or blurred CPK model. Figure 27.6 shows an example of the solvent-excluded surface (otherwise known as the molecular surface) of a portion of PDB entry 4ins.

Lee-Richards surfaces were independently rediscovered by Connolly in 1983.

27.2.7 *Higher Order Structure — Schematic Diagrams*

When the structures of myoglobin and hemoglobin appeared almost 50 years ago, it was clear that there were profound similarities in their three-dimensional structures. Moreover, the most interesting similarities were in the global aspects of the structure, the shared overall folding pattern. To appreciate this, the large amount of detail in the full atomic structure of the individual residues is a distraction and a source of confusion.

It was clear that something was needed to extract and display the overall topology of the structure. An early laboratory approach to this was to thread tygon tubing along the main chain of a molecular model, and to pump an aliquot of fluorescent dye through the tube, visually "tracing the chain." In 1970, Rubin invented a device ("Byron's Bender") that would bend a wire into the form of the backbone of a protein structure. These were simple, convenient, and popular.

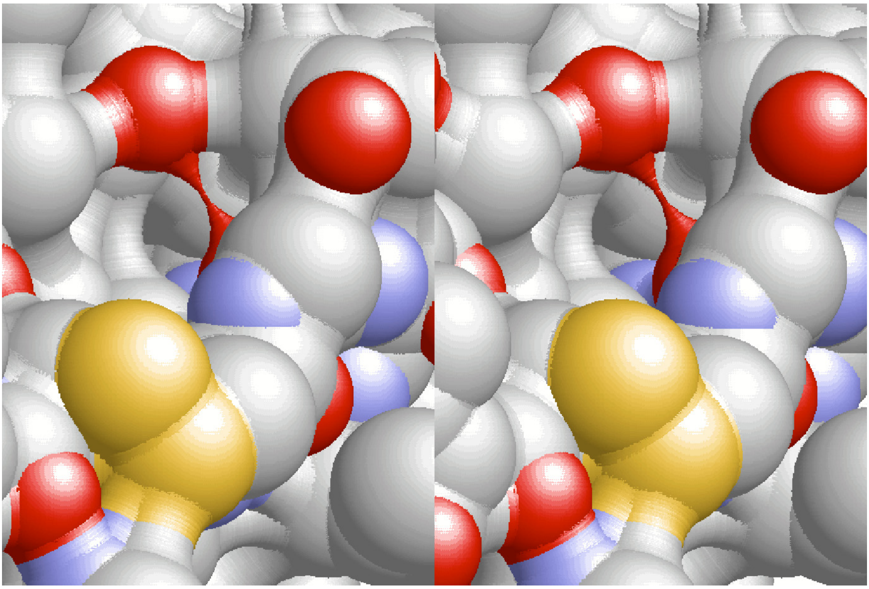


Fig. 27.6 RasMol cross-eyed stereo rendering of a Lee-Richards molecular surface model of a portion of 4ins.

Perhaps the first computer program designed specifically to present a simplified representation of protein structures created the "ribbon" diagram of McLachlan. This represented each residue by a trapezoid in the plane of the peptide and made a complicated three-dimensional structure intelligible by use of hidden-line removal.

Rossmann in the US and Furugren in Sweden developed the "cartoon" representation in which cylinders represented *α*-helices, and large arrows represented strands of *β*-sheet. Intervening loops appeared, relatively inconspicuously — as in most cases, they deserved to — as narrow tubes. Richardson took up this idea and executed such drawings for many structures. Her collection, in "Advances in Protein Chemistry," remains an influential source.

Aside from being labor-intensive and requiring manual skill, hand-drawn diagrams are ultimately limited by the difficulty of reorienting the viewpoint, of drawing stereo pairs, and especially, of superposition.

This has led to the writing of many computer programs for representations of proteins.

27.2.8 *Superposition*

Physical molecular models are fine, especially if one is interested primarily in one or a few structures. For many years, the "model room" of the Laboratory of Molecular Biology in Cambridge, England, contained wireframe structures of oxy- and deoxy-hemoglobin. Each occupied a cube of edge about four feet, mounted on a table. One problem was that if one wanted to compare details of the two structures all one could do would be to take a yardstick, determine an interatomic distance on one model, then go to the other model and measure the corresponding distance. The essential tool, superposition, was impossible. Over the years, the models were also subject to partial degradation.

27.3 Computational Techniques

The problems that molecular graphics can help us solve arise from the size, complexity, and variety of the systems we want to examine. Recalling that a typical protein structure contains thousands of residues and tens of thousands of atoms arranged with a well-defined spatial organization, the first requirement is to be able to look at such an object in an intelligible way. The two classical types of molecular models: the representation of every atom by a sphere and the representation of every chemical bond by a line segment, are entirely inadequate for structures so large. If what one wants to do is to be able to follow the course of the chain in three dimensions, a representation of a space curve following the backbone is something that does make sense to a molecular biologist, particularly if additional aids to three-dimensional perception are available, such as stereo and real-time rotation.

But simplification is too simple an answer. First of all, to show such a space-curve representation, a lot of information has been lost. Secondly, we not only want to be able to look at a structure, we want to be able to address a variety of architectural questions. We want to be able to tinker with a molecule — to ask, for example, what will be the effect on the structure of the replacement of one or more of the side-chains, which can occur either through natural mutation or laboratory synthesis. We may want to design grafts of part of one protein structure into a different, unrelated protein structure, e.g. the transfer of the combining site from a rat antibody to a human antibody for therapeutic purposes. These would be typical computer exercises in the field of biotechnology.

In practice, user interfaces provide facilities for interacting with a picture and with the underlying model that it represents. Many programs make it possible to design and alter pictures, to select and label atoms, to choose different colors for different regions or atom types, to "clip" portions of three-dimensional space (for example, to display only the portion of a structure within a sphere around an atom in order to show the neighbors with which it interacts), or to translate and reorient the current viewpoint. Enhancements of perception of spatial relationships within a large molecule are achieved by perspective, stereo, depth cueing, interactive real-time shifts in clipping planes, and the kinetic depth effect upon viewing an object in a simulated state of rotation. Many of these are standard elementary operations provided for in hardware or systems software in modern workstations.

An essential feature of the user interface is the ability to control and record the orientation of an object or set of objects being displayed. The two principal problems are: 1) how to specify numerically the orientation displayed, and 2) how to communicate to the computer the orientation desired. (A rigid three-dimensional object has three rotational degrees of freedom, and a mouse appears to have one component too few.)

27.3.1 *Interactive Control of What is Displayed and How it is Displayed*

Interactive graphics involves real-time user control of several aspects of a picture:

- What material is being shown
- How the material is represented
- The apparent orientation of the representation

Many modern graphics programs allow control over the choice of material through menus. Users can select which structures or substructures to include in the picture. For proteins, the user can select whether to display a trace (polygonal or smoothed) of successive C_α atoms only, or all backbone atoms, or all atoms, or a simplified representation or "cartoon." If more than one protein structure, or substructure, is displayed, it is possible to control the way the structures are superposed. Interactive search for well-fitting regions is an important tool for exploring relationships between structures. This involves separate choices of regions to superpose and regions to display. In some cases, automatic structural superposition programs can select the atoms for superposition.

For proteins or nucleic acids, choices of representation include: all-atom representations, skeletal representations showing bonding frameworks, ball-and-stick representations combining representations of atoms and bonds, chain traces, or schematic representations. Experience has shown that it is important to be able to combine different representations for different parts of a structure. A standard example is a picture that shows an active site in full detail, but the overall context of the entire structure in reduced detail. Subsidiary drawings in which a selected region is blown up are sometimes useful in this context (see Fig. 27.7).

Control over orientation is an essential component of interactive graphics. Various analog input devices have been used in computer graphics for specification of orientation, as well as for control of other aspects of a display. These include buttons, dials, slidebars, trackballs, and three-dimensional mice. A fairly obvious assignment was to allow three dials to control rotations around *x*-, *y*-, and *z*-axes. Perhaps the most complex was the SpaceBall, which combined three rotational and three translational degrees of freedom plus buttons.

It is possible to do an adequate simulation of a trackball using a mouse. Because mouse and keyboard are standard, and the other devices are not, most contemporary programs use the mouse and keyboard exclusively for interactive control of the display. Video gamers, generally better equipped than molecular biologists, use joysticks.

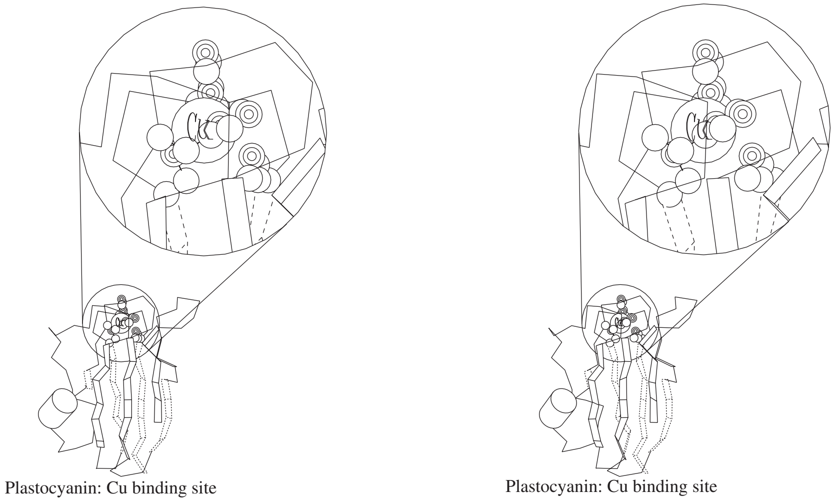


Fig. 27.7 Poplar leaf plastocyanin, line drawing. This picture shows the utility of showing in great detail the binding site for copper, and also in depicting the overall structural context of the binding site.

27.3.2 *Techniques*

An essential goal of a molecular picture is to produce something intelligible to a scientist. All possible techniques of computer graphics — whether realistic or imaginary — have been enlisted in this effort.

27.3.2.1 *Color and pseudocolor*

Some proteins have natural color, such as oxy- and deoxy-hemoglobin. This accounts for the difference in appearance of arterial and venous blood. However, most of the color in computer-generated diagrams is artificial. Corey, Pauling, and Koltun assigned black to carbon, blue to nitrogen, and red to oxygen. (Note that this is not consistent with the colors of litmus paper in the presence of basic and acidic groups.) Many graphics programs use different colors to distinguish different atom types and provide some variation on the Corey, Pauling, and Koltun (CPK) color assignments. The major change from the original CPK colors is that carbon is now more commonly presented as grey, rather than black, to avoid problems with black backgrounds.

Many programs permit selection of color as part of the selection of material and representation. For instance, it may be deemed helpful to color helices red, strands blue, and loops yellow. [Readers are urged not to use red and green for important structural distinctions, as the pictures thereby produced lose their point for a significant portion of the audience that has the most common form of color-blindness. Microarray pictures that conventionally show upregulated genes in green and downregulated genes in red are common offenders.]

Another common use of pseudocolor is the rainbow effect, in which the color in a chain trace varies from red at the *N*-terminus, through the colors of the spectrum, to violet at the *C*-terminus. This is a useful way to show the direction of the chain. An alternative is to draw chevrons on the chain (see Fig. 27.1).

27.3.2.2 *Light sources, shadows, shading and depth cueing, texture, transparency*

An atom is not a ball with a crisp glossy surface. A bond is not a metal wire nor is it a plastic rod. However, the most commonly used physical models of atoms are plastic balls, and of bonds, metal wires or plastic rods.

Therefore, many graphics programs produce pictures that appear as realistic pictures of macroscopic plastic models of chemical structures. Modern rendering software allows the choice of positions and other characteristics of multiple light sources, with different beam dispersion characteristics, e.g. simulating diffuse lights or spotlights. In addition to assigning color to elements of the picture, more complex textures are possible, including mosaic patterns and simulations of natural materials such as stone and wood. It is now easy to make extremely fancy pictures. Different colors and textures can provide a means to represent atom and residue types, charge, temperature factors and other important parameters, but care is needed to avoid overwhelming the chemistry with decoration.

It is important to give the viewer a sense of three dimensions, of relations between the depths of different parts of the picture. Shadows within the picture do this, as does hidden-surface removal (the simulation of opacity). Translucency can be simulated also but is not easily effective. (In line drawings, simulation of translucency can be achieved by not removing hidden lines but converting them to broken lines.)

Depth cueing is the reduction of intensity of objects according to their distance from the viewpoint. It is a simulation of viewing the model through fog.

27.3.2.3 *Stereo*

Stereoscopic vision is an important component of our depth perception in everyday life. Because our eyes are (typically) 2.5 inches (63 mm) apart, they receive slightly different views of the scene we are looking at. Our brains integrate the information and interpret the differences in terms of depth: the larger the difference, the closer the element of the scene.

To simulate this effect in a drawing, it is necessary to: (1) create two different views appropriate for the left and right eyes, and (2) deliver the two views separately to the left and right eyes. Typical simulated geometries place an object in a box 12 inches (305 mm) wide, 9 inches (229 mm) high, and 6 inches (252 mm) deep, viewed from a distance of 30 inches (762 mm), with an eye-separation of 2.56 inches (65 mm).[52] Often the two images are created simply by rotating the object slightly for each eye, say by ±2.95°. Especially when done without a perspective correction, this simple rotation creates an impression of a distant image. Thomas[53] argues for accurate geometry in reproduction. Done properly, such accurate geometry can create the impression of a real object coming out from the display screen. Such images can be so real that viewers try to grasp them.

Even approximately, it is difficult but not impossible to draw stereo pictures without computer automation. An early example in our field was a set of crystal structure illustrations by von Laue,

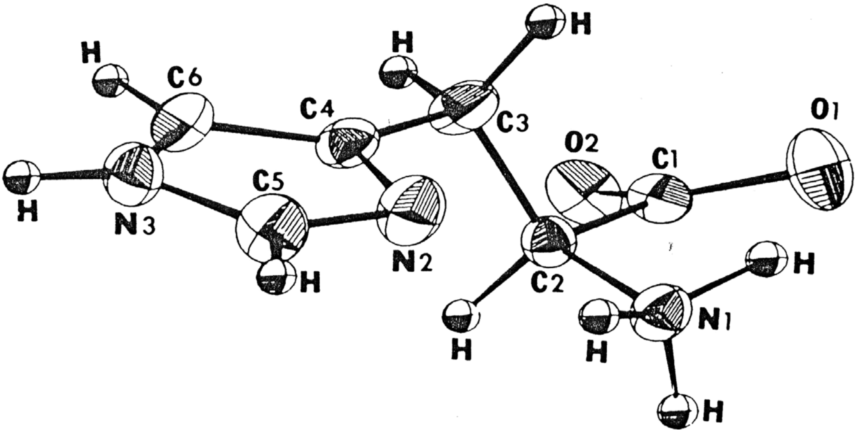


Fig. 27.8 ORTEP (Oak Ridge Thermal Ellipsoid Program) picture of the amino acid histidine. (Courtesy of Dr. J. Madden.)

Menzer, von Simson, Verständig, and von Mises.[54] Modern computer graphics makes it easy to generate the two views. An early program used by several generations of crystallographers was Johnson's Oak Ridge-Thermal Ellipsoid Program (ORTEP). PLUTO, by Motherwell,[55] followed five years later. These programs calculated the appropriate line segments and produced output on a pen plotter. A particular algorithmic challenge was "hidden-line removal," the simulation of opacity. Later, interactive computer graphic devices did the geometrical calculations, but not the hidden line removal, in hardware. (See Fig. 27.8 for a typical ORTEP drawing.)

Delivery of separate views to separate eyes is achievable by physical separation (side-by-side), temporal separation (time-sliced), filters (e.g. red-green overlays), or with a lenticular screen ("naked-eye" stereo displays). Side-by-side pictures can be viewed by standard lens systems, or with a little practice, without aids. Side-by-side figures place the left-eye image either to the left of the right-eye image (requiring divergence of the eyes — "wall-eyed stereo") or to the right ("cross-eyed stereo"), or both (allowing the viewer to choose whichever action is easier). Time-sliced stereo requires some kind of goggles that coordinate the display of left and right images with the change from opacity to transparency of filters in front of the two eyes. In the worst case, these goggles are bulky, heavy, fragile, and expensive. Projecting polarized images onto a suitable screen requires only simple polaroid filters. Many installations now have viewing "caves," immersing the viewer within the scene. Coupling these with motion or force sensors gives a tactile (colloquially, a "touchie-feelie") illusion.

27.3.2.4 *Kinetic depth effect*

It is much more difficult to sense the depth in a static scene than in a moving one. Our brains infer positions for objects from their relative motion, seeing three dimensions in a sequence of two-dimensional images that show realistic projections of what we would expect to see from a moving scene. Stopping the motion usually causes a loss of this kinetic depth perception.

The combination of the kinetic depth effect with depth cueing and/or with perspective projection helps to avoid confusion between the intended scene and one in which the front and back of the scene are exchanged, producing the enantiomorph.

27.3.2.5 *Computing surfaces, contours, isosurfaces, and tesselations*

Johnson's ORTEP could produce atom-by-atom surfaces, either as distinct atom-by-atom thermal ellipsoids, or with the 1971 version, as van der Waals surfaces by representing atoms as overlapping spheres. At that time, the plots were in black and white, making it difficult to produce a full rendering of a CPK model in this manner.

Raster-based interactive graphics systems had color and could do full renderings of CPK models, but, in the early 1970s, such systems were expensive, and therefore, rare. In general, renderings of isosurfaces as meshes of contour lines were used for interactive presentation of surfaces.

Over the next two decades, performance improved and prices dropped. Raster-based graphics became the norm, displacing most uses of vector-based graphics for interactive work. For raster-based graphics, hidden line-removal is a much simpler problem than for vector-based graphics. Commercial firms such as Evans and Sutherland and Silicon Graphics made moderate-cost systems that were capable of doing high quality surface renderings in real-time, using hardware to allow software to manage surfaces in terms of triangular decompositions (tesselations). In 1991, the program GRASP[56] made the rendering of property-colored solvent-excluded surfaces in an interactive graphics environment accessible to those with Silicon Graphics workstations. In 1992, Sayle's program RasMol[57–58] made rendering of full CPK models accessible to anyone with almost any raster-based graphics workstation even while doing pixel-by-pixel (as opposed to triangle-by-triangle) renderings of van der Waals surfaces.

27.4 Recent History and Current Practice of Molecular Graphics

The following is an imperfect effort to provide some of the major highlights in the history of molecular graphics since the late 1980s, with apologies to the authors of the many important efforts not mentioned. Our focus is on examples of work in the past two decades that has had a significant impact on current practice. The threads we will explore start with ORTEP, MIDAS, Molscript, Kinemage, RasMol.

27.4.1 *ORTEP*

One of the first molecular graphics programs, ORTEP, is still used. It is the gold standard for the renderings of small molecules and ligands for macromolecules, providing deep scientific insight into thermal librations of such molecules. The current version, ORTEP III,[59] retains all the features and functionality of the original ORTEP and has added "semi-interactive" capabilities, accepting some input from the user in response to prompts and optionally displaying output on the screen instead of only as a paper plot. The authors have done a remarkable job of keeping the program compatible with data sets from 40 years ago while providing an essential tool for current studies that produces molecular drawings of remarkable scientific accuracy, clarity and beauty.

27.4.2 *MIDAS and Chimera*

In the late 1980s, UCSF Computer Graphics Laboratory's MIDAS[60] provided graphics tools for drug design by visualization and docking of molecules. This package was widely used throughout structural biology and helped to make high quality graphical representations of molecules an essential element of scientific research in structural biology. The successor to MIDAS is Chimera,[61] which is one of the most mature and feature-rich of the currently used interactive molecular graphics programs with startlingly realistic images, especially when viewed in stereo.

27.4.3 *Insight and VMD, SWISS-MODEL and DeepView*

Insight,[62] developed at UCSF and by Biosym Technologies, and, later, as Insight II at Accelrys, is an important example of a molecular graphics and molecular modeling program. The distinction between molecular graphics programs and molecular modeling programs is fuzzy, but important. A graphics program focuses on the ability to visualize a model of a molecule. A modeling program focuses on the ability to assemble and change a model of a molecule. There are many modeling programs drawing tools from physics and chemistry to synthesize small molecules (and increasingly macromolecules) from scratch or to combine experimental results with semi-classical and quantum-mechanical models. Some of the modeling tools are coupled with high quality rendering capabilities, as in Insight, and in other cases, rely on external molecular graphics packages for the visualization of results. The molecular graphics program VMD[63] is very well designed to serve as an interactive visualization front end to molecular dynamics modeling programs. The molecular graphics program DeepView (Swiss-PDBViewer) serves as the visualization front end for the homology modeling server SWISS-MODEL.[64]

27.4.4 *Molscript, Bobscript, Raster3D, and POV-Ray*

In 1991, Kraulis released Molscript.[65] Molscript is a rendering system both for visualization and export to other rendering programs. When "photorealistic" renderings of models are required, Molscript or an extended form by Esnouf, Bobscript,[66] are typically used with Merritt's program Raster3D.[67] This is the standard for high-quality renderings of macromolecules as static images. Raster3D provides a powerful set of rendering tools for quality images of ribbons, space-filling atoms, etc., but most importantly it does a careful ray-tracing from light sources to the molecular model and onto the viewer. In recent years, the ray-tracing engine started by Buck in 1986 for a game computer (the Amiga) and which became a full-fledged multi-platform ray-tracing engine in the early 1990s, has, in the form of POV-Ray, also started to be used as a ray-tracing alternative to Raster3D (see http://www.povray.org/) for photorealistic images. Lesk rewrote the Lesk-Hardman program to make use of general rendering software. Many pictures produced by this software appear in Ref. 68.

27.4.5 *Kinemage, Movie-Making*

In 1992, Richardson and Richardson released a Macintosh-based visualization package for kinemages (kinetic images) of macromolecules.[69] The package consists of several programs that are now available on a wide range of platforms (see http://kinemage.biochem.duke.edu/kinemage/magepage.php#defined):

"A 'kinemage' (kinetic image) is a scientific illustration presented as an interactive computer display. […] A kinemage is prepared in order to better communicate ideas that depend on three-dimensional (or more) information. The kinemages are distributed as plain text files of commented display lists and accompanying explanations. […] They are viewed and explored in an open-ended way by the reader using either the Mage or KiNG graphics program. A kinemage file can be generated either by a program or hand-cobbled. A utility called Prekin makes a starting kinemage from a PDB-format coordinate file. […]".

A kinemage is not a movie, but it validated the use of movies as a routine mechanism for communication about molecules. It is now routine to make short movies and to embed them in presentations. Almost any molecular graphics program is capable of exporting a sequence of images, each one showing a small incremental change in position from the previous one, so that they can be assembled into a flip-book movie.

In the 1970s, movie-making was a slow and daunting process, involving hours to days of work for even a short movie, as frames were transferred one at a time to film. This changed gradually as computer processing speeds and disk capacity increased, so that movies could be stored and viewed in real time, at first locally on a computer in the late 1980s and early 1990s using the mpeg format, and then in the 1990s, as the world-wide-web developed, on the Internet. By 1996, Netscape had introduced a web-page image format called “animated GIF” that allowed short movies to be introduced into web pages.

Using a package from the early 1990s, called ImageMagick (see http://www.imagemagick.org/script/history.php), any user could use almost any molecular graphics program to produce the images for a flipbook that ImageMagick would assemble into a short movie. While making such movies is now commonplace, the process can be daunting to non-experts and can demand resources that may not be available on local computers. Both problems have been solved by a variety of servers on the Internet that accept PDB ID codes or accept uploads of PDB files. One of the first of these was the “PDB to MultiGIF” web site by Bohne[70] at http://www.glycosciences.de/modeling/pdb2mgif/ that offers the user a simple web form to generate an animated GIF using RasMol. Short movies are of particular value for “morphing” (a concept dating back to the imagery in[71] molecules, i.e. for showing changes in molecular conformation during various biological interactions. The Yale Morph Server at http://www.molmovdb.org/morph/ generates the necessary intermediate images from two end-point configurations. There are various movie-making servers with differing capabilities. The “World Index of Molecular Visualization Resources” at http://molvis.sdsc.edu/visres/ established by Martz and Kramer is a visitor-maintained site that includes a list of such “Molecular 3D Visualization Servers.” Another useful compendium appears at http://www.hhmi.swmed.edu/external/crystallography/gr1.html.

27.4.6 *RasMol, Chime, Jmol, PyMol, and ccp4mg*

In 1992, Sayle described the program RasMol, which uses a highly efficient rendering algorithm to allow high quality interactive rendering of macromolecules on a wide variety of platforms. The program was released in 1993 and is still heavily used. Sayle’s maintenance and development of RasMol was supported by Glaxo for several years. RasMol is still actively maintained.[72] RasMol is a particularly easy-to-use program oriented towards rapid and simple display of PDB entries and coordinate sets from a variety of packages. It provides simple menus and a highly intuitive command language. It is used both for high-end research and for education down to the kindergarten level. RasMol was first released just as the world-wide-web was becoming popular, and was adopted as a helper application for browsers. Helper applications get their own windows on the screen, competing for screen “real estate” with the original browser window and creating some confusion in the handling of the mouse and keyboard.

MDL Systems, Inc. recognized the need for a version of RasMol that was better integrated with web browsers and created Chemscape Chime (see http://www.mdl.com/chemscape/chime), a derivative of RasMol that works as a web browser “plug-in.” The virtue of tight coupling with browsers created compatibility issues as browsers changed. This problem was solved when the developers of an existing, open source, java-based molecular graphics program, Jmol by Gezelter,[73] adopted the RasMol command language and turned Jmol into a full replacement for Chime that was able to work with a wide range of web browsers (see http://jmol.sourceforge.net/history/).

It is important to understand the role of non-interactive scripting in modern interactive molecular graphics. A script can be generated as a record of the mouse manipulations and menu selections used to create a desired image and then read back in at a later date to recover the state of the program in preparation for further manipulations.

Then, often with appropriate hand editing, such scripts can become tutorials on the interesting features of the same molecules or initialization scripts to select appropriate color schemes and renderings for other molecules. In the decade and a half that RasMol has been available, many tutorial scripts have been written for both RasMol and for Chime. Many originally written for Chime are now being converted to Jmol versions, and because both RasMol and Jmol are open source, it is expected that fairly complete compatibility will be achieved between RasMol and Jmol scripts.

Because RasMol is written in C, a fully compiled language, it achieves higher performance than Jmol, which is written in java, a compiled and then interpreted language with performance issues. Python, itself a powerful scripting language, achieves much higher performance in graphics applications than java with good platform independence. PyMol by DeLano[74] and ccp4mg by Potterton *et al.*[75] are two full-featured molecular graphics programs written in Python. Both produce very well-rendered images. As of this writing, ccp4mg is one of the most comprehensive of the currently available packages.

27.5 Current Choices in Hardware and Software

The improvement in performance of commodity personal computers with high performance graphics capabilities has resulted in wider access to suitable platforms for molecular graphics. This has resulted in a certain degree of standardization for both hardware and software, and has significantly reduced the use and availability of specialized molecular graphics systems. Most currently supported molecular graphics systems now assume one of three major hardware/operating-system platforms: an x86 CPU running some variant of Linux, an x86 CPU running some version of Microsoft Windows, or an x86 or PowerPC running some version of Mac OS X. There is still a need for applications that run on Silicon Graphics, Sun, Hewlett Packard, and IBM graphics workstations, but the approach to control and display is now determined by the demands and capabilities of Linux, Windows, and Mac OS in relatively inexpensive mass-marketed personal computers. Such personal computers are called “commodity” personal computers.

Until the second half of the first decade of the twenty-first century there was some advantage to doing molecular graphics on large high-resolution cathode ray tubes using specialized (and expensive) graphics cards to drive them. Inexpensive personal computer monitors and graphics hardware now have similar capabilities, except in the rendering of flicker-free high-resolution stereo, an application for which high-resolution cathode ray tubes remain better suited than flat panel displays.

The common use of commodity personal computers has resulted in a simplification and standardization of approaches to control and display molecular graphics, to make use of the type of keyboard, mouse, and display found on such computers, using the approach that originated as the Xerox Palo Alto Research Center Windows-Icons-Mouse-Pointer (Xerox PARC WIMP) paradigm.[76] That approach is now common to personal computers, so much so that some newer applications are weak in their support for other approaches to control, such as scripting with command files, and lack capabilities for high quality rendering in media other than on commodity displays.

The remaining major exception to the near universal use of commodity personal computers for molecular graphics is when high performance interactive stereo display is needed. In order for polarized shutter glasses to present 60 complete images per second without significant flicker, the display must be capable of 120 images per second (60 per second for the left eye and 60 per second for the right eye in alternation). Modern LCD displays are still too slow to change images that quickly.

27.5.1 *Current Applications*

There are numerous currently used molecular graphics programs, many of which are available for download on the Internet. Some are used for experimental research, some for theoretical work, some for education, and some for all three. Space does not permit a discussion of all the programs and packages currently available. Fortunately, there are many useful resources on the Internet to assist in locating and comparing relevant packages, e.g. http://www.liv.ac.uk/Chemistry/Links/refmodl.html, http://molvis.sdsc.edu/visres/index.html and http://www.rcsb.org/pdb/static.do?p=software/software_links/molecular_graphics.html.

As noted above, ORTEP, Chimera, Raster3D, Molscript, Bobscript, RasMol, jmol, PyMol, and ccp4mg cover a large portion of the current types of molecular graphics applications, but there are many other applications and servers. Both open source and proprietary packages now provide capabilities that allow scientists throughout the structural biology community to communicate their results with almost the same clarity, precision, artistry, and scientific precision that Dickerson and Geis achieved four decades ago.

27.6 The Future

It is likely that personal computers will continue to become more capable and more cost-effective with higher performance graphics. This will allow molecular graphics programs to assume the availability of more of the resources needed to draw high quality, complex molecular images in real time. Network bandwidths will continue to increase. The combination of these trends should allow three-dimensional, dynamic photo-realistic images to become the norm in the communication of results in structural biology. A less obvious trend, but arguably more important trend is towards greater commonality and increasing cross-communication among graphics programs. The PDB format from the 1970s for macromolecular structures and the Crystallographic Information File (CIF) from the 1990s[77] for small molecules have allowed easy interchange of data among molecular graphics programs. The scripting language from RasMol has shown promise as a mechanism to allow rendering commands to be shared. It seems likely that a language combining the clarity and simplicity of the RasMol command set with the power and extensibility of the Python-based language used by PyMol will be developed and will allow users to move freely among a wide range of graphics packages.[78]

References

1. Katz L, Levinthal C. (1972) Interactive computer graphics and representation of complex biological structures. *Ann Rev Biophys Bioeng* **1**: 465–504.
2. Corey RB, Pauling L. (2004) Molecular models of amino acids, peptides, and proteins. *Rev Sci Inst* **24**: 621–627.
3. Linnaeus C. (1753) *Species Plantarum*, Stockholm.
4. Ray J. (1670) *Catalogus Plantarum Angliae*, London.
5. Darwin C. (1859) *On the Origin of Species by Natural Selection*. Murray.
6. Mendel G. (1865) *Versuche über Pflanzen-Hybriden* (Experiments with plant hybrids). *Proceedings of the National History Society of Brunn* (Bohemia, now Czech Republic).
7. Franklin RE, Gosling RG. (1953) Molecular configuration in sodium thymonucleate. *Nature* **171**: 740–741.
8. Watson JD, Crick FHC. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**: 964–967.
9. Blow D. (2002) Max Perutz (1914–2002). *Quart Rev Biophysics* **35**: 201–204.
10. Reichert ET, Brown AP. (1909) *The Differentiation and Specificity of Corresponding Proteins and Other Vital Substances in Relation to Biological Classification and Organic Evolution: The Crystallography of Hemoglobins.* Carnegie Institution, Washington, DC.
11. Stenö N. (1669) *De Solido Intra Solidum Naturaliter Contento (Concerning a solid body enclosed by process of nature within a solid): Ddissertationis Prodromus.* Jacobum Moukee.
12. Paternò E. (1869) Intorno all'azione del percloruro di fosforo sul clorale (On the action of phosphorus pentachloride on chloral). *Giornale di scienze naturali ed economiche di Palermo* **5**: 117–122.
13. van't Hoff JH. (1874) Voorstel tot Uitbreiding der Tegenwoordige in de Scheikunde gebruikte Structuurformules in de Ruimte, benevens een daarmee samenhangende Opmerking omtrent het Verband tusschen Optisch Actief Vermogen en chemische Constitutie van Organische Verbindingen (Proposal for the extension of current chemical structural formulas into space, together with related observation on the connection between optically active power and the chemical constitution of organic compounds). *Arch Neerl Sci Exact Nat* **9**: 445–454.
14. Le Bel JA. (1874) Sur les relations qui existent entre les formules atomiques des corps organiques, et le pouvoir rotatoire de leurs dissolutions (On the relations that exist between the atomic formulas of organic compounds and the rotatory power [i.e. the optical activity] of their solutions). *Bull Soc Chim Fr* **22**: 337–347.
15. Haüy RJ. (1822) *Traité de cristallographie* (Treatise on Crystallography). Bachelier et Huzard.

16. Nagendrappa G. (2007) Pasteur — the harbinger of stereochemistry. *Resonance* **12**: 38–48.
17. Pasteur L. (1860) *Note Relative au Penicillium Glaucum et à la Dissymétrie Moléculaire des Produits Organiques Naturels.* Mallet-Bachelier, Imprimeur-Libraire.
18. Lehn JM. (1995) *Supramolecular Chemistry: Concepts and Perspectives: A Personal Account Built Upon the George Fisher Baker Lectures in Chemistry at Cornell University [and] Lezioni Lincee, Accademia Nazionale Dei Lincei, Roma*. Wiley-VCH.
19. Perrin J. (1909) Mouvement brownien et réalité moléculaire (Brownian movement and molecular reality). *Annales de Chimie et de Physique* **18**: 1–114.
20. White HE. (1931) Pictorial representations of the electron cloud for hydrogen-like atoms. *Phys Rev* **37**: 1416–1424.
21. Feldmann RJ, Bing DH. (1980) *TAMS: teaching aids for macromolecular structure. Teachers manual.* Division of Computer Research and Technology (DCRT), NIH/PHS/DHEW.
22. Lee B, Richards FM. (1971) The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**: 379–400.
23. Dickerson RE, Geis I. (1969) *The Structure and Action of Proteins.* W. A. Benjamin.
24. Geis I. (1958) Drawing of sperm whale myoglobin. Available at: http://www.math.fsu.edu/~quine/IntroMathBio_05/Proteins/myoglobin_geis.jpg.
25. Holbrook JJ, Liljas AA, Steindel SJ, Rossmann MG. (1975) Lactate dehydrogenase. *The Enzymes* **11**: 191–292.
26. Brändén CI, Jörnvall H, Eklund H, Furugren B. (1975) Alcohol dehydrogenases. *The Enzymes* **11**: 103–190.
27. Levitt M, Warshel A. (1975) Computer simulation of protein folding. *Nature* **253**: 694–698.
28. Richardson JS. (1981) The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34**: 167–339.
29. Lesk AM, Hardman KD. (1982) Computer-generated schematic diagrams of protein structures. *Science* **216**: 539–540.
30. Lesk VI, Lesk AM. (1989) Schematic diagrams of nucleic acids and protein-nucleic acid complexes. *J Appl Cryst* **22**: 569–571.
31. Cruickshank DWJ. (1959) Fourier synthesis and structure factors. In: *International Tables for X-ray Crystallography*, Vol. II, pp. 317–340. Kynoch Press.
32. A Society of Gentlemen in Scotland. *Encyclopaedia Britannica or a New Dictionary of Arts and Sciences*, Vol. II. Bell and Macfarquhar, 1771.
33. Mathews FS. (1983) Interactive graphics in the study of molecules of biological interest. In: *Crystallography in North America*, pp. 235–240. American Crystallographic Association.

34. Connolly ML. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**: 709–713.
35. Patterson AL. (1934) A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys Rev* **46**: 372–376.
36. Bernstein FC, Koetzle TF, Williams GJB *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **112**: 535–542.
37. Berman HM, Westbrook J, Feng Z *et al.* (2000) The Protein Data Bank. *Nucl Acids Res* **28**: 235–242.
38. Baker EN, Blundell TL, Cutfield JF *et al.* (1988) The Structure of 2Zn Pig Insulin Crystals at 1.5 Å Resolution. *Philos Trans Roy Soc London. Series B, Biol Sci* **319**: 369–456. PDB ID: 4ins.
39. Dodge FD. (1931) US Patent No. 1,851,159, Means for Constructing Stereochemical Models.
40. Kennard O, Gamblin Doré CF. (1966) US Patent No. 3,286,339, Construction of Models Representing Molecular and Other Structures.
41. Kendrew JC, Bodo G, Dintzis HM *et al.* (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**: 662–666.
42. Newham RE. (1983) Crystallography at Penn State. In: *Crystallography in North America*, pp. 88–91. American Crystallographic Association.
43. Auerbach on Digital Plotters and Image Digitizers. Auerbach Publishers, 1972.
44. Dayhoff MO. (1963) A contour-map program for X-ray crystallography. *Commun ACM* **6**: 620–622.
45. Levinthal C. (1966) Molecular model-building by computer. *Scientific American* **214**: 42–52.
46. Bernstein HJ, Andrews LC, Berman HM *et al.* (1974) CRYSNET — a Network of Intelligent Remote Graphics Terminals. In: *Second Annual AEC Scientific Computer Information Exchange Meeting, Proceedings of the Technical Program*, pp. 149–161. Brookhaven National Laboratory Report #18803.
47. Meyer EF Jr, Morimoto CN, Villarreal J *et al.* (1974) CRYSNET, a crystallographic computing network with interactive graphics display. *Fed Proc* **33**: 2402–2405.
48. Diamond R. (1980) BILDER: a computer graphics program for biopolymers and its application to the interpretation of the structure of tobacco mosaic virus protein discs at 2.8 Å resolution. *Biomol Struct Conform Funct Evol* **1**: 567–588.
49. Jones TA. (1985) Diffraction methods for biological macromolecules. Interactive computer graphics: FRODO. *Meth Enzymol* **115**: 157–171.
50. Jones TA, Bergdoll M, Kjeldgaard M. (1990) O: a macromolecular modeling environment. In: *Crystallographic and Modeling Methods in Molecular Design*, pp. 189–195. Springer-Verlag, New York.
51. Chothia C. (1984) Principles that determine the structure of proteins. *Ann Rev Biochem* **53**: 537–572.

52. Johnson CK. (1965) ORTEP. Report ORNL-3794. Oak Ridge National Laboratory, Tennessee, USA.
53. Thomas DJ. (1993) Toward more reliable printed stereo. *J Mol Graph* **11**: 15–22.
54. von Laue M, Menzer G, von Simson C, Verständig E, von Mises R. (1926) *Stereoskopbilder von Kristallgittern (Stereoscopic Drawings of Crystal Lattices)*. Springer.
55. Motherwell WDS. (1992) *PLUTO, a Program for Plotting Molecular and Crystal Structures.* University Chemical Laboratory, Cambridge, England, 1976.
56. Nicholls AJ. *GRASP Manual: Graphical Representation and Analysis of Surface Properties.* Columbia University.
57. Sayle R, Bissell A. (1992) RasMol: a program for fast realistic rendering of molecular structures with shadows. *Proceedings of the 10th Eurographics UK '92 Conference, University of Edinburgh, Scotland.*
58. Sayle RA, Milner-White EJ. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem Sci* **20**: 374.
59. Burnett MN, Johnson CK. (1996) ORTEP-III: Oak Ridge Thermal Ellipsoid Plot Program for Crystal Structure Illustrations. *Report ORNL-6895.* Oak Ridge National Laboratory, Oak Ridge, TN, USA.
60. Ferrin TE, Huang CC, Jarvis LE, Langridge R. (1988) The MIDAS display system. *J Mol Graph* **6**: 13–27.
61. Pettersen EF, Goddard TD, Huang CC *et al.* (2004) UCSF Chimera — a visualization system for exploratory research and analysis. *J Comput Chem* **25**: 1605–1612.
62. Dayringer HE, Tramontano A, Sprang SR, Fletterick RJ. (1986) Interactive program for visualization and modeling of proteins, nucleic acids, and small molecules. *J Mol Graph* **4**: 82–87.
63. Humphrey W, Dalke A, Schulten KK. (1996) VMD-visual molecular dynamics. *J Mol Graph* **14**: 33–38.
64. Guex N, Peitsch MC. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**: 2714–2723.
65. Kraulis PJ. (1991) Molscript. *J Appl Crystallogr* **24**: 946–950.
66. Esnouf RM. (1999) An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J Mol Graph* **15**: 132–134.
67. Merritt EA, Bacon DJ. (1997) Raster3D photorealistic graphics. *Meth Enzymol* **277**: 505–524.
68. Lesk AM. (2001) *Introduction to Protein Architecture: the Structural Biology of Proteins.* Oxford University Press.
69. Richardson DC, Richardson JS. (1992) The Kinemage: a tool for scientific communication. *Protein Sci* **1**: 3–9.
70. Bohne A. (1998) PDB2multiGIF: a web tool to create animated images of molecules. *J Mol Model* **4**: 344–346.
71. Isaiah. 2:4.

72. Bernstein HJ. (2000) Recent changes to RasMol, recombining the variants. *Trends Biol Sci* **25**: 453–455.
73. Gezelter D. Jmol: an open source Java program. See http://www.Openscience.org/jmol.
74. DeLano WL. (2002) The PyMOL Molecular Graphics System. DeLano Scientific, Palo Alto, CA, USA. http://www.pymol.org.
75. Potterton L, McNicholas S, Krissinel E *et al.* (2004) Developments in the CCP4 molecular-graphics project. *Acta Cryst* **D60**: 2288–2294.
76. Thacker CP, McCreight EM, Lampson BW *et al.* (1979) Alto: a personal computer, Xerox PARC. Technical Report CSL-79-11.
77. Hall SR, Allen FH, Brown ID. (1991) The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Cryst* **A47**: 655–685.
78. Westin C, Hanson B, Bernstein HJ *et al.* (2007) SBEVSL: communicating Scripts between Molecular Visualization Programs Poster presentation TP172, abstract E0003, American Crystallographic Association 2007 Meeting, 21–26 July 2007, Salt Lake City, Utah.
79. Ludwig ML, Pattridge KA, Metzger AL *et al.* (1997) Control of oxidation-reduction potentials in flavodoxin from *Clostridium beijerinckii*: the role of conformation changes. *Biochemistry* **36**: 1259–1280. PDB ID: 5nll.
80. Cheung YY, Lam SY, Chu WK *et al.* (2005) Crystal structure of a hyperthermophilic archaeal acylphosphatase from *Pyrococcus horikoshii* — structural insights into enzymatic catalysis, thermostability, and dimerization. *Biochemistry* **44**: 4601–4611. PDB ID: 1w2i.

Index