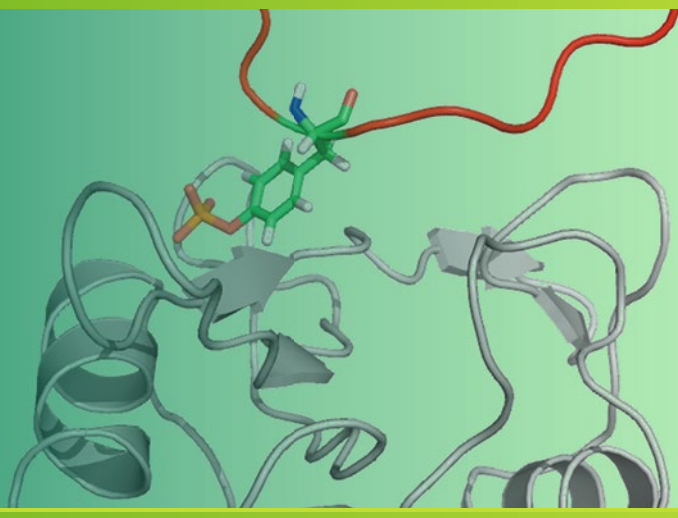
Methods in Molecular Biology 1268

Springer Protocols



Peng Zhou
Jian Huang *Editors*

Computational Peptidology

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Computational Peptidology

Edited by

Peng Zhou

Center of Bioinformatics (COBI), School of Life Science and Technology, University of Electronic Science and Technology of China (UESTC), Chengdu, China

Jian Huang

Center of Bioinformatics (COBI), School of Life Science and Technology, University of Electronic Science and Technology of China (UESTC), Chengdu, China

Humana Press

Editors
Peng Zhou
Center of Bioinformatics (COBI)
School of Life Science and Technology
University of Electronic Science and Technology of China (UESTC)
Chengdu, China

Jian Huang
Center of Bioinformatics (COBI)
School of Life Science and Technology
University of Electronic Science and Technology of China (UESTC)
Chengdu, China

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-4939-2284-0 ISBN 978-1-4939-2285-7 (eBook)
DOI 10.1007/978-1-4939-2285-7
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014958477

Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Computational peptidology is a newly emerging subfield that focuses on the use of computational, theoretical, and bioinformatic approaches to treat peptide-related problems. In *Computational Peptidology: Methods and Protocols*, expert researchers in relevant fields detail in silico methods and techniques widely used to study peptides. These include methodologies covering the database, molecular docking, dynamics simulation, data mining, and de novo design and structure modeling of peptides and protein fragments. Chapters also include the integration and application of these technologies to analyze, model, identify, predict, and design a wide variety of bioactive peptides and peptide drugs as well as peptide-based biomaterials. Written in the successful Methods in Molecular Biology™ series format, chapters include introductions to their respective topics, lists of the necessary methods and tools, step-by-step, readily reproducible protocols, and notes on troubleshooting and avoiding known pitfalls.

Chengdu, China ***Peng Zhou***
Jian Huang

Contents

Contributors

Alexandre M.J.J. Bonvin • *Computational Structural Biology Group, Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, The Netherlands*

Kumardeep Chaudhary • *Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India*

Shereen Cynthia D'Cruz • *Department of Biochemistry and Molecular Biology, School of Life Sciences, Pondicherry University, Pondicherry, India*

Marc Devocelle • *Department of Chemistry, Royal College of Surgeons in Ireland, Dublin, Ireland*

Fergal J. Duffy • *School of Medicine and Medical Science, University College Dublin, Dublin, Ireland; Complex and Adaptive Systems Laboratory, University College Dublin, Dublin, Ireland; Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland*

Richard J. Edwards • *School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia; Centre for Biological Sciences, University of Southampton, Southampton, UK; Institute for Life Sciences, University of Southampton, Southampton, UK*

Loan Ping Eng • *Department of Biochemistry, National University of Singapore, Singapore, Singapore*

Ankur Gautam • *Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India*

Sudheer Gupta • *Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India*

Mannu Jayakanthan • *Centre for Bioinformatics, School of Life Sciences, Pondicherry University, Pondicherry, India; Department of Plant Molecular Biology and Bioinformatics, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India*

Rajamanickam Jubendradass • *Department of Biochemistry and Molecular Biology, School of Life Sciences, Pondicherry University, Pondicherry, India*

Pallavi Kapoor • *Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India*

Krzysztof Kuczera • *Departments of Chemistry and Molecular Biosciences, University of Kansas, Lawrence, KS 66045, USA*

Rahul Kumar • *Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India*

Mariagrazia Di Luca • *NEST, Istituto Nanoscienze-CNR and Scuola Normale Superiore, Pisa, Italy*

Giuseppe Maccari • *Center for Nanotechnology Innovation @NEST, Istituto Italiano di Tecnologia, Pisa, Italy*

Daniela Marasco • *Department of Pharmacy, CIRPEB: Centro Interuniversitario di Ricerca sui Peptidi Bioattivi, University of Naples "Federico II", Naples, Italy*

Premendu P. Mathur • *Centre for Bioinformatics, School of Life Sciences, Pondicherry University, Pondicherry, India; Department of Biochemistry and Molecular Biology, School of Life Sciences, Pondicherry University, Pondicherry, India; KIIT University, Bhubaneshwar, India*

Adrien S.J. Melquiond • *Computational Structural Biology Group, Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, The Netherlands*

Ram V. Mohan • *Department of Nanoengineering, Joint School of Nanoscience and Nanoengineering, North Carolina A&T State University, Greensboro, NC, USA*

Giancarlo Morelli • *Department of Pharmacy, CIRPEB: Centro Interuniversitario di Ricerca sui Peptidi Bioattivi, University of Naples "Federico II", Naples, Italy*

Gautier Moroy • *Molécules Thérapeutiques In Silico, Inserm UMR-S 973, Université Paris Diderot, Sorbonne Paris Cité, Paris, France*

Riccardo Nifosì • *NEST, Istituto Nanoscienze-CNR and Scuola Normale Superiore, Pisa, Italy*

Nicolas Palopoli • *Centre for Biological Sciences, University of Southampton, Southampton, UK*

Gajendra P.S. Raghava • *Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India*

Julien Rey • *Molécules Thérapeutiques In Silico, Inserm UMR-S 973, Université Paris Diderot, Sorbonne Paris Cité, Paris, France*

Kristen L. Rhinehardt • *Department of Nanoengineering, Joint School of Nanoscience and Nanoengineering, North Carolina A&T State University, Greensboro, NC, USA*

Pasqualina Liana Scognamiglio • *Department of Pharmacy, CIRPEB: Centro Interuniversitario di Ricerca sui Peptidi Bioattivi, University of Naples "Federico II", Naples, Italy; Center for Advanced Biomaterials for Health Care, Istituto Italiano di Tecnologia (IIT),Naples, Italy*

Tongye Shen • *Department of Biochemistry, Cellular & Molecular Biology, University of Tennessee, Knoxville, TN, USA*

Denis C. Shields • *School of Medicine and Medical Science, University College Dublin, Dublin, Ireland; Complex and Adaptive Systems Laboratory, University College Dublin, Dublin, Ireland; Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland*

Johannes Söllner • *Emergentec Biodevelopment GmbH, Vienna, Austria*

Goundla Srinivas • *Department of Nanoengineering, Joint School of Nanoscience and Nanoengineering, North Carolina A&T State University, Greensboro, NC, USA*

Tin Wee Tan • *Department of Biochemistry, National University of Singapore, Singapore, Singapore*

Pierre Thévenet • *Molécules Thérapeutiques In Silico, Inserm UMR-S 973, Université Paris Diderot, Sorbonne Paris Cité, Paris, France*

Joo Chuan Tong • *Department of Biochemistry, National University of Singapore, Singapore, Singapore; Institute of High Performance Computing, Singapore, Singapore*

Mikael Trellet • *Computational Structural Biology Group, Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Utrecht, The Netherlands*

Pierre Tuffery • *Molécules Thérapeutiques In Silico, Inserm UMR-S 973, Université Paris Diderot, Sorbonne Paris Cité, Paris, France*
Guangshun Wang • *Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE, USA*
Chung F. Wong • *Department of Chemistry and Biochemistry, University of Missouri St Louis, St Louis, MO, USA*

Chapter 1

De Novo Peptide Structure Prediction: An Overview

Pierre Thévenet, Julien Rey, Gautier Moroy, and Pierre Tuffery

Abstract

Peptide structure identification is an important contribution to the further characterization of the residues involved in functional interactions. De novo structure peptide prediction has, in the past few years, made significant progresses that make reasonable, for peptides up to 50 amino acids, its use for the fast identification of their structural topologies. Here, we introduce some of the concepts underlying approaches of the field, together with their limits.

Key words Structure prediction, In silico, Structural bioinformatics, Large scale, Soluble peptides

1 Introduction

1.1 3D Modeling of Peptide Structure as a Mean to Assist and Rationalize Functional Analysis

The identification of the important residues involved in peptide stability interactions and function usually relies on well-established experimental approaches, such as directed mutagenesis or alanine scanning, that have repeatedly proven valuable in the past years. For instance, among well-known illustrations, using alanine scanning over all positions 13–40 of the 40 amino acid αβ peptide amyloid fibril, Williams and coworkers have shown that one single amino acid substitution at position 17, 27, or 34 could result in a completely unfolded peptide [1]. As well, Van Craenenbroeck and coworkers [2] have shown the important role of one particular amino acid using a similar approach to study the affinity of the interaction between the ghrelin and the recombinant human ghrelin receptor. However, these experiments are costly, time consuming. Usually, the knowledge of the 3D structure of the peptide could constitute a basis to better rationalize further exploration by focusing on the role of particular positions likely to affect peptide activity. Here again, such exploration can be undertaken using experimental techniques but also, and in a more and more efficient manner, using in silico techniques [3, 4]. Finally, in a context where the scan of complete genomes or proteomic pipelines are able to produce large amount of peptide sequences, there is also a

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_1,

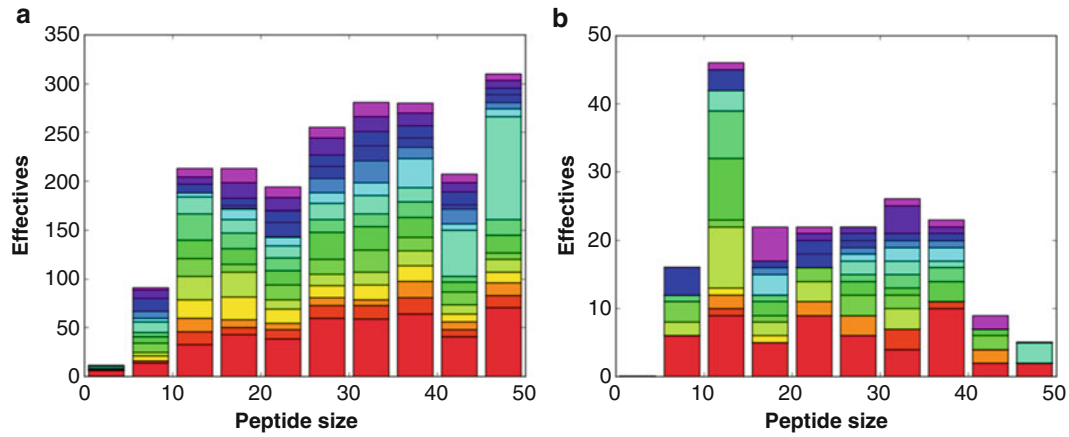


Fig. 1 Number of experimental peptide structures deposited in the Protein Data Bank per year, for (**a**) any type of peptides and (**b**) for peptides containing at least one nonnatural residue. The effective for the years before 2000 are cumulated and correspond to *red cells* (*bottom*), then each consecutive year is detailed using a color gradient up to 2013—*purple*

need to categorize and organize the amount of information available so as to focus on the candidates offering the best perspectives of development, and once more, information about the 3D structure of the sequences can be a significant contribution. Despite important progress particularly using NMR spectroscopy, the number of peptide structures experimentally solved remains low. As illustrated in Fig. 1, the number of Protein Data Bank (PDB) entries corresponding to only one asymmetric unit and of size below 50 residues, was, on date of October 1st, 2013 of only 2,055 [5]. This number falls down to 1,265 removing the sequences that have more than 30 % of sequence identity—note that this redundancy elimination is only effective for a size of sequence between 20 and 50 amino acids, hence redundancies for smaller sizes are not removed. Thus, means for the large scale and high throughput in silico prediction of peptide structure are timely.

1.2 Why Peptide De Novo Modeling Instead of Homology Modeling?

For proteins, the best option to get a 3D model, as illustrated during the Critical Assessment of Techniques for Protein Structure Prediction experiments [6], is presently comparative protein modeling, also called homology modeling. These techniques rely on the observation that proteins of similar sequences usually have a similar fold, accepting that the lower the sequence identity, the larger the error on the modeling [7]. However, those well-established approaches for protein modeling face several limitations that prevent their use at a large scale for peptides.

Firstly, as previously mentioned, the number of peptide structures that have been experimentally determined remains low. Secondly, compared to proteins, peptides usually have an increased

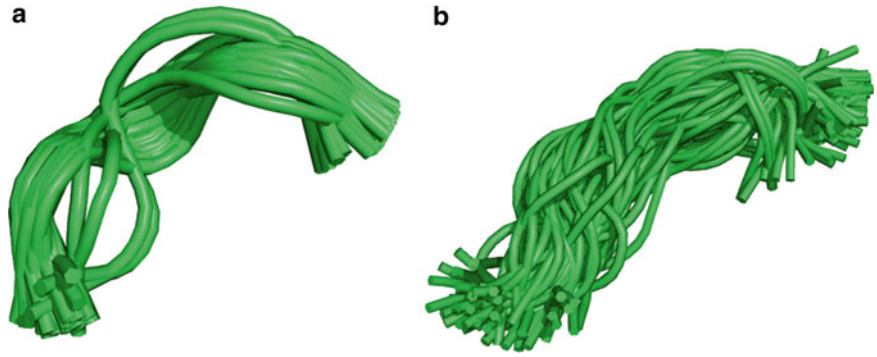


Fig. 2 Acceptable peptide conformations according experimental NMR constraints for (**a**) Met-Enkephalin 1 (1plw) and (**b**) MUC2 Muncin Domain Peptide (2li2)

Fig. 3 (**a**) α-conotoxin GI, with different disulfide bonds (1xga in *green*, 1xgb in *cyan* and 1xgc in *magenta*), (**b**) engineered spider venom peptide (4b2u in *green*, 4b2v in *cyan*) and (**c**) antibacterial designed peptides (2m2x in *green*, 2m2y in *cyan*) with different sequences but same fold

flexibility, which results from a smaller number of stabilizing interactions. Consequently, the conformations can be fuzzier, as illustrated in Fig. 2 that shows the conformations identified by NMR for two peptides of five and seven residues. As a consequence, the impact of substitutions can be more dramatic than for proteins, which makes the use of homology modeling more hazardous. Thirdly, a large part of the activity about peptides is related to peptide design experiments which largely escape the context of homology modeling. For instance, as illustrated in Fig. 3a for the alpha conotoxin GI (PDB entries 1xga, 1xgb, and 1xgc), engineering makes possible to stabilize the exact same amino acid sequence of 13 amino acids, but with different disulfide bond patterns, which result in large conformational changes [8]. Figure 3b illustrates a study by Loening and coworkers [9]. Here, sequence modifications allowed to modify a spider venom peptide by shortening a β-turn. Finally, Fig. 3c shows an example of the design of an antibacterial beta hairpin, transformed into an analogue but cyclized hairpin incorporating six nonstandard amino acids [10]. Such design clearly could not be performed using

homology techniques, and besides, as illustrated Fig. 1, the number of peptide entries of the PDB including nonstandard amino acids is of only 209 so far.

2 Approaches to De Novo Peptide Modeling

In silico approaches to identify peptide structure given its sequence can be divided into two categories. The first one corresponds to molecular dynamics (MD) simulation techniques, Nobel prized in 2013. Despite effective, these approaches remain presently limited to rather small peptide sizes due to calculation times. The second one corresponds to approaches specifically designed for peptides. They rely on knowledge based rules learnt from peptides or proteins that bias the conformational search performed to identify the best predicted structure. By contrast to MD simulations techniques, they are much faster, but usually return a limited set of conformations and they do not provide information about the thermodynamics of the peptides. Their ease of use—some of them are online—make them suitable for quick glance at peptide predicted structure, but also in the more usage intensive context of peptide design (Table 1).

2.1 Molecular Dynamics Simulation Approaches

MD simulation is based on the numerical integration of the classical Newtonian equations of motions for all the atoms in a system. The interactions between atoms are described by empirical potential functions or force fields derived from experiments and from gas-phase quantum mechanical calculations. MD simulations are a simple and accurate method for sampling the energy landscape of a system in an unbiased way.

The first 1 μs MD simulation with full representation of peptide and solvent has been performed on the villin headpiece subdomain, a 36-residue peptide [11]. Starting with an unfolded extended state, the peptide reached during 150 ns a stable state, which has significant resemblance to the native one. During one whole year; the MD simulation was computed by 256-CPU on a Cray T3E, one of the most powerful supercomputer in the world in the 1990s. In order to decrease the large number of atoms and gain computational time, several implicit solvent models can be used for MD simulations. The de novo folding of villin headpiece subdomain has been later studied by thousands of independent MD simulations started from an extended structure in implicit GB/SA (Generalized Born/Solvent Area) solvent [12]. The total simulated time exceeded 220 μs. To obtain this remarkable value, the authors used thousands of PCs distributed throughout the world. Based on these MD simulations, the folding time scale of villin headpiece subdomain was estimated to be approximately 5 μs. Another success of MD simulations was performed on the TRP

Table 1
De novo peptide structure prediction approaches, with their main features

Program	Heuristic	Model generation	Model selection	Peptide size	User constraints	Online web server
PEPstr	Secondary structure and β-turns prediction	MD simulation	Amber all-atom force field	7–25	None	http://www.imtech.res.in/raghava/pepstr
Bhageerath	Secondary structure prediction	Monte Carlo	Biophysical filters all-atom force field	<70	Local secondary structures	http://www.scfbio-iitd.res.in/bhageerath
PepLook	Generalized secondary structure 16 states	Iterative Boltzmann stochastic procedure	All-atom force field	5–30	Disulfide bonds nonstandard amino acids (offline)	http://peplook.gembloux.ulg.ac.be
PEP-FOLD	Generalized secondary structure 27 letters states	Greedy algorithm Monte Carlo	sOPEP coarse-grained force field	9–36	Disulfide bonds inter-residue contacts	http://bioserv.rpbs.univ-paris-diderot.fr/PEP-FOLD
Rosetta	3–9 residues fragments	Fragment assembly	Rosetta score	N/A	NMR constraints	http://robetta.bakerlab.org
I-Tasser	9 fragments prediction templates	Template fragment assembly	C-score	N/A	Distance restraints Templates specification	http://zhanglab.ccmb.med.umich.edu/I-TASSER/

cage. Based on MD simulations using another implicit solvent based on GB/SA model, Simmerling and coworkers were able to predict de novo the structure of a 20-residue peptide, called trp cage, prior to the release of the experimentally determined structure [13]. The predicted structure has a low 0.97 Å Cα RMSD and 1.4 Å for all heavy atoms in comparison with the experimental one. Several MD simulations of 100 ns at different temperatures were performed to converge to identical families of conformation close to the NMR-based structures.

In general however, due to limited simulation time, MD simulations often fail to sample the conformational space of protein and peptide efficiently enough so as to identify the native conformation. Replica Exchange MD (REMD) method are an attempt to address this issue. It is based on exchanges of temperatures after a given time between copies of simulated system, identical except for the temperature. One main advantage of such approach is a better sampling due to the facilitated energy barrier crossings at higher temperature. REMD has been applied successfully to investigate the structure of a five-residue peptide, called Met-Enkephalin, in implicit solvent [14] and in explicit water molecules [15]. A 23-residue has also reached, starting from an extended conformation, the folded structure using REMD in GB/SA implicit solvent [16].

Implicit representations of water and lipid molecules have been employed to study de novo folding of membrane-bound peptide. Using implicit membrane GB model combined with REMD simulations, the major pVIII coat protein of the filamentous fd bacteriophage, a 50-residue peptide, has been predicted successfully [17]. The predicted conformations and positions into membranes were consistent with experimental data from NMR spectroscopy. A similar protocol has been applied on eight peptides (16–30 residues) deriving from synthetic tryptophan-flanked transmembrane peptides with different hydrophobic lengths [18]. In agreement with experimental data, these peptides have been correctly predicted in conformations able to insert spontaneously into membrane through both analogous mechanism.

Another attempt to overcome this limitation is technological. Recently, Shaw and coworkers have designed specialized supercomputer called Anton to greatly accelerate the execution of MD simulations. In explicit solvent, Anton has been able to perform simulations as long as 1 ms. Two peptides, i.e., a variant of the villin headpiece subdomain and a variant of a WW domain, have been folded correctly, with a backbone RMSD of ~1 Å from the crystal structure [19]. Moreover, Shaw and coworkers studied 12 proteins domains that range in size from 10 to 80 amino acid residues [20]. They succeeded to predict the experimentally determined native structures for 11 of the 12 proteins.

These impressive results, while full of promises, are still to be transposed to the biological end user. Two major obstacles are to be addressed. Firstly, MD approaches presently remain rather complex to use, which makes their today's use by everybody unlikely. So far, no online resource proposing to run MD simulations to fold a peptide sequence could be setup. Secondly, their scalability to large collections of sequences remains out of reach. For instance, they do not seem be a relevant approach to process a myriad of peptide sequences, such as those expected for venom peptides [21] and short sequences in prokaryote genomes, both estimated of several millions.

2.2 Peptide Specific Approaches

The search for specific methods able to identify the native conformation of short amino acid sequences was pioneered in the 1990s, where Ishikawa and Dill proposed Geocore [22], an approach growing the polypeptidic chain using a limited number of ϕ/ψ dihedral angle combinations, which made possible to sample the possible conformations to propose several folds. Since then, several methods have gradually been developed. These methods generally consider conformational constraints predicted from the amino acid sequence so as to bias the generation of the conformations.

PEPStr [23] relies on the observation that β-turns are commonly observed in bioactive peptides. Consequently, it relies on the prediction of the secondary structure using PSIPRED [24] supplemented by a neural network to predict the β-turns [25] to propose likely ϕ and ψ angles values. Models are then generated and refined by a small MD using the AMBER6 force field. Bhageerath [26] is another approach that relies on the prediction of the secondary structure from the amino acid sequence. Secondary structures being assigned from the prediction, the conformational flexibility allowed by the loops is sampled and biophysical filters are applied to discard the less likely structures. In practice, a Monte Carlo algorithm is applied, using an all-atom force field to select 100 of conformations of low energy, from which ten best models are returned. An updated version of Bhageerath has recently been proposed [27]. The main improvement brought is the usage of a "divide and conquer" strategy, using both de novo and homology modeling prediction, which makes possible to escape pure de novo modeling for longer sizes.

PepLook [28] and PEP-FOLD [29, 30] rely on a more accurate description of local conformation, named "structural alphabet", which can be assimilated to some generalized secondary structure. Here, the structure is seen as a series of fragments of short size (in practice four or five amino acids), each of which can be assigned a limited set of canonical conformations, larger than the standard three secondary structure states (α-helix, β-sheet and random coil). This increased number of states allows to describe more accurately the conformations of the loops. The way the states

are identified and the underlying model to describe fragment geometry condition their use for de novo modeling. PepLook [28] relies on a structural alphabet of 16 canonical states that describe the conformations of fragments of five residues [31] in terms of ϕ and ψ angle values. Given a sequence, the probabilities of each state are predicted and likely pairs of ϕ/ψ values are identified. An iterated procedure then generates decoys and progressively modifies the probabilities of the pairs of ϕ/ψ values depending on the energy of the decoys, so as to promote the lowest energy conformations [32]. At convergence, the lowest energy conformation is returned. This approach has recently been updated to accept cyclized peptides, and some nonnatural amino acids [33].

PEP-FOLD [29, 30] relies on a structural alphabet that uses 27 different canonical states to describe the conformation of fragments of four residues [34] using geometrical—not angular—descriptors. Given an amino acid sequence, only the eight best predicted states are selected at each position in the sequence. Then a procedure based on a greedy algorithm is used to assemble the complete peptide [35], growing the peptide one residue by one residue, similarly to Geocore. The assembly is driven by a specific coarse grained force field [36]. Originally limited to linear 25-residue peptides, PEP-FOLD has been recently updated to larger sizes (50 amino acids) and peptides cyclized by disulfide bonds. Residue contact information can also be specified [37].

Not surprisingly, the design of the different approaches has led to an improvement in performance. Benchmarked on a dataset of 42 linear peptides of size between 9 and 20 residues, mixing peptides in aqueous and membrane environment, PepStr observed average precision on the Cα was of 4.0 Å RMSD. Bhageerath observed precision on a dataset of 50 linear peptides of size between 17 and 70 residues was of 4.1 Å RMSD, among which the subset of 26 peptides between 17 and 50 for peptides approximate by 4.2 Å RMSD the experimental conformation. PepLook initial results were of 3.8 Å RMSD over four peptides of size between 20 and 27 amino acids. Finally, over the PepStr dataset, the accuracy of PEP-FOLD was of 3.0 Å and of 2.7 Å for the subset of 15 peptides in solution.

Figure 4 illustrates how the approaches perform for four peptides of size between 17 and 47 residues, of different topologies, and for which the experimental structure has been deposited in the PDB in 2013. The peptides have been submitted to the PepStr, Bhageerath, PepLook, and PEP-FOLD servers. Only the models close enough to the experimental structure are depicted. Note that PepStr did not provide any native like model for all the queries, which is understandable due to the focus of the approach on beta hairpins. For the enterocin JSB, Bhageerath best model corresponds to a mirror topology, a topology also returned by

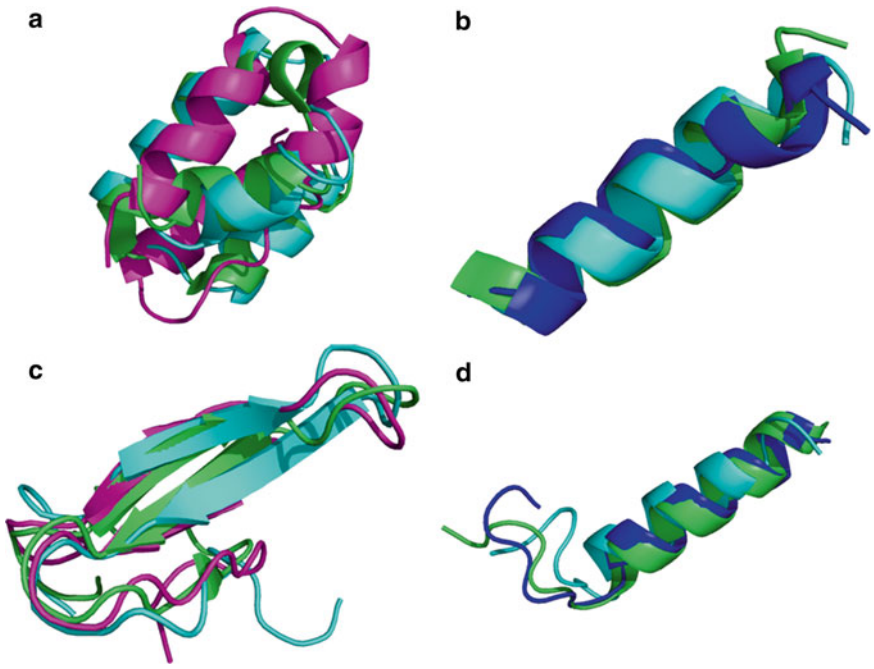


Fig. 4 Models generated for four peptides: (**a**) Enterocin JSB (2m60), (**b**) N-terminal cytosolic tail of the V-ATPase a2-subunit (2lx4), (**c**) ppiase pin1 (2m9i) (**d**) sensory box protein fragment (2yom). *Green*: experimental structures. *Cyan*, *Magenta*, *Dark blue*: models generated by PEP-FOLD, Bhageerath and PepLook, respectively. Only native like models are depicted, except for Enterocin JSB

PEP-FOLD (not shown) along with a native like topology. For the PPIase pin1, a native like model was returned by both Bhageerath and PEP-FOLD. Finally PepLook and PEP-FOLD returned a native like model for both the V-ATPase a2 subunit and the sensory box protein fragment, when Bhageerath models corresponded to broken helices.

Looking at the recent advances for peptides cyclized by disulfide bonds, PepLook and PEP-FOLD best model accuracy was of 3.8 Å and 2.7 Å, respectively, over a test set of 34 peptides having from one to three disulfide bonds [37]. However, it is important to note that the performance decreases as the number of disulfide bonds increases, the corresponding results being of 2.7 Å and 1.7 Å for peptides having one disulfide bond, and of 5.2 Å and 3.4 Å, for peptides having three disulfide bonds. Thus, it is reasonable to consider that, presently, successful predictions are only reached for a number disulfide bond of one or two.

One can ask how these peptide specific approaches compare with more generalist approaches developed for proteins such as Rosetta [38] and I-Tasser [39], although these approaches have not been specifically benchmarked for small protein sizes. Preliminary results indicate that PEP-FOLD and Rosetta perform

similarly on a test set of 56 linear peptides in solution, and of size between 25 and 52 amino acids [40], PEP-FOLD performing slightly better for α-helical topologies. On the dataset of cyclized peptides, I-Tasser average performance was of 2.5 Å, but it was not possible to assess how the existence of a template in the PDB could condition the results [33]. Thus, peptide specific approaches seem to compare favorably with generic de novo approaches developed for proteins, while being faster. For instance PEP-FOLD typical execution times are of only 40 min for peptides of size 36 amino acids. Recent results [41] would indicate it is possible to improve this processing time by a factor of 10, opening the door to very large scale peptide 3D generation.

Finally, it is also interesting to note that these approaches have been used successfully by experimentalists. For instance, PEPstr has been used in a study of the interaction between the lumen enzymes and the termini of shell proteins (17 amino acids). The helical predicted structure of the 17mer was confirmed by circular dichroism spectra [42]. PepLook has been used to study the conformational stability of fragments of 19 amino acids of the polymorphism of IL-2R receptor chains [43]. Some PEP-FOLD structure predictions of LLP peptides [44] and of the NFL-TBS.40-63 fragment [45] have been confirmed by circular dichroism measurements.

3 Present Limits and Perspectives

As illustrated in the previous sections, the two classes of peptide de novo structure prediction approaches appear to have complementary possibilities. Molecular simulation techniques now reach a high degree of sophistication, and an accuracy that make their use possible for most of the biological systems and conditions in which de novo peptide structure prediction could be necessary. These include standard but also nonstandard amino acids, aqueous but also lipidic environment, or pH dependence. They can also be employed to study assemblies of peptides, and peptide interactions with their partners, providing to some extent information about the free energy of binding, or informations about the relative stability of peptides resulting from substitutions [46]. They usually require specific skills however, and they cannot yet be applied to large scale collections of sequences due to the computational effort they require. On the opposite, more specialized approaches now provide an answer in only a few minutes. These approaches however cannot presently address the de novo prediction of all kinds of peptides. They are noteworthy limited in their ability to identify the correct fold of peptides having a large number of disulfide bonds, which prevents their large scale application in the context of venomics for instance. Due to the lack of structural information to

derive specific rules, they are also presently mostly parameterized for the prediction of the structure peptides in solution, at neutral pH, for peptides having a sequence composed of standard amino acids. For such peptides they are however able to propose in most cases a structure that corresponds to a reasonable approximation of the experimental structure. This makes them well suited to analyze larger collections of candidate sequences, as can be identified from the genome wide inspection of sequences, or to assist the design of more specific experiments.

Acknowledgements

This work has been supported by the French IA bioinformatics BipBip grant, by INSERM UMR-S 973 recurrent funding.

References

1. Williams AD, Shivaprasad S, Wetzel R (2006) Alanine scanning mutagenesis of Abeta(1–40) amyloid fibril stability. J Mol Biol 357(4): 1283–1294
2. Van Craenenbroeck M, Gregoire F, De Neef P et al (2004) Ala-scan of ghrelin (1–14): interaction with the recombinant human ghrelin receptor. Peptides 25(6):959–965
3. Vanhee P, van der Sloot AM, Verschueren E et al (2011) Computational design of peptide ligands. Trends Biotechnol 29(5):231–239
4. Audie J, Boyd C (2010) The synergistic use of computation, chemistry and biology to discover novel peptide-based drugs: the time is right. Curr Pharm Des 16(5):567–582
5. Berman HM, Westbrook J, Feng Z et al (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
6. Kryshtafovych A, Monastyrskyy B, Fidelis K (2013) CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 82(Suppl 2):7–13
7. Marti-Renom MA, Stuart AC, Fiser A et al (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325
8. Gehrmann J, Alewood PF, Craik DJ (1998) Structure determination of the three disulfide bond isomers of alpha-conotoxin GI: a model for the role of disulfide bonds in structural stability. J Mol Biol 278(2):401–415
9. Loening NM, Wilson ZN, Zobel-Thropp PA et al (2013) Solution structures of two homologous venom peptides from Sicarius dolichocephalus. PLoS One 8(1):e54401
10. Conibear AC, Rosengren KJ, Daly NL et al (2013) The cyclic cystine ladder in theta-defensins is important for structure and stability, but not antibacterial activity. J Biol Chem 288(15):10830–10840
11. Duan Y, Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 282(5389):740–744
12. Zagrovic B, Snow CD, Shirts MR et al (2002) Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. J Mol Biol 323(5): 927–937
13. Simmerling C, Strockbine B, Roitberg AE (2002) All-atom structure prediction and folding simulations of a stable protein. J Am Chem Soc 124(38):11258–11259
14. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314(1–2):141–151
15. Sanbonmatsu KY, Garcia AE (2002) Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. Proteins 46(2):225–234
16. Rhee YM, Pande VS (2003) Multiplexed-replica exchange molecular dynamics method for protein folding simulation. Biophys J 84(2 Pt 1):775–786
17. Im W, Brooks CL III (2004) De novo folding of membrane proteins: an exploration of the structure and NMR properties of the fd coat protein. J Mol Biol 337(3):513–519
18. Im W, Brooks CL III (2005) Interfacial folding and membrane insertion of designed peptides

studied by molecular dynamics simulations. Proc Natl Acad Sci U S A 102(19):6771–6776

19. Shaw DE, Maragakis P, Lindorff-Larsen K et al (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330(6002):341–346
20. Lindorff-Larsen K, Piana S, Dror RO et al (2011) How fast-folding proteins fold. Science 334(6055):517–520
21. Vetter I, Davis JL, Rash LD et al (2011) Venomics: a new paradigm for natural products-based drug discovery. Amino Acids 40(1):15–28
22. Ishikawa K, Yue K, Dill KA (1999) Predicting the structures of 18 peptides using Geocore. Protein Sci 8(4):716–721
23. Kaur H, Garg A, Raghava GP (2007) PEPstr: a de novo method for tertiary structure prediction of small bioactive peptides. Protein Pept Lett 14(7):626–631
24. Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287(4):797–815
25. Kaur H, Raghava GP (2004) A neural network method for prediction of beta-turn types in proteins using evolutionary information. Bioinformatics 20(16):2751–2758
26. Jayaram B, Bhushan K, Shenoy SR et al (2006) Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. Nucleic Acids Res 34(21):6195–6204
27. Jayaram B, Dhingra P, Lakhani B (2012) Bhageerath-targeting the near impossible: pushing the frontiers of atomic models for protein tertiary structure prediction. J Chem Sci 124(1):83–91
28. Thomas A, Deshayes S, Decaffmeyer M et al (2009) PepLook: an innovative in silico tool for determination of structure, polymorphism and stability of peptides. Adv Exp Med Biol 611:459–460
29. Maupetit J, Derreumaux P, Tuffery P (2009) PEP-FOLD: an online resource for de novo peptide structure prediction. Nucleic Acids Res 37(Web Server issue):W498–W503
30. Maupetit J, Derreumaux P, Tuffery P (2010) A fast method for large-scale de novo peptide and miniprotein structure prediction. J Comput Chem 31(4):726–738
31. Etchebest C, Benros C, Hazout S et al (2005) A structural alphabet for local protein structures: improved prediction methods. Proteins 59(4):810–827
32. Glick M, Rayan A, Goldblum A (2002) A stochastic algorithm for global optimization and for best populations: a test case of side chains in proteins. Proc Natl Acad Sci U S A 99(2): 703–708
33. Beaufays J, Lins L, Thomas A et al (2012) In silico predictions of 3D structures of linear and cyclic peptides with natural and non-proteinogenic residues. J Pept Sci 18(1):17–24
34. Camproux AC, Gautier R, Tuffery P (2004) A hidden Markov model derived structural alphabet for proteins. J Mol Biol 339(3): 591–605
35. Tuffery P, Guyon F, Derreumaux P (2005) Improved greedy algorithm for protein structure reconstruction. J Comput Chem 26(5): 506–513
36. Maupetit J, Tuffery P, Derreumaux P (2007) A coarse-grained protein force field for folding and structure prediction. Proteins 69(2): 394–408
37. Thevenet P, Shen Y, Maupetit J et al (2012) PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. Nucleic Acids Res 40(Web Server issue):W288–W293
38. Simons KT, Bonneau R, Ruczinski I et al (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl 3:171–176
39. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40
40. Thévenet P, Shen Y, Maupetit J et al (2012) Delivering the native structures of peptides from computer simulations and predicted NMR proton chemical shifts. In: Abstract of 32nd European Peptides Society Symposium, Megaron, Athens, Greece, 2–8 Sept 2012
41. Thevenet P, Tuffery P. submitted
42. Fan C, Cheng S, Sinha S et al (2012) Interactions between the termini of lumen enzymes and shell proteins mediate enzyme encapsulation into bacterial microcompartments. Proc Natl Acad Sci U S A 109(37): 14995–15000
43. Charlois Y, Lins L, Brasseur R (2011) A new in-silico method for determination of helical transmembrane domains based on the PepLook scan: application to IL-2Rbeta and IL-2Rgammac receptor chains. BMC Struct Biol 11:26
44. Steckbeck JD, Craigo JK, Barnes CO et al (2011) Highly conserved structural properties of the C-terminal tail of HIV-1 gp41 protein

despite substantial sequence variation among diverse clades: implications for functions in viral replication. J Biol Chem 286(31): 27156–27166

45. Berges R, Balzeau J, Takahashi M et al (2012) Structure-function analysis of the glioma targeting NFL-TBS.40-63 peptide corresponding to the tubulin-binding site on the light neurofilament subunit. PLoS One 7(11):e49436. doi:10.1371/journal.pone.0049436, PONE-D-12-10940 [pii]
46. Liu Z, Dominy BN, Shakhnovich EI (2004) Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential. J Am Chem Soc 126(27):8515–8528

Chapter 2

Molecular Modeling of Peptides

Krzysztof Kuczera

Abstract

This article presents a review of the field of molecular modeling of peptides. The main focus is on atomistic modeling with molecular mechanics potentials. The description of peptide conformations and solvation through potentials is discussed. Several important computer simulation methods are briefly introduced, including molecular dynamics, accelerated sampling approaches such as replica-exchange and metadynamics, free energy simulations and kinetic network models like Milestoning. Examples of recent applications for predictions of structure, kinetics, and interactions of peptides with complex environments are described. The reliability of current simulation methods is analyzed by comparison of computational predictions obtained using different models with each other and with experimental data. A brief discussion of coarse-grained modeling and future directions is also presented.

Key words Molecular potential, Metadynamics, Coarse-grained modeling, Force field, Solvation, Peptide

1 Introduction

Peptides are an important class of biological molecules made up of relatively short chains of amino acids. On their own, they perform a wide range functions in metabolic regulation and signaling [1, 2]. Additionally, they serve as simpler models for the larger and more complex proteins. Understanding the behavior of peptides is important for understanding fundamental biological processes involved in normal metabolism, as well as its perturbation in disease states. Molecular modeling of peptides is aimed at providing microscopic explanations of their function in terms of structure, dynamics and interactions with environment. Two opposing effects determine the properties of peptides in solution. First, in common with proteins, they have the ability to form specific three-dimensional structures, including α-helices, β-hairpins, and turns. Second, due to their smaller size, peptides tend to be more flexible than proteins and dynamics plays a greater role in peptide properties. Thus, the main aspects of peptide modeling are the

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_2,

characterization of the range of possible structures and predicting correct populations of the different conformers as well as the rates of their interconversions. This chapter provides a brief review of the main methods used in peptide modeling and describes a selection of interesting recent applications. An important approach not covered in this review is docking, which covers a wide range of methods that have been described elsewhere [3].

2 Peptide Structure

Peptides are linear chains of amino acids connected by peptide bonds. The most important characteristic of a peptide is its primary structure, i.e., the sequence of amino acids. The goal of modeling is to predict structure and dynamics based on sequence alone. The detailed three-dimensional structure of a peptide may be determined by specifying the values of all flexible dihedrals. The backbone conformation of each amino acid residue is defined by specification of the dihedral angles φ and ψ and the sidechain conformation by consecutive dihedrals χ_1, χ_2, …, while the peptide dihedral ω is usually taken to be in the trans conformation [4]. Due to the similarity in chemical composition, protein secondary structure elements—α-helices, β-hairpins, and turns also appear in peptides (Fig. 1). Additionally, the linear peptide chains may be cross-linked or cyclized, either by S–S disulfide bonds in a pair of cysteine residues, or through a peptide bond involving the N- and C-terminal residues.

3 Molecular Potentials and Solvation

3.1 Atomistic Potentials

The most basic approach to weighting molecular conformations is through a potential energy function U. For historical and physical reasons, most often used atomistic potentials have a similar form, of the type [5]

$$U = \frac{1}{2}\sum_{\text{bonds}} k_b\left(b-b_0\right)^2 + \frac{1}{2}\sum_{\text{angles}} k_\theta\left(\theta-\theta_0\right)^2 + \frac{1}{2}\sum_{\text{dihedrals}} k_\phi\left[1+\cos\left(n\phi+\delta\right)\right] + \sum_{\text{atom pairs}\,ij}\left\{k\frac{q_i q_j}{R_{ij}} + 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{R_{ij}}\right)^{6}\right]\right\} \quad (1)$$

where the first three terms describe deformations of molecular structure (bonded terms) and the last three represent interactions between atoms which are not chemically bonded (nonbonded terms). The quantities b, θ, φ, and R_{ij} describe the molecular structure—values of bond lengths, bond angles, dihedral angles, and interatomic distances, respectively. The remaining

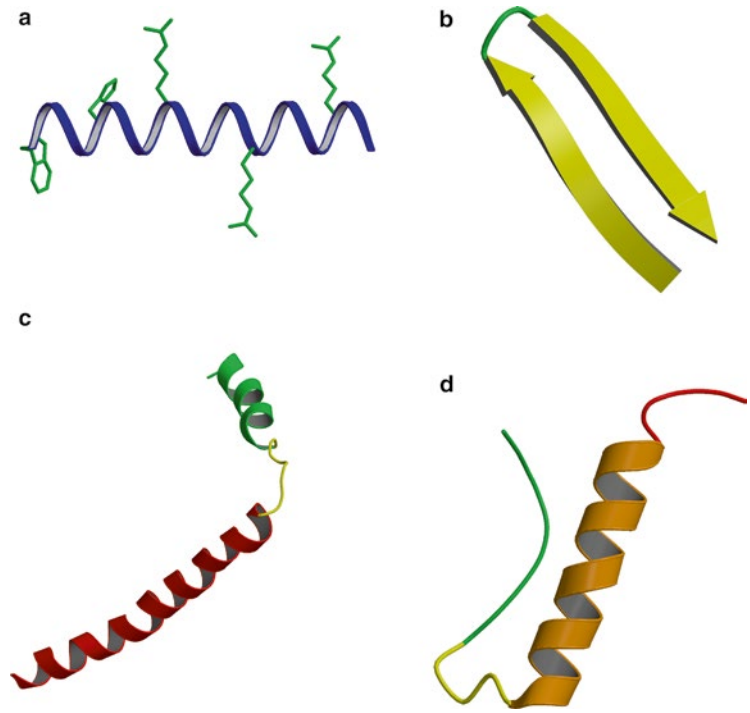


Fig. 1 Examples of peptide secondary structures. (**a**) WH21 peptide—a model-built ideal α-helix structure [93]. (**b**) The GB1 β-hairpin peptide—residues 41–56 from the second hairpin of the immunoglobulin-binding domain of streptococcal protein G [151]. (**c**) Phospholamban, a transmembrane peptide regulating calcium signaling, PDB structure 1FJK [152]. (**d**) Human peptide YY, a member of the neuropeptide Y family, PDB structure 2DEZ [153]

quantities—k_b, b_0, k_θ, θ_0, k_φ, n, δ, q_i, ε_{ij}, and σ_{ij} are parameters of the force field. The parameters have specific physical meanings: e.g., k_b and b_0 describe the stiffness and equilibrium length of chemical bonds, k_φ and n are the heights and periodicities of barriers for rotation around single bonds, q_i are atomic partial changes and σ_{ij} are van der Waals radii [5]. To speed up computations, typically further approximations are introduced. First, the pairwise interactions are included only for atoms within a cutoff distance $R_{ij} \leq R_{\text{cut}}$. Since this leads to relatively large errors for the slowly decaying Coulomb interaction $q_i q_j / R_{ij}$, a long-range correction is added to this term, most often in the form of the Ewald sum. It has been proposed that corrections to the van der Waals interactions (the last two terms in Eq. 1) should also be included. Additionally, to decrease the numbers of parameters in the force field, the so-called combination rules are introduced, which allow the calculation of van der Waals parameters for unlike atom pairs ε_{ij} and σ_{ij} from combination of values for like pairs—ε_{ii}, ε_{jj} and σ_{ii}, σ_{jj}, respectively. An often used form are Lorentz–Berthelot mixing rules [5]. Even with these approximations, evaluation of nonbonded terms remains the major contributor to the cost of U computation.

Several force field variants for peptide modeling have been developed. These differ in a number of aspects, including minor differences on the form of Eq. 1, the cutoff scheme and the method of obtaining the model parameters. Thus, AMBER [6] and CHARMM [5] use a combination of first principles quantum chemistry calculations and experimental data, while OPLS/AA [7] and GROMOS [8, 9] parameters are based solely on experimental measurements. An effect not described by the fixed-charge models in Eq. 1 is polarization, i.e., the change to a charge distribution in a molecule in response to presence of other charges. This is taken into account in polarizable force fields, such as AMOEBA [10], FLUCQ [11], or Drude-oscillator models [12], and improves treatment of highly charged systems, such as multiply charged ions, though at increased computational cost.

3.2 Solvation

Since biological processes occur in aqueous solution, including the effects of solvation is a crucial part of peptide modeling. The conceptually simplest approach is to include explicit solvent, i.e., introduce atomistic models of water molecules to solvate the peptide, with water internal deformations and interactions treated by the terms in Eq. 1. Several water models have been developed over the years, with the most popular being the TIP3P [13], TIP4P [13], and SPC [14]. Comparisons between different models [15] and special variants optimized for use with Ewald sums have been published [16]. Presence of explicit water molecules provides the most detailed description of the specific peptide–water interactions, including hydrogen bonding to charged/polar groups and the relatively weaker interactions with nonpolar parts of the solute. However, this approach is quite expensive, as typically the peptide, which is the molecule we want to model, makes up only about 10 % of the system, and most of the computational time is taken up by calculations of nonbonded interactions between pairs of water molecules. Approximate methods that help avoid this cost include implicit solvent models and coarse-grained simulations.

Implicit (or continuum) solvent models include the presence of solvent through an effective correction to the solute energy. Two approximate models that are often employed are Generalized Born (GB) [17, 18] and Poisson-Boltzmann (PB) [19]. The GB method evaluates the electrostatic part of the solvation free energy by assigning solute atoms effective radii. The PB method calculates electrostatic free energies using classical electrodynamics, by considering the solute interior to be a medium of low dielectric constant, $\varepsilon = 1-4$ and the external solvent as a medium of high dielectric constant, ca. 80 for water. An additional nonpolar contribution to the effective solvation free energy is approximated by a term proportional to the solvent accessible surface area (SASA) of the solute, of the form γ (SASA) + b, where γ and b are model parameters [20]. With the surface area correction, the Generalized

Born and Poisson-Boltzmann methods are denoted by GB/SA and PB/SA, respectively. A further simplification of solvation treatment is the introduction of purely SASA-based models, in which separate solvation parameters are assigned to individual atoms [21]. All these implicit solvent models have the advantage of speeding up calculations of energy and forces, but at the cost of introducing average, effective treatment of solvation and loss of specific solute–solvent interactions.

3.2.1 Coarse-Grained Models

The general approach taken by coarse-grained models is to extend the size of elementary building blocks beyond individual atoms to their groups and often also introduce implicit solvation. At the simplest level, we have united atom models, in which nonpolar groups (e.g., the methyl group) are treated as effective atoms, like CHARMM19 [22] or G53a6 [23], while polar groups retain their all-atom representation. A general review of coarse-grained models in computational biology may be found in [24, 25], while a review of force fields applicable to peptides and proteins is in [26]. Three examples are described here—UNRES [27], OPEP [26], and ROSETTA [28]. The UNRES coarse-grained force field uses two interaction sites per protein residue, representing the peptide group and the sidechain. The form of the potential is quite sophisticated, based on a cluster-cumulant expansion of the free energy of a solvated peptide, and the parameters are obtained based on experimental data for model systems [27]. Thanks to its careful design, UNRES has been highly successful in predicting the three-dimensional structures of peptides and proteins [29–31]. The OPEP protein force field employs an atomistic representation of the backbone and a coarse-grained one for sidechains, together with implicit solvation, and is parameterized using the observed stable structures and thermodynamic properties [26]. This CG force field correctly describes the native structures of peptides and small proteins [26]. ROSETTA uses a mostly knowledge-based force field, based on statistics obtained from structural databases [28]. The energy (or scoring function) includes low resolution terms (related to such properties as secondary structure and radius of gyration) and high resolution terms (describing effective solvation, van der Waals interactions, and hydrogen bonding). ROSETTA contains tools for a variety of tasks, including protein structure prediction, homology modeling, various docking protocols, as well as protein design and engineering [28]. The main advantage of coarse-graining is the decrease in the number of interaction centers, especially by eliminating the explicit solvent. This decreases the cost of evaluating the energy, even though the effective potential may become much more complicated than the simple Eq. 1. The second advantage is that the energy landscape is smoother in the coarse-grained picture, making conformational search faster. The disadvantage is the loss of atomic detail, such as

specific hydrogen bonding interactions, which are often only treated in an average way.

The following sections present some of the main methods that are used in potential energy based modeling of peptides.

4 Modeling Methods

4.1 Molecular Dynamics

4.1.1 Basic Algorithm

Molecular dynamics (MD) simulations involve solving of Newton's equations of motion for atoms moving in a potential U, such as given by Eq. 1. For atom i of mass m_i, the acceleration $\vec{a}_i$ is related to the force $\vec{F}_i$

$$m_i\vec{a}_i = \vec{F}_i = -\left(\frac{\partial U}{\partial x_i}, \frac{\partial U}{\partial y_i}, \frac{\partial U}{\partial z_i}\right) \tag{2}$$

where x_i, y_i, z_i are the Cartesian coordinates of the atom. The equations are integrated iteratively with a fixed time step Δt. For example, in the velocity-Verlet algorithm, the propagation of coordinates and velocities from time t to time $t+\Delta t$ is obtained by [32, 33]:

$$\begin{aligned}\vec{r}_i(t+\Delta t) &= \vec{r}_i(t) + \vec{v}_i(t)\Delta t + \frac{1}{2}\vec{a}_i(t)\Delta t^2 \\ \vec{v}_i(t+\Delta t) &= \vec{v}_i(t) + \frac{1}{2}\left[\vec{a}_i(t) + \vec{a}_i(t+\Delta t)\right]\Delta t\end{aligned} \tag{3}$$

This algorithm tells us that to calculate the new positions at $t+\Delta t$ we need the current positions, velocities and accelerations at t, and for the new velocities we need the current velocities as well as the current and new accelerations. The need to evaluate the accelerations (obtained from forces, i.e., derivatives of the potential) at every step is the main computational cost of MD simulations. The outcome of an MD simulation is the trajectory—a record of positions (and possibly velocities) of all the atoms at a set of time points 0, Δt, $2\Delta t$, …. This allows for a complete description of the system from the point of view of classical mechanics, including sampled structures and their populations, types of motions and their amplitudes and time scales, geometrical, energetic, thermodynamic, and kinetic description of the system.

4.1.2 Time Scales

MD simulations involve three basic time scales: the integration time step Δt, the total simulation time t_{sim}, and the time scale of the process of interest τ_r. In order to describe the correct motion of the particle on the given potential U, Δt needs t be small enough so that changes in the forces over one step are small. The highest frequency motions in organic molecules are stretches of C–H, N–H, and O–H bonds, with a frequency of ca. 3.000 cm^{-1}, corresponding to a period of 20 fs. This leads to the requirement that $\Delta t = 1$ fs

(1 fs = 10^{-15} s) for stable trajectories at room temperature. It is generally believed that the motions at fs time scales do not have biological significance, so several approaches such as SHAKE [34, 35], RATTLE [36], SETTLE [37], and LINCS [38] have been developed for constraining (i.e., fixing at equilibrium values) of the fastest degrees of freedom. With constraints on all bond lengths, or at least on bonds involving the lightest H atoms, stable trajectories with $\Delta t = 2$ fs can be obtained. A separate approach to increasing the effective length of Δt is multiple time stepping (MTS) [39, 40]. MTS takes into account the observation that there are a small number of energy terms that vary quickly with time—e.g., internal deformations and short-range nonbonded interactions, while a majority of the terms are relatively slowly varying long-range interactions between atom pairs that are not nearest neighbors. In this scheme we can update the fast forces every 1 fs, and the slow forces every 4 fs, yielding significant savings in computational effort. The total simulation time determines the computational cost, as the number of force evaluations is roughly $t_{sim}/\Delta t$. We thus need about 10^6 steps to generate a 1 ns trajectory and about 10^9 steps for a 1 μs simulation. The simulation length t_{sim} determines what kind of processes we can sample. Examples of process time scales τ_r are 1 ns for a single hydrogen bond formation, 100 ns for folding of a small α-helix and 10 μs for folding of a small β-hairpin, 1 ms for folding of a small protein. In order to model these events with MD simulations, we must have at least $t_{sim} > \tau_r$ and preferably $t_{sim} >> \tau_r$. Thus, although direct atomistic MD simulations yield the most complete information about the behavior of the studied system, including structure, dynamics and thermodynamics, MD is very expensive to pursue for large systems over long time scales. Current peptide simulations typically involve systems of 5,000–10,000 atoms over 100 ns–10 μs time scales. For larger systems or longer times, specialized algorithms have been developed that are more efficient.

4.2 Free Energy Simulations

In free energy simulations (FES) the object is to calculate directly changes in free energy due to various molecular processes, such as conformational change, folding, binding or mutation. The advantages of FES methods are calculation of experimentally relevant quantities, the ability to describe slow processes and relatively low statistical errors. The shortcomings are that kinetics may only be indirectly obtained, and the difficulty in obtaining converged results. Several approaches aimed at specific kinds of processes are briefly described, without detailed derivations, which require delving into statistical mechanics [41].

4.2.1 Coupling Parameters and Mutation

We consider two states of the studied system, the initial state 0 (wild type) with potential energy U_0 and the final state 1 with potential U_1 (mutant). We introduce the coupling parameter λ and the hybrid potential $U(\lambda)$, which has the property that $U(0) = U_0$

and $U(1) = U_1$. In the simplest linear coupling approach the equation used is [42, 43]:

$$U(\lambda) = (1-\lambda)U_0 + \lambda U_1 \tag{4}$$

The free energy difference between the initial and final states $\Delta G = G_1 - G_0$ may then be obtained from a series of simulations with intermediate values of λ. In the thermodynamic perturbation method (TP) partial free energy changes along the path from $\lambda = 0$ to $\lambda = 1$ are found:

$$\Delta G\left(\lambda' \to \lambda''\right) = -kT\left\langle \exp\left(-\frac{U(\lambda'') - U(\lambda')}{kT}\right)\right\rangle_{\lambda'} \tag{5}$$

where T is the absolute temperature, k is the Boltzmann constant and the angled brackets denote an average over a sample of conformations corresponding to thermal equilibrium for the system with $\lambda = \lambda'$. This sample is typically generated by a molecular dynamics simulation for the hybrid system with $U = U(\lambda')$. The total free energy change is then evaluated as the sum of contributions that span the full range of λ, from 0 to 1. In the thermodynamic integration approach, the derivative of G with respect to λ is obtained [44]

$$\frac{\partial G}{\partial \lambda} = \left\langle \frac{\partial U}{\partial \lambda}\right\rangle_{\lambda} = \langle \Delta U_{\lambda} \rangle \tag{6}$$

where the first equality is general and the second true only for linear coupling, and involves the averaging of the potential energy difference $\Delta U = U_1 - U_0$ over the sample of structures corresponding to thermal equilibrium for $U(\lambda)$. The overall ΔG is then obtained by numerical integration of the derivative from $\lambda = 0$ to $\lambda = 1$. An important property of the TI method is that ΔG is a linear function of the potential energy, opening the possibility for decomposition of decomposing the overall free energy change into contributions from different parts of the system—solute, solvent, individual residues, or terms in the potential—internal deformation, electrostatic, van der Waals. This leads to insights into the microscopic effects that contribute to the observed free energy changes. It must be noted that only the overall ΔG is an observable physical quantity and the decompositions are only qualitative aids in our understanding [45].

Because classical simulations do not include the energetic effects of changes in covalent structure upon mutation, actual simulations typically involve thermodynamic cycles of the type shown in Fig. 2, in which the difference $\Delta\Delta G$ between free energy changes in two processes is evaluated [46]. Recent applications typically use more complex schemes, with separate switching of different terms in the potential and corrections for changes in the number of atoms [47].

a

$$\begin{array}{lcl} & \overset{\Delta G_{b1}}{\rightarrow} & \\ \mathbf{Protein} + \mathbf{L}_1 & & \mathbf{Protein{:}L}_1 \\ \Delta G_{free} \downarrow & & \downarrow \Delta G_{bound} \\ \mathbf{Protein} + \mathbf{L}_2 & \underset{\Delta G_{b2}}{\rightarrow} & \mathbf{Protein{:}L}_2 \end{array}$$

$$\Delta\Delta G = \Delta G_{free} - \Delta G_{bound} = \Delta G_{b1} - \Delta G_{b2}$$

b

$$\begin{array}{lcl} & \overset{\Delta G_{b1}}{\rightarrow} & \\ \mathbf{Protein}_1 + \mathbf{L} & & \mathbf{Protein}_1\mathbf{{:}L} \\ \Delta G_{free} \downarrow & & \downarrow \Delta G_{bound} \\ \mathbf{Protein}_2 + \mathbf{L} & \underset{\Delta G_{b2}}{\rightarrow} & \mathbf{Protein}_2\mathbf{{:}L} \end{array}$$

$$\Delta\Delta G = \Delta G_{free} - \Delta G_{bound} = \Delta G_{b1} - \Delta G_{b2}$$

Fig. 2 (**a**) Thermodynamic cycle for calculating difference of binding free energies between two ligands (e.g., two different inhibitors) of the same protein. (**b**) Thermodynamic cycle for calculating difference of binding free energies between two proteins (e.g., wild type and a mutant) and the same ligand. The *horizontal arrows* represent physical processes for which experimental measurements may be made and the *vertical arrows* represent alchemical transformations, which may be calculated by free energy simulations. The quantity $\Delta\Delta G$ may be obtained both ways

4.2.2 Conformational Processes

The basic tool of conformational free energy simulations is umbrella sampling (US) [48, 49]. This approach explores the free energy changes along a reaction coordinate Z through a series of constrained simulations which use the biased potential $U^* = U + U_c$, where U is the molecular potential and U_c is a harmonic constraint:

$$U_c = \frac{1}{2} f \left(Z - Z_0 \right)^2 \tag{7}$$

Each individual simulation thus samples a neighborhood of its equilibrium position Z_0, generating the biased probability distribution $P^*(Z)$ for Z in that neighborhood, corresponding to

potential U^*, which is related to the biased potential of mean force (PMF) $G^*(Z) = -kT\ln(P^*(Z))$, and finally, recovering of the unbiased PMF $G(Z)$:

$$G(Z) = -kT \ln P^*(Z) - U_c + C \tag{8}$$

where C is an arbitrary constant. The PMF in the Z range of interest may be recovered by adjusting the values of the constants in regions where the different simulation windows overlap. An automated approach, the weighted histogram analysis method (WHAM), has been developed to calculate the full range of $G(Z)$ making optimal use of all the windows [50].

Conformational free energies may also be explored with a variant of the TI method, in which the parameter λ represents a conformational coordinate, such as distance between two atoms or a dihedral angle. Simulations at fixed values of λ are used to calculate the derivative of free energy with respect to λ, analogous to Eq. 6 but more generally including a correction for the coordinate transformation between λ and Cartesian coordinates [51, 52].

Metadynamics [53] and related techniques such as the adaptive biasing potential method [54] present a different approach to conformational free energy simulations. Metadynamics starts with selection of one or more collective variables—functions of coordinates that represent the important structural changes in the system. A simulation is then performed in which a positive biasing potential, or "hill," is added to repel the system from already explored regions of the collective variable space. At the end of the simulation, when the full range of coordinates has been explored, the biased free energy landscape is flat and the unbiased potential of mean force of the system is obtained by inverting the sum of the Gaussian constraints.

4.2.3 Binding Free Energies

The binding free energy ΔG_b is defined as the free energy change for the formation of a solvated complex P:L from the separate solvated components P and L:

$$\mathrm{P(aq)} + \mathrm{L(aq)} \rightarrow \mathrm{P:L(aq)}$$

$$\Delta G_b = G(\mathrm{P:L}) - G(\mathrm{P}) - G(\mathrm{L})$$

Due to potential applications in drug design, ΔG_b calculation has garnered significant interest in the scientific community, and a number of methods have been developed to estimate this important quantity [41, 55]. Initial studies of binding free energies employed the coupling parameter approach and thermodynamics cycles, such as Fig. 2. These methods directly yield values of $\Delta\Delta G_b$, the changes in the binding free energy ΔG_b due to changes in protein or ligand.

More recently, the double-decoupling method has been developed, which directly yields the binding free energy as [56–58]

$$\Delta G_{\mathrm{b}} = \Delta G^{*}(\mathrm{P:L}) - \Delta G^{*}(\mathrm{L})$$

where ΔG^{*}(P:L) and ΔG^{*}(L) are the free energies of decoupling, i.e., switching off interactions between the ligand and its environment in the P:L complex and in solution, respectively. Recently, an alternative approach has been developed to calculate the binding free energy from potentials of mean force obtained with appropriate restraints [56]. It has been suggested that double decoupling should be applied for ligands in deep binding pockets, while the PMF approach is preferred for surface binding. The double decoupling and PMF approaches are currently the most accurate and also most complex and expensive ways of obtaining binding free energies.

A widely used and intuitively appealing approximate approach to binding free energies is the MM-PBSA method developed in the Kollman group [59, 60], in which

$$\Delta G_{\mathrm{b}} = \langle \Delta U_{\mathrm{MM}} \rangle + \langle \Delta G_{\mathrm{PBSA}} \rangle - \langle T\Delta S \rangle$$

Here ΔU_{MM} is the molecular mechanics energy difference between the P:L complex and the free protein (P) and ligand (L), ΔG_{PBSA} is the analogous change in solvation free energy obtained using the PB/SA methods (see above), while $T\Delta S$ represents an entropic contribution, typically using the harmonic approximation. The angled brackets <…> denote averaging over a sample of representative conformations, such as from an MD simulation. A variant of this approach, called MM-GBSA differs by the use of the Generalized Born solvation model rather than PB.

The Linear Interaction Energy (LIE) method is a more simplified approach to ΔG_{b}, using the relation [61]

$$\Delta G_{\mathrm{b}} = \beta\left(\langle U^{\mathrm{elec}}_{\mathrm{bound}} \rangle - \langle U^{\mathrm{elec}}_{\mathrm{free}} \rangle\right) + \alpha\left(\langle U^{\mathrm{vdW}}_{\mathrm{bound}} \rangle - \langle U^{\mathrm{vdW}}_{\mathrm{free}} \rangle\right) + \gamma$$

where α, β, γ are empirical parameters obtained by fitting to a training data set; elec and vdW denote the electrostatic and van der Waals components of the potential energy; bound and free denote values obtained for the complex and the free ligand; and <…> denotes averaging over a representative conformational sample.

4.3 Replica Exchange

The development of the replica-exchange method has led to significant progress in enhancing the rate of sampling conformations in molecular simulations [62, 63]. This method has now become a standard component in the toolbox of molecular modeling. To simplify discussion, the temperature replica-exchange molecular dynamics (REMD) approach is discussed here. The basic idea is to

concurrently propagate several trajectories, or replicas of the system, each one at a different temperature T_i, $i=1,\ldots,$ NR. Periodically, exchange attempts are made between neighboring replicas, with the acceptance probability depending of the difference in temperatures of the replicas and the energies of the conformers being compared, based on the principle of microscopic reversibility. We can thus think of the time evolution of conformations in a REMD simulation as a combination of Newtonian dynamics along the replicas and a random walk between replicas. As a result of the exchange, states with lower relative energies explored at the higher temperatures are introduced into the low-temperature replicas. The conformational sampling rate at the lower temperatures is thus enhanced. Additionally, each replica tends to sample conformers according to the Boltzmann distribution, so the method may be applied to study of temperature dependence of physical properties—such as peptide melting curves.

Replica-exchange methods are a topic of intensive current research and numerous variants of the method are in use. The original application involved Monte Carlo simulations, rather than MD within the replicas. Multiple trajectories are propagated at each temperature in multiplexed REMD [64]. Other properties than temperature may be used as the basis of exchange in Hamiltonian replica exchange [65]. Theoretical properties of the method have been extensively explored [66]. The method is quite elegant and useful but has several drawbacks. First, it only predicts thermodynamics, with no kinetic information. Second, while the conformational sampling is enhanced compared to a single low-temperature MD trajectory, the acceleration appears to be limited to about an order of magnitude. Third, the cost of the method becomes prohibitive for large systems, due to a confluence of unfavorable factors. Because overlap of energy distributions sampled by neighboring replicas is needed for effective exchange, a larger number of replicas NR are needed for systems with more atoms. For these systems the cost per unit time of each replica propagation is also higher, as is the trajectory length needed to explore representative conformers. Finally, the time for the random walk to traverse the set of replicas also increases with NR. All these factors make it very expensive and inefficient to use replica-exchange for systems of more than 10,000 atoms or so, thus limiting applications to small solvated explicitly peptides. However, the same factors make it highly efficient to use REMD for implicit solvent models and even more so for coarse-grained simulations.

4.4 Accelerated Sampling

To increase the rate of sampling of rare events compared to direct MD, a number of approaches have been proposed, of which several are briefly discussed. Hyperdynamics, introduced by Voter, uses deformation of the potential energy surface by elevating the low-energy regions, effectively lowering the heights of barriers

separating neighboring conformations, leading to speedup of transitions [67]. Variants of this approach have recently been implemented in the NAMD and AMBER software packages [68–70]. A simulation of the 58-residue protein BPTI showed that a 500 ns aMD trajectory [69] explored the same conformational range as a 1 ms MD trajectory generated on the specialized Anton computer system [71]. A more general approach to deforming potential energy surfaces is the idea of generalized-ensemble simulations, where the attempt is to achieve a flat energy landscape, without barriers that slow down transitions [72]. Self-guided dynamics is another approach in this class, where the transitions are eased by modifying the forces acting in the system to include information about average forces from previous simulation steps [73]. A different approach to the problem is taken by Parallel Replica Dynamics (PRD), in which a series of trajectories with different initial velocities is started in parallel in a single energy basin. When the first transition to a neighboring basin/conformer is detected, all replicas are re-initialized in the new basin, effectively accelerating transition rates [74].

4.5 Transition Paths and Kinetic Networks

The methods discussed in this section all aim to increase the rate of sampling rare events through special algorithms that focus on describing transitions. The different path methods are based on transition state theory of chemical reaction rates. The basic transition path sampling method (TPS) generates a statistical ensemble of transition paths between two states (such as a folded and unfolded peptide) by perturbing existing paths [75]. The transition rate constant is calculated by averaging over this ensemble. Further developments are transition interface sampling (TIS) and the forward flux methods, in which the transition is divided into stages and the full rate constant is obtained by following transitions between the intermediates [76, 77]. Such path methods are rather involved, but they have been successfully applied in peptide simulations, e.g., the study of beta-hairpin folding [107].

The methods of Markov State Models (MSM) [78] and Milestoning [79] move beyond considering simple reaction paths and generate kinetic networks that describe the dynamics of complex processes. In the directional Milestoning method [80], the starting point is the generation of a set of representative structures of the system, called anchors. In simple cases these anchors are generated as a grid, in larger spaces they are based on an extensive previous simulation, such as REMD. Next, the milestones are introduced as the interfaces (hypersurfaces) separating the domains of the neighboring anchors. A series of n_α MD trajectories are then initiated at each milestone α, of which $n_{\alpha\beta}(t)$ reach milestone β during time t, allowing the calculation of the probability that a trajectory initiated at α will reach β at time t as $K_{\alpha\beta}(t) = n_{\alpha\beta}(t)/n_\alpha$. From the kinetic matrix $K_{\alpha\beta}$ it is possible to calculate the mean first

passage time for the overall process and its different stages, the stationary flux and the stationary probability distribution (potential of mean force). The Milestoning approach has a number of attractive properties. The underlying MD is atomistic, providing a high-level description of system motions. The kinetic network is defined in the coarse-grained space of milestones and its analysis provides a simplified and insightful description of the mechanism of the process. Breaking up the overall transition into small fragments speeds up the computation for both activated-type and diffusive-type processes. The elementary transitions between pairs of milestones are typically quite fast, on the ps–ns time scale. Running multiple short independent trajectories is highly parallelizable on modern computer systems. Typically, 10–1,000 interfaces may be explored with hundreds of trajectories initiated per interface in a total simulation time shorter than a single transition between reactant and product states [81].

The Markov State Modeling (MSM) method takes a somewhat different approach to generating a kinetic network [78, 82]. The starting point is a sampling of the system conformational space, typically by a long MD trajectory. This generates a set of explored conformations, the microstates, and also probabilities of transitions between pairs of microstates. This stationary short-term transition matrix is then used to characterize the long-term dynamics of the system. Further, clustering of the microstates is performed to generate a coarse-grained network of transitions between metastable intermediates and this network is analyzed to determine the mechanism of the overall complex process. Connections between MSM predictions and results of kinetic experimental measurements have been analyzed in detail [82]. Recently, EMMA, an automated procedure for generating MSMs has been introduced [83].

5 Examples of Applications

5.1 Conformational Exploration

Early simulations of fundamental properties involved nanosecond-length MD trajectories exploring the conformational space of small peptides. These included the alanine dipeptide [84], used as a test for empirical force fields, and other short peptides such as Ala_3, Val_3, or Tyr-Pro-Gly-Asp-Val [85]. For pharmaceutical applications, effects of cyclization on preferred conformations were explored for the enkephalin DPDPE [86] and for RGD peptidomimetics and prodrugs [87, 88]. More recently, microsecond-length MD simulations of short peptides have become possible, leading to characterization of the full conformational landscapes of larger systems, such as Ala_5 [89, 90]. MD trajectories of 2 μs length were generated for angiotensins I and II, short peptides involved in blood-pressure regulation [91]. A 4 μs simulation of the 36-residue neuropeptide hPYY reproduced the observed structural features and predicted

the major types of expected fluctuations that could be used to explain the effects of peptide modifications [92]. REMD simulations have been applied to characterize conformations of medium-size peptides such as Ala_{21}, the Fs peptide, and WH21 [64, 93].

5.2 Structure Formation

Secondary structure is a crucial feature of peptides and proteins. Predicting secondary structure and the mechanism of structure formation has been the focus of significant modeling efforts. Beginning simulations using direct MD, free energy simulations and elevated temperatures focused on early events of α-helix formation/breaking [94–97]. Later, more extensive MD simulations were performed with implicit solvent models, indicating the presence of several stages of helix folding for model peptides based on $(AAQAA)_n$, $(AAKAA)_n$, and $(AARAA)_n$ motifs [98–100]. The application of REMD allowed calculation of melting curves for helices in explicit solvent and analysis of their conformational landscapes [93, 101]. REMD results with modern force fields showed generally good agreement of calculated helix content with experimental measurements and also presented a complex statistical pathway for helix folding, initiated at several locations along the chain, the presence of a characteristic "off-center" single-helix intermediate and ending with folding of the peptide termini, as shown in Fig. 3a [93]. Recent progress in peptide design and characterization has led to the discovery that even quite short peptides with terminal blocking groups, such as ac-A_5-NH_2 or ac-WA_3H-NH2 (WH5), can take on helical conformations, Fig. 3b [90, 102]. For these short peptides very good agreement with experimentally observed helix content was found in MD and REMD simulations using several force fields [90, 93].

Analogous simulations have also been performed for β-hairpins—including simplified models [103, 104], explicit solvent REMD [105, 106], and transition path sampling [107]. This work has led to formulation of several possible models for β-hairpin formation. In the zipper model folding starts at the turn, followed by formation of consecutive backbone hydrogen bonds and establishment of sidechain contacts. In the hydrophobic collapse model the folding starts with the formation of a compact structure with native sidechain contacts, while in the broken zipper model the establishment of the turn is followed by formation of loose hydrophobic contacts, then the hairpin hydrogen bonds and finally the native hydrophobic core [30].

Thanks to its careful design, the UNRES coarse-grained model has been highly successful in predicting the three-dimensional structures of numerous peptides and proteins [29–31]. The small number of interaction sites and implicit solvation, coupled with the multiplexed REMD simulation approach make this a very effective tool for protein and peptide modeling [31, 108].

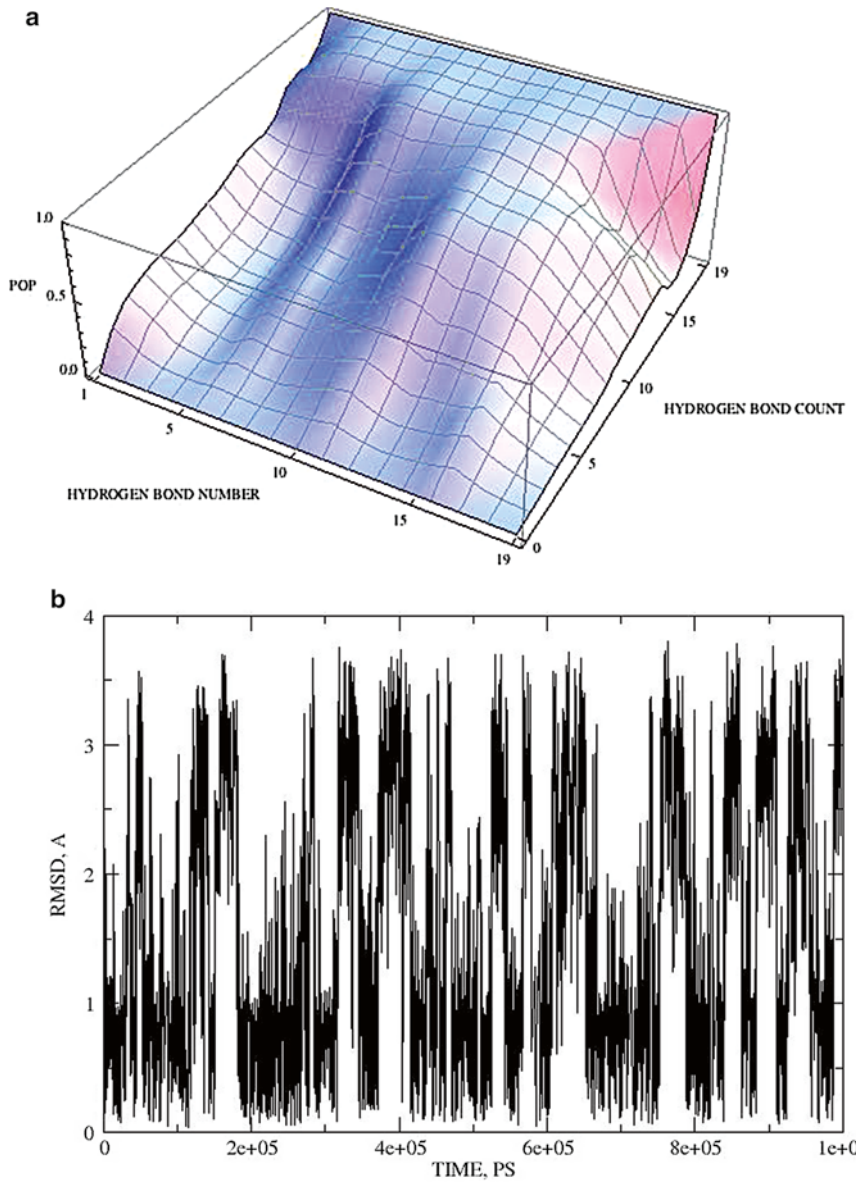


Fig. 3 (**a**) Statistical folding pathway for the WH21 α-helical peptide showing presence of "off-center" intermediate involving hydrogen bonds 11–16. The figure shows populations of individual helical hydrogen bonds, numbered 1–19, as a function of the total number of hydrogen bonds present, from 0 (in coil) to 19 (in fully helical state). (**b**) Root-mean-square deviations from ideal helix in a 1 μs MD simulation of the pentapeptide WH5 showing multiple folding and unfolding events. Both figures reprinted with permission from ref. [93]. Copyright Taylor and Francis, 2012

Analogously, the OPEP CG force field enables generation of long-time stable MD trajectories of folded proteins and correct REMD predictions of structures of α-helical and β-hairpin peptides and small proteins [26].

5.3 Kinetics

While kinetics on ms time scales has been simulated for small proteins on specialized computer systems [109], most peptide simulations are currently in the 100 ns–10 μs range. Thus folding of medium size helices, observed to occur at ca. 300 ns time scale at room temperature [110] is within reach of direct MD, while folding of β-hairpins, observed at ca. 6 μs, is at the upper limit of range [111]. Numerous simulations so far have been made with implicit solvent, which yield unrealistic time scales, but produce interesting qualitative insights. Thus, much faster folding of helices compared to hairpins has been reproduced in implicit solvent [112, 113], as has the observation of significant differences of relaxation times of individual hydrogen bonds along the peptide chain [114]. Recent high-resolution kinetics measurements have found two relaxation times of ca. 1 and 5 ns for the WH5 pentapeptide [102], sparking renewed interest in explicit-solvent MD of helix folding kinetics. These time constants have been well reproduced in MD simulations of WH5, with the longer time assigned to overall helix formation and the shorter to formation of the central hydrogen bond [115]. Analogous time scales for WH5 were found using directional Milestoning, which additionally indicated two folding pathways—a direct formation of the full helix and folding through an intermediate involving the central and N-terminal hydrogen bonds, Fig. 4 [115]. Milestoning simulations for the larger WH21 found that an unfolding initiated at the N-terminal corresponded to a folding time similar to experimental observations, while paths starting at C-terminal and center of peptide led to much longer relaxations [116]. For the 16-residue GB1 hairpin a 5 μs folding time was found with transition path sampling, in excellent agreement with experimental data [107].

5.4 Complex Environments, Binding and Interactions

A number of simulations have pursued description of peptide behavior related to biological activity. REMD in implicit membrane model based on the Generalized Born approach was used to suggest that peptides of the WALP and TMX families bind to membrane surface and form partially helical structures before insertion [117]. In contrast, explicit solvent REMD suggested that the surface-bound helix is a trap and that insertion of the unstructured peptide followed by folding is the favored pathway [118]. MD simulations were employed to characterize bending, flexing, and torsional motions of the two domains of phospholamban, a membrane-embedded peptide which regulates the SERCA calcium pump, Fig. 5a [119, 120]. Of special interest is the MD study of the HIV TAT peptide, which is involved in translocating cargoes across cell membranes [121]. This work proposed an interesting mechanism of TAT action that involves interaction of charged residues with lipid phosphate groups and then insertion of long positively charged side chains to form transient pores. Similar behavior was found for the model peptide Arg_9 [122]. A review of

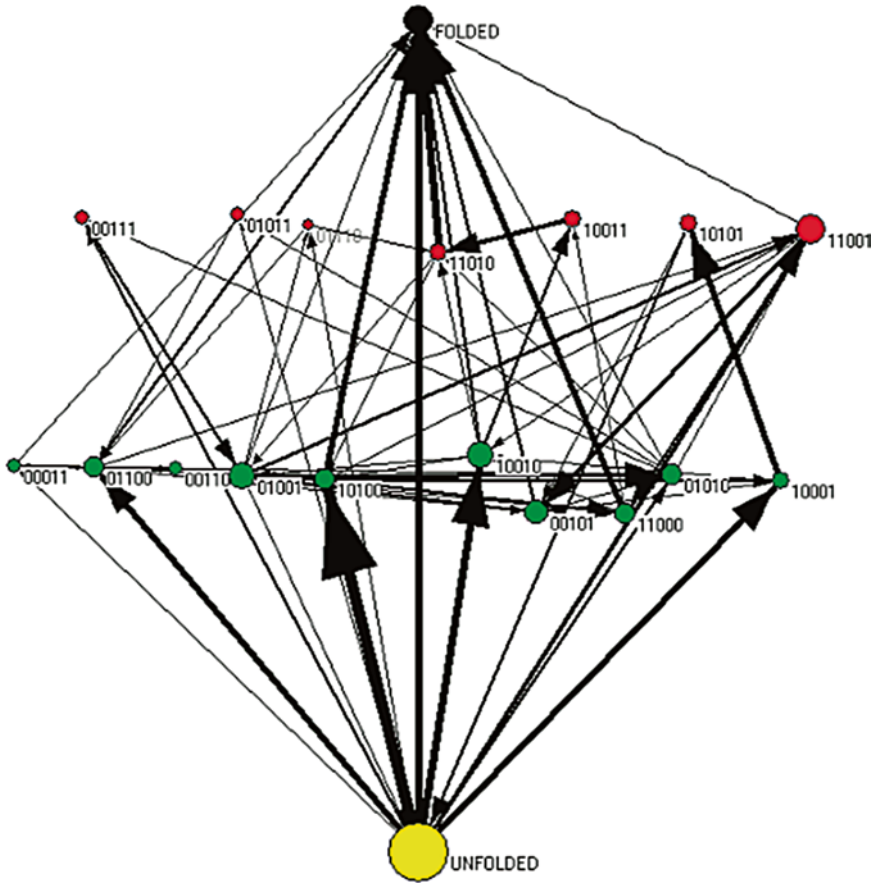


Fig. 4 Network representation of structural transitions of WH5 pentapeptide based on a Directional Milestoning simulation with 90 anchors. The symbols denote the structure of each residue, e.g., "11001" is the state in which residues 1, 2, and 5 are folded (in helical region of Ramachandran map), while residues 3 and 4 are unfolded. *Arrows* denote direct transitions between states, with *thicker arrows* corresponding to larger flux. Reprinted with permission from ref. [115]. Copyright American Chemical Society, 2012

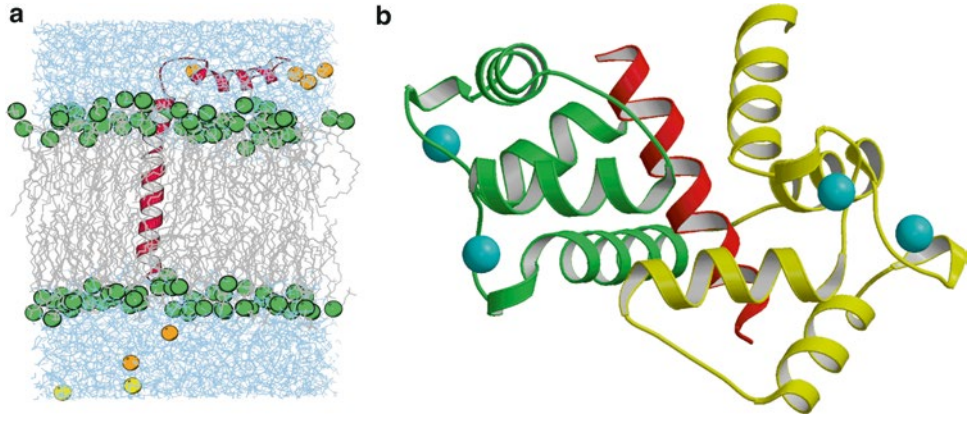


Fig. 5 (**a**) Snapshot from MD simulation of phospholamban (PLB) in a membrane bilayer. PLB in *red*, membrane headgroups in *green*, tails in *grey*, water in *blue*, Cl– ions in *orange*, Na+ ions in *yellow*. Reprinted from ref. [119]. (**b**) Interactions of calmodulin with a peptide derived from the smooth muscle myosin light chain kinase (smMLCK), PDB structure 1CDL [154]

simulations of antimicrobial and cell-penetrating peptides is available [123]. Protein–protein interactions play an important role in regulation and signaling. For calmodulin, a small protein that binds a wide range of targets in response to changes in calcium ion concentration, the free energies of binding have been estimated for interactions with the smMLCK peptide (Fig. 5b) and peptides derived from Death Associated Protein kinases, using approaches of the MM/PSBA type showed the importance of hydrophobic interactions for ligand recognition [124, 125]. The binding free energy of the phosphorylated YEEI peptide to the SH2 domain of LcK kinase was calculated using an approach based on a potential of mean force calculation with restraining potentials [56, 126]. Analyses of interactions of the drug gleevec with two kinases, Abl and c-Src, suggested that the solvent accessibility of the DFG motif of Abl is responsible for its effective inhibition by the drug [127]. A significant simulation effort has been devoted to studies of peptide aggregation. A recent review discusses modeling of Alzheimer's β-peptide (Aβ), which forms aggregates that are linked to symptoms of the neurodegenerative Alzheimer's disease [128]. Simulations were able to explore the structural flexibility of Aβ, its ability to complex with short hydrophobic peptides, the formation of fibrils from Aβ fragments and interactions with aggregation inhibitors [128, 129]. In a related work, the aggregation of a 7-residue sequence from the prion protein was investigated by long-term MD with the OPEP coarse-grained force field, showing the formation of a nucleus of 4–6 peptides followed by a growth phase with definite polymorphism of structures [130].

5.5 Modeling of Experimental Observables

Besides structure, dynamics time scales and binding free energies discussed above, a wide range of experimental observables have been simulated for peptide systems. Among thermodynamic properties modeled are melting curves, which typically do not yield T_m values or shapes similar to experimental data when obtained using all-atom models [4, 64, 90], with better agreement obtained by CG methods [26, 27]. Simulations of a wide range of NMR spectroscopic features have been reported [131–133]. Fluorescence anisotropy signals, describing molecular reorientations, were simulated for aromatic peptides and their models [134], as well as for angiotensins [91]. Generally, simulations tend to overestimate the reorientation rates of solutes in water due to underestimation of viscosity by popular water models [91], while motions in nonaqueous solvents are in excellent agreement with observations [135, 136]. MD simulations of the YGGWL enkephalin peptide showed that while average calculated donor-acceptor distances were in excellent agreement with Förster Resonance Energy Transfer (FRET) data, the underlying distance distributions were very wide, reflecting a wide range of sampled conformers, with some dependence of sampled orientations on the distance [137].

5.6 Force Field Dependence of Modeling Results

A wide range of experimental data, structural, thermodynamic, and kinetic, are used as input in force field parameterization and also for validation, i.e., checking of model reliability. However, the most interesting simulation results are those that provide complementary information that is difficult to obtain from experiments. Results falling into this category are the microscopic mechanisms of observed effects—conformational distributions of flexible molecules (e.g., the ensemble of unfolded states), detailed pathways of structural transitions (e.g., helix folding). An important way of accessing the reliability of model predictions is to check for their independence of the employed force field. Over the recent years, as MD simulations have been extended beyond single nanoseconds, some definite trends have been detected in the results. Importantly, most modern empirical force fields are able to correctly predict the stable folded structures and time scales of folding of peptides and small proteins [90, 93, 138]. However, some bias in conformational preferences has been detected, e.g., CHARMM27 and AMBER03 force fields tend to overestimate the stability of α-helical structure while OPLS/AA tends to over-stabilize extended/sheet forms [90, 92, 139]. Further, folding pathways differing in important details have been reported for peptides and small proteins when several force fields were applied to study the same molecular system, e.g., for helix-forming pentapeptides [90] and a small protein [138]. This is quite disappointing, as it is these mechanistic insights, which are potentially the most interesting aspects of modeling, that are the features most difficult to investigate experimentally.

6 Conclusions and Future Directions

Based on the current perspective, several future directions are emerging in peptide modeling, involving simulated systems, algorithms, force fields and computers. In terms of systems studied, the trend for larger and more complex systems and longer simulation times may be expected to continue [140]. This will provide models for complex biological processes, leading to progress in fundamental understanding as well as more practical applications in drug design and new materials. As can be seen from the above review, development of new simulation methods is an area of active research, which contributes significantly to the range of systems and properties that can be modeled. Methods such as replica exchange, free energy simulations, Milestoning and Markov State Modeling have enabled extensive thermodynamic and kinetic description of processes occurring on miliseconds to hours. Similarly, with continuing current trends of use of parallel computing, graphics processors, (GPUs) and other novel architectures [69, 141], progress in computer design is expected to continue contributing to the potential for simulating large systems over long

time scales. Further development and applications of coarse-grained models is another trend that is moving modeling in this direction [25].

With the current stage of methods and computer resources, it appears that the problem of sampling representative conformations for peptides and small proteins is essentially solved. The critical issue that emerges is the accuracy of the potential energy functions. Improved parameterization will be needed for all simulation details to converge. This includes both improvements in force fields and generation of a wider range of experimental data. Two current trends in this area are development of improved electrostatics and use of quantum mechanics (QM) to calculate forces and energies. The new AMOEBA force field is an example of going beyond the static point charge model of electrostatics, including both higher multipoles and polarization [10]. Several polarizable force fields are already in use. These tend to give improved descriptions of energy surfaces at the cost of higher computational effort. The application of QM eliminates the need for fixed parameters altogether, as energies and forces are evaluated from first principles [142, 143]. Methods of this type are applicable to simulating a wide range of processes, including chemical reactions, but are significantly more expensive that molecular mechanics (MM) approaches. Additionally, the levels of approximation required for QM calculations to be feasible for systems of thousands of atoms lead back to the accuracy problem. The way to reconcile these difficulties is through multiscale modeling, where events in small regions over short time scale are described with a higher level, more accurate method, while those over larger distances and longer times are modeled at a lower, less accurate level [142–145].

Acknowledgements

The figures were prepared using MolScript [146, 147], Raster3D [148], ImageMagick [149], and Grace [150].

References

1. Zambrowicz A, Timmer M, Polanowski A, Lubec G, Trziszka T (2013) Manufacturing of peptides exhibiting biological activity. Amino Acids 44:315–320
2. Kastin AJ (2006) Handbook of biologically active peptides. Academic Press, Amsterdam, Boston, pp 1–1636
3. Scior T, Bender A, Tresadern G, Medina-Franco JL, Martinez-Mayorga K, Langer T, Cuanalo-Contreras K, Agrafiotis DK (2012) Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 52:867–881
4. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616

5. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30:1545–1614
6. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24:1999–2012
7. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105:6474–6487
8. Van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) Gromacs: fast, flexible, and free. J Comput Chem 26:1701–1718
9. Poger D, Van Gunsteren WF, Mark AE (2010) A new force field for simulating phosphatidylcholine bilayers. J Comput Chem 31:1117–1125
10. Shi Y, Xia Z, Zhang J, Best R, Wu C, Ponder JW, Ren P (2013) The polarizable atomic multipole-based AMOEBA force field for proteins. J Chem Theory Comput 9:4046–4063
11. Rick SW, Stuart SJ, Berne BJ (1994) Dynamical fluctuating charge force-fields – application to liquid water. J Chem Phys 101:6141–6156
12. Jiang W, Hardy DJ, Phillips JC, MacKerell AD, Schulten K, Roux B (2011) High-performance scalable molecular dynamics simulations of a polarizable force field based on classical drude oscillators in NAMD. J Phys Chem Lett 2:87–92
13. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
14. Berendsen H, Postma J, Van Gunsteren W, Hermans J (1981) Interaction models for water in relation to protein hydration. In: Pullman B (ed) Intermolecular forces. Reidel, Dordrecht, p 331
15. Mark P, Nilsson L (2001) Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. J Phys Chem A 105:9954–9960
16. Rick SW (2004) A reoptimization of the five-site water potential (TIP5P) for use with Ewald sums. J Chem Phys 120:6085–6093
17. 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
18. Onufriev A, Case DA, Bashford D (2002) Effective Born radii in the generalized Born approximation: the importance of being perfect. J Comput Chem 23:1297–1304
19. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A 98:10037–10041
20. Honig B, Nicholls A (1995) Classical electrostatics in biology and chemistry. Science 268:1144–1149
21. Ferrara P, Apostolakis J, Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins 46: 24–33
22. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) Charmm – a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217
23. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25:1656–1676
24. Saunders MG, Voth GA (2013) Coarse-graining methods for computational biology. Annu Rev Biophys 42:73–93
25. Noid WG (2013) Perspective: coarse-grained models for biomolecular systems. J Chem Phys 139(9):090901
26. Chebaro Y, Pasquali S, Derreumaux P (2012) The coarse-grained OPEP force field for non-amyloid and amyloid proteins. J Phys Chem B 116:8741–8752
27. Liwo A, Khalili M, Czaplewski C, Kalinowski S, Oldziej S, Wachucik K, Scheraga HA (2007) Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins. J Phys Chem B 111:260–285
28. Kaufmann KW, Lemmon GH, DeLuca SL, Sheehan JH, Meiler J (2010) Practically useful:

what the ROSETTA protein modeling suite can do for you. Biochemistry 49:2987–2998

29. Maisuradze GG, Senet P, Czaplewski C, Liwo A, Scheraga HA (2010) Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field. J Phys Chem A 114:4471–4485
30. Lewandowska A, Oldziej S, Liwo A, Scheraga HA (2010) beta-hairpin-forming peptides; models of early stages of protein folding. Biophys Chem 151:1–9
31. He Y, Mozolewska MA, Krupa P, Sieradzan AK, Wirecki TK, Liwo A, Kachlishvili K, Rackovsky S, Jagiela D, Slusarz R, Czaplewski CR, Oldziej S, Scheraga HA (2013) Lessons from application of the UNRES force field to predictions of structures of CASP10 targets. Proc Natl Acad Sci U S A 110:14936–14941
32. Swope WC, Andersen HC, Berens PH, Wilson KR (1982) A computer-simulation method for the calculation of equilibrium-constants for the formation of physical clusters of molecules – application to small water clusters. J Chem Phys 76:637–649
33. Verlet L (1967) Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. Phys Rev 159:98–103
34. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical-integration of cartesian equations of motion of a system with constraints – molecular-dynamics of N-alkanes. J Comput Phys 23:327–341
35. Barth E, Kuczera K, Leimkuhler B, Skeel RD (1995) Algorithms for constrained molecular-dynamics. J Comput Chem 16:1192–1209
36. Andersen HC (1983) Rattle – a velocity version of the shake algorithm for molecular-dynamics calculations. J Comput Phys 52: 24–34
37. Miyamoto S, Kollman PA (1992) Settle – an analytical version of the Shake and Rattle algorithm for rigid water models. J Comput Chem 13:952–962
38. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18:1463–1472
39. Tuckerman M, Berne BJ, Martyna GJ (1992) Reversible multiple time scale molecular-dynamics. J Chem Phys 97:1990–2001
40. Martyna GJ, Tuckerman ME, Tobias DJ, Klein ML (1996) Explicit reversible integrators for extended systems dynamics. Mol Phys 87:1117–1157
41. Mobley DL, Klimovich PV (2012) Perspective: alchemical free energy calculations for drug discovery. J Chem Phys 137:230901
42. Mezei M, Beveridge DL (1986) Free energy simulations. Ann N Y Acad Sci 482:1–23
43. Seeliger D, de Groot BL (2010) Protein thermostability calculations using alchemical free energy simulations. Biophys J 98:2309–2316
44. Mobley DL, Liu S, Cerutti DS, Swope WC, Rice JE (2012) Alchemical prediction of hydration free energies for SAMPL. J Comput Aid Mol Des 26:551–562
45. Boresch S, Karplus M (1995) The meaning of component analysis: decomposition of the free energy in terms of specific interactions. J Mol Biol 254:801–807
46. Gao J, Kuczera K, Tidor B, Karplus M (1989) Hidden thermodynamics of mutant proteins: a molecular dynamics analysis. Science 244:1069–1072
47. Mugnai ML, Elber R (2012) Thermodynamic cycle without turning off self-interactions: formal discussion and a numerical example. J Chem Theory Comput 8:3022–3033
48. Torrie GM, Valleau JP (1974) Monte-Carlo free-energy estimates using non-Boltzmann sampling – application to subcritical lennard-jones fluid. Chem Phys Lett 28:578–581
49. Roux B (1995) The calculation of the potential of mean force using computer-simulations. Comput Phys Commun 91:275–282
50. 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
51. Karplus M, Elber R, Gao J, Kuczera K, Tidor B (1989) Dynamics and thermodynamics of myoglobin and hemoglobin. Cytochrome P-450: Biochemistry and Biophysics. pp 258–265
52. Kuczera K (1996) Free energy simulations of axial contacts in sickle-cell hemoglobin. Biopolymers 39:221–242
53. Laio A, Gervasio FL (2008) Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Rep Prog Phys 71:126601
54. Rodriguez-Gomez D, Darve E, Pohorille A (2004) Assessing the efficiency of free energy calculation methods. J Chem Phys 120: 3563–3578
55. Merz KM (2010) Limits of free energy computation for protein–ligand interactions. J Chem Theory Comput 6:1769–1776
56. Jiang W, Hodoscek M, Roux B (2009) Computation of absolute hydration and binding free energy with free energy perturbation distributed replica-exchange molecular dynamics. J Chem Theory Comput 5:2583–2588

57. Boresch S, Tettinger F, Leitgeb M, Karplus M (2003) Absolute binding free energies: a quantitative approach for their calculation. J Phys Chem B 107:9535–9551
58. Mobley DL, Chodera JD, Dill KA (2006) On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. J Chem Phys 125
59. Kollman PA, Massova I, Reyes C, Kuhn B, Huo SH, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Accounts Chem Res 33:889–897
60. Gohlke H, Case DA (2004) Converging free energy estimates: MM-PB(GB)SA studies on the protein–protein complex Ras-Raf. J Comput Chem 25:238–250
61. Carlsson J, Ander M, Nervall M, Aqvist J (2006) Continuum solvation models in the linear interaction energy method. J Phys Chem B 110:12034–12041
62. Swendsen RH, Wang JS (1986) Replica Monte-Carlo simulation of spin-glasses. Phys Rev Lett 57:2607–2609
63. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151
64. Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu KY, Garcia AE (2003) Peptide folding simulations. Curr Opin Struct Biol 13:168–174
65. Hritz J, Oostenbrink C (2008) Hamiltonian replica exchange molecular dynamics using soft-core interactions. J Chem Phys 128
66. Kouza M, Hansmann UHE (2011) Velocity scaling for optimizing replica exchange molecular dynamics. J Chem Phys 134
67. Voter AF (1997) Hyperdynamics: accelerated molecular dynamics of infrequent events. Phys Rev Lett 78:3908–3911
68. Sinko W, de Oliveira CAF, Pierce LCT, McCammon JA (2012) Protecting high energy barriers: a new equation to regulate boost energy in accelerated molecular dynamics simulations. J Chem Theory Comput 8:17–23
69. Pierce LCT, Salomon-Ferrer R, de Oliveira CAF, McCammon JA, Walker RC (2012) Routine access to millisecond time scale events with accelerated molecular dynamics. J Chem Theory Comput 8:2997–3002
70. Wang Y, Markwick PRL, de Oliveira CAF, McCammon JA (2011) Enhanced lipid diffusion and mixing in accelerated molecular dynamics. J Chem Theory Comput 7:3199–3207
71. Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ, Chao JC, Eastwood MP, Gagliardo J, Grossman JP, Ho CR, Ierardi DJ, Kolossvary I, Klepeis JL, Layman T, Mcleavey C, Moraes MA, Mueller R, Priest EC, Shan YB, Spengler J, Theobald M, Towles B, Wang SC (2008) Anton, a special-purpose machine for molecular dynamics simulation. Commun ACM 51:91–97
72. Okamoto Y (2011) Drug design by generalized-ensemble simulations. Curr Pharm Design 17:1758–1772
73. Wu XW, Brooks BR (2011) Toward canonical ensemble distribution from self-guided Langevin dynamics simulation. J Chem Phys 134
74. Voter AF, Germann TC (1998) Accelerating the dynamics of infrequent events: combining hyperdynamics and parallel replica dynamics to treat epitaxial layer growth. Mater Res Soc Symp Proc 528:221–236
75. Bolhuis PG, Chandler D, Dellago C, Geissler PL (2002) Transition path sampling: throwing ropes over rough mountain passes, in the dark. Annu Rev Phys Chem 53:291–318
76. van Erp TS, Bolhuis PG (2005) Elaborating transition interface sampling methods. J Comput Phys 205:157–181
77. Allen RJ, Frenkel D, ten Wolde PR (2006) Forward flux sampling-type schemes for simulating rare events: efficiency analysis. J Chem Phys 124
78. Pande VS, Beauchamp K, Bowman GR (2010) Everything you wanted to know about Markov State Models but were afraid to ask. Methods 52:99–105
79. West AMA, Elber R, Shalloway D (2007) Extending molecular dynamics time scales with milestoning: example of complex kinetics in a solvated peptide. J Chem Phys 126
80. Majek P, Elber R (2010) Milestoning without a reaction coordinate. J Chem Theory Comput 6:1805–1817
81. Cardenas AE, Jas GS, DeLeon KY, Hegefeld WA, Kuczera K, Elber R (2012) Unassisted transport of N-acetyl-l-tryptophanamide through membrane: experiment and simulation of kinetics. J Phys Chem B 116: 2739–2750
82. Prinz JH, Chodera JD, Pande VS, Swope WC, Smith JC, Noe F (2011) Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models

of biomolecular dynamics. J Chem Phys 134(24):244108
83. Senne M, Trendelkamp-Schroer B, Mey ASJS, Schutte C, Noe F (2012) EMMA: a software package for Markov model building and analysis. J Chem Theory Comput 8:2223–2238
84. Mackerell AD, Feig M, Brooks CL (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 25:1400–1415
85. Brooks CL, Case DA (1993) Simulations of peptide conformational dynamics and thermodynamics. Chem Rev 93:2487–2502
86. Wang Y, Kuczera K (1996) Molecular dynamics simulations of cyclic and linear DPDPE: influence of the disulfide bond on peptide flexibility. J Phys Chem 100:2555–2563
87. Wang Y, Goh SY, Kuczera K (1999) Molecular dynamics study of disulfide bond influence on properties of an RGD peptide. J Pept Res 53(2):188–200
88. Mahadevan J, Xu C, Siahaan T, Kuczera K (2002) Molecular dynamics simulations of conformational behavior of linear RGD peptidomimetics and cyclic prodrugs in aqueous and octane solutions. J Biomol Struct Dyn 19:775–788
89. Buchete NV, Hummer G (2008) Coarse master equations for peptide folding dynamics. J Phys Chem B 112:6057–6069
90. Hegefeld WA, Chen SE, DeLeon KY, Kuczera K, Jas GS (2010) Helix formation in a pentapeptide experiment and force-field dependent dynamics. J Phys Chem A 114: 12391–12402
91. Deleon KY, Patel AP, Kuczera K, Johnson CK, Jas GS (2012) Structure and reorientational dynamics of angiotensin I and II: a microscopic physical insight. J Biomol Struct Dyn 29:671–690
92. Hegefeld WA, Kuczera K, Jas GS (2011) Structural dynamics of neuropeptide hPYY. Biopolymers 95:487–502
93. Jas GS, Kuczera K (2012) Computer simulations of helix folding in homo- and heteropeptides. Mol Simulat 38:682–694
94. Tiradorives J, Jorgensen WL (1991) Molecular-dynamics simulations of the unfolding of an alpha-helical analog of ribonuclease-A S-peptide in water. Biochemistry 30:3864–3871
95. Soman KV, Karimi A, Case DA (1991) Unfolding of an alpha-helix in water. Biopolymers 31:1351–1361
96. Brooks CL (1993) Molecular simulations of peptide and protein unfolding – in quest of a molten globule. Curr Opin Struct Biol 3:92–98
97. Young WS, Brooks CL (1996) A microscopic view of helix propagation: N and C-terminal helix growth in alanine helices. J Mol Biol 259:560–572
98. Ferrara P, Apostolakis J, Caflisch A (2000) Computer simulations of protein folding by targeted molecular dynamics. Proteins 39: 252–260
99. Chowdhury S, Zhang W, Wu C, Xiong GM, Duan Y (2003) Breaking non-native hydrophobic clusters is the rate-limiting step in the folding of an alanine-based peptide. Biopolymers 68:63–75
100. Zhang W, Lei HX, Chowdhury S, Duan Y (2004) Fs-21 peptides can form both single helix and helix-turn-helix. J Phys Chem B 108:7479–7489
101. Garcia AE, Sanbonmatsu KY (2002) Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. Proc Natl Acad Sci U S A 99:2782–2787
102. Lin MM, Mohammed OF, Jas GS, Zewail AH (2011) Speed limit of protein folding evidenced in secondary structure dynamics. Proc Natl Acad Sci U S A 108:16622–16627
103. Dinner AR, Lazaridis T, Karplus M (1999) Understanding beta-hairpin formation. Proc Natl Acad Sci U S A 96:9068–9073
104. Klimov DK, Thirumalai D (2000) Mechanisms and kinetics of beta-hairpin formation. Proc Natl Acad Sci U S A 97:2544–2549
105. Garcia AE, Sanbonmatsu KY (2001) Exploring the energy landscape of a beta hairpin in explicit solvent. Proteins 42:345–354
106. Zhou RH, Berne BJ, Germain R (2001) The free energy landscape for beta hairpin folding in explicit water. Proc Natl Acad Sci U S A 98:14931–14936
107. Bolhuis PG (2003) Transition-path sampling of beta-hairpin folding. Proc Natl Acad Sci U S A 100:12129–12134
108. Czaplewski C, Kalinowski S, Liwo A, Scheraga HA (2009) Application of multiplexed replica exchange molecular dynamics to the UNRES force field: tests with alpha and alpha plus beta proteins. J Chem Theory Comput 5:627–640
109. Piana S, Lindorff-Larsen K, Dirks RM, Salmon JK, Dror RO, Shaw DE (2012) Evaluating the effects of cutoffs and treatment of long-range electrostatics in protein folding simulations. PloS One 7:e39918
110. Thompson PA, Munoz V, Jas GS, Henry ER, Eaton WA, Hofrichter J (2000) The helix-coil

kinetics of a heteropeptide. J Phys Chem B 104:378–389

111. Munoz V, Thompson PA, Hofrichter J, Eaton WA (1997) Folding dynamics and mechanism of beta-hairpin formation. Nature 390: 196–199
112. Ferrara P, Apostolakis J, Caflisch A (2000) Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. J Phys Chem B 104: 5000–5010
113. Ferrara P, Caflisch A (2000) Folding simulations of a three-stranded antiparallel beta-sheet peptide. Proc Natl Acad Sci U S A 97:10780–10785
114. Ihalainen JA, Paoli B, Muff S, Backus EHG, Bredenbeck J, Woolley GA, Caflisch A, Hamm P (2008) alpha-Helix folding in the presence of structural constraints. Proc Natl Acad Sci U S A 105:9588–9593
115. Jas GS, Hegefeld WA, Majek P, Kuczera K, Elber R (2012) Experiments and comprehensive simulations of the formation of a helical turn. J Phys Chem B 116:6598–6610
116. Kuczera K, Jas GS, Elber R (2009) Kinetics of helix unfolding: molecular dynamics simulations with milestoning. J Phys Chem A 113:7461–7473
117. Im W, Brooks CL (2005) Interfacial folding and membrane insertion of designed peptides studied by molecular dynamics simulations. Proc Natl Acad Sci U S A 102:6771–6776
118. Nymeyer H, Woolf TB, Garcia AE (2005) Folding is not required for bilayer insertion: replica exchange simulations of an alpha-helical peptide with an explicit lipid bilayer. Proteins 59:783–790
119. Houndonougbo Y, Kuczera K, Jas GS (2005) Structure and dynamics of phospholamban in solution and in membrane bilayer: computer simulations. Biochemistry 44:1780–1792
120. Houndonougbo Y, Kuczera K, Jas GS (2008) Effects of CMAP and electrostatic cutoffs on the dynamics of an integral membrane protein: the phospholamban study. J Biomol Struct Dyn 26:17–34
121. Herce HD, Garcia AE (2007) Molecular dynamics simulations suggest a mechanism for translocation of the HIV-1 TAT peptide across lipid membranes. Proc Natl Acad Sci U S A 104:20805–20810
122. Herce HD, Garcia AE, Litt J, Kane RS, Martin P, Enrique N, Rebolledo A, Milesi V (2009) Arginine-rich peptides destabilize the plasma membrane, consistent with a pore formation translocation mechanism of cell-penetrating peptides. Biophys J 97:1917–1925
123. Bond PJ, Khalid S (2010) Antimicrobial and cell-penetrating peptides: structure, assembly and mechanisms of membrane lysis via atomistic and coarse-grained molecular dynamic simulations. Protein Pept Lett 17:1313–1327
124. Yang C, Jas GS, Kuczera K (2004) Structure, dynamics and interaction with kinase targets: computer simulations of calmodulin. Biochim Biophys Acta 1697:289–300
125. Kuczera K, Kursula P (2012) Interactions of calmodulin with death-associated protein kinase peptides: experimental and modeling studies. J Biomol Struct Dyn 30:45–61
126. Gan WX, Roux B (2009) Binding specificity of SH2 domains: insight from free energy simulations. Proteins 74:996–1007
127. Lin YL, Meng YL, Jiang W, Roux B (2013) Explaining why Gleevec is a specific and potent inhibitor of Abl kinase. Proc Natl Acad Sci U S A 110:1664–1669
128. Lemkul JA, Bevan DR (2012) The role of molecular simulations in the development of inhibitors of amyloid beta-peptide aggregation for the treatment of Alzheimer's disease. ACS Chem Neurosci 3:845–856
129. Shea JE, Urbanc B (2012) Insights into A beta aggregation: a molecular dynamics perspective. Curr Top Med Chem 12:2596–2610
130. Nasica-Labouze J, Mousseau N (2012) Kinetics of amyloid aggregation: a study of the GNNQQNY prion sequence. Plos Comput Biol 8(11):e1002782
131. Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE (1999) Peptide folding: when simulation meets experiment. Angew Chem Int Edit 38:236–240
132. Daura X, Gademann K, Schafer H, Jaun B, Seebach D, van Gunsteren WF (2001) The beta-peptide hairpin in solution: conformational study of a beta-hexapeptide in methanol by NMR spectroscopy and MD simulation. J Am Chem Soc 123:2393–2404
133. Gattin Z, Zaugg J, van Gunsteren WF (2010) Structure determination of a flexible cyclic peptide based on NMR and MD simulation 3J-coupling. Chemphyschem 11:830–835
134. Kuczera K, Unruh J, Johnson CK, Jas GS (2010) Reorientations of aromatic amino acids and their side chain models: anisotropy measurements and molecular dynamics simulations. J Phys Chem A 114:133–142
135. Jas GS, Wang Y, Pauls SW, Johnson CK, Kuczera K (1997) Influence of temperature

and viscosity on anthracene rotational diffusion in organic solvents: molecular dynamics simulations and fluorescence anisotropy study. J Chem Phys 107:8800–8812

136. Jas GS, Larson EJ, Johnson CK, Kuczera K (2000) Microscopic details of rotational diffusion of perylene in organic solvents: molecular dynamics simulation and experiment vs Debye-Stokes-Einstein theory. J Phys Chem A 104:9841–9852
137. Unruh JR, Kuczera K, Johnson CK (2009) Conformational heterogeneity of a leucine enkephalin analogue in aqueous solution and sodium dodecyl sulfate micelles: comparison of time-resolved FRET and molecular dynamics simulations. J Phys Chem B 113: 14381–14392
138. Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? Biophys J 100:L47–L49, Erratum in 2011 Aug 17;101(4):1015
139. Matthes D, de Groot BL (2009) Secondary structure propensities in peptide folding simulations: a systematic comparison of molecular mechanics interaction schemes. Biophys J 97:599–608
140. Piana S, Lindorff-Larsen K, Shaw DE (2012) Protein folding kinetics and thermodynamics from atomistic simulation. Proc Natl Acad Sci U S A 109:17845–17850
141. Ruymgaart AP, Cardenas AE, Elber R (2011) MOIL-opt: energy-conserving molecular dynamics on a GPU/CPU system. J Chem Theory Comput 7:3072–3082
142. van der Kamp MW, Mulholland AJ (2013) Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. Biochemistry 52: 2708–2728
143. Dracinsky M, Moller HM, Exner TE (2013) Conformational sampling by ab initio molecular dynamics simulations improves NMR chemical shift predictions. J Chem Theory Comput 9:3806–3815
144. Vreven T, Morokuma K, Farkas O, Schlegel HB, Frisch MJ (2003) Geometry optimization with QM/MM, ONIOM, and other combined methods. I. Microiterations and constraints. J Comput Chem 24:760–769
145. de Pablo JJ (2011) Coarse-grained simulations of macromolecules: from DNA to nanocomposites. Annu Rev Phys Chem 62:555–574
146. Kraulis PJ (1991) Molscript – a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 24: 946–950
147. Esnouf RM (1999) Further additions to MolScript version 1.4, including reading and contouring of electron-density maps. Acta Crystallogr D 55:938–940
148. Merritt EA, Bacon DJ (1997) Raster3D: photorealistic molecular graphics. Method Enzymol 277:505–524
149. http://www.imagemagick.org/
150. http://plasma-gate.weizman.ac.il/Grace/
151. Blanco FJ, Jimenez MA, Pineda A, Rico M, Santoro J, Nieto JL (1994) NMR solution structure of the isolated N-terminal fragment of protein-G B1 domain. Evidence of trifluoroethanol induced native-like beta-hairpin formation. Biochemistry 33:6004–6014
152. Lamberth S, Schmid H, Muenchbach M, Vorherr T, Krebs J, Carafoli E, Griesinger C (2000) NMR solution structure of phospholamban. Helv Chim Acta 83:2141–2152
153. Nygaard R, Nielbo S, Schwartz TW, Poulsen FM (2006) The PP-fold solution structure of human polypeptide YY and human PYY3-36 as determined by NMR. Biochemistry 45:8350–8357
154. Meador WE, Means AR, Quiocho FA (1992) Target enzyme recognition by calmodulin – 2.4-angstrom structure of a calmodulin-peptide complex. Science 257:1251–1255

Chapter 3

Improved Methods for Classification, Prediction, and Design of Antimicrobial Peptides

Guangshun Wang

Abstract

Peptides with diverse amino acid sequences, structures, and functions are essential players in biological systems. The construction of well-annotated databases not only facilitates effective information management, search, and mining but also lays the foundation for developing and testing new peptide algorithms and machines. The antimicrobial peptide database (APD) is an original construction in terms of both database design and peptide entries. The host defense antimicrobial peptides (AMPs) registered in the APD cover the five kingdoms (bacteria, protists, fungi, plants, and animals) or three domains of life (bacteria, archaea, and eukaryota). This comprehensive database (http://aps.unmc.edu/AP) provides useful information on peptide discovery timeline, nomenclature, classification, glossary, calculation tools, and statistics. The APD enables effective search, prediction, and design of peptides with antibacterial, antiviral, antifungal, antiparasitic, insecticidal, spermicidal, anticancer activities, chemotactic, immune modulation, or antioxidative properties. A universal classification scheme is proposed herein to unify innate immunity peptides from a variety of biological sources. As an improvement, the upgraded APD makes predictions based on the database-defined parameter space and provides a list of the sequences most similar to natural AMPs. In addition, the powerful pipeline design of the database search engine laid a solid basis for designing novel antimicrobials to combat resistant superbugs, viruses, fungi, or parasites. This comprehensive AMP database is a useful tool for both research and education.

Key words Ab initio design, Database filtering tech, Database screen, Peptide design, Peptide prediction, Universal peptide classification

1 Introduction

There are at least two good reasons for our current focus on host defense antimicrobial peptides (AMPs). First, AMPs have remained potent for millions of years. Therefore, AMPs constitute useful templates for developing a new generation of antimicrobials to meet the growing antibiotic resistance problem worldwide. Second, AMPs are key components of the innate immune system universally required for the survival of both invertebrates and vertebrates. Thus, research in this direction improves our understanding of

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_3,

innate immunity and its relationships with the adaptive immune system in vertebrates [1–6].

Lysozyme, discovered by Alexander Fleming in 1922, is now recognized as the first antimicrobial peptide. However, there was little research on AMPs until the discoveries of cecropins, defensins, and magainins in the 1980s [7–9]. Since then, AMPs have been identified from a variety of living species. Select AMPs identified during 1922–2012 are listed in the discovery timeline page of the antimicrobial peptide database (APD) [10, 11]. In earlier days when the number of AMPs was limited, these peptides were handled in review articles. With a rapid increase in the number of such peptides, it became impractical to continue to manage them manually. As a consequence, several databases have been established to categorize these peptides [10–31]. AMSDb appears to be the first such database available online in 1998 [12]. The information format of this database is identical to the SWISS-Prot (UniProt) [32]. It contains 895 antimicrobial peptides, proteins, and their precursors from plants and animals. Unfortunately, AMSDb is no longer updated. To meet the need of better databases with a broad scope, two general databases were published side by side in 2004. ANTIMIC reported more than 1,700 entries [13], while a new version of ANTIMIC called DAMPD [14] contains 1,232 entries. In 2004, the first version of the APD [10] reported 525 peptide entries. These peptides were manually collected from the literature with the aid of public search engines such as PubMed, Swiss-Prot, and PDB [32–34]. The peptide number reached 1,228 entries in the second version of the APD [11] and there are 2,329 peptide entries in the current version.

Since the publication of APD and ANTIMIC, several specialized databases have been established to emphasize certain aspects of natural, synthetic, or recombinant AMPs from a special peptide family (circular peptides, defensins, and thiopeptides) or source (e.g., bacteria, plants, shrimps, amphibians) [15–28]. For example, defensin knowledgebase is dedicated to defensins only, while DADP contains only polypeptides from frogs. More recently, the CAMP [29], YADAMP [30], and LAMP [31] have also been built. Table 1 lists major databases dedicated to AMPs. Among these databases, the APD [10, 11] stands out. This article highlights the unique aspects of the APD as well as new developments since the publication of the second version in 2009.

2 Database Design and Search Functions

2.1 Criteria for Peptide Collections

In terms of peptide registration, the APD database [11] follows a set of self-defined criteria. First, the peptide must have a known amino acid sequence, at least partially. Second, the peptide should have demonstrated antimicrobial activity. Third, the peptide

Table 1

A chronological list of the databases for antimicrobial peptides[a]

Year	Database	URL (http://)	Scope	Country	Reference
1998	AMSDb	www.bbcm.univ.trieste.it/~tossi/amsdb.html	Plant/animal AMPs	Italy	[12]
2002	SAPD	oma.terkko.helsinki.fi:8080/~SAPD/	Synthetic AMPs	Finland	[25]
2003, 2004	Peptaibols	www.cryst.bbk.ac.uk/peptaibol/home.shtml	Fungal peptaibols	England	[26]
2004, 2009	APD	aps.unmc.edu/AP/	AMPs	USA	[10, 11]
2004, 2012	DAMPD	apps.sanbi.ac.za/dampd/	AMPs	South Africa/Saudi Arabia	[13, 14]
2006	PenBase	penbase.immunaqua.com	Shrimp AMPs	France	[15]
2006, 2008	Cybase	research1t.imb.uq.edu.au/cybase/	Circular proteins	Australia	[18]
2006, 2010	BAGLE	bioinformatics.biol.rug.nl/websoftware/bagel/bagel_start.php	Bacterial AMPs	Netherland	[21]
2007	AMPer	marray.cmdr.ubc.ca/cgi-bin/amp.pl	Like AMSDb	Canada	[24]
2007, 2010	BACTIBASE	bactibase.pfba-lab-tun.org/main.php	Bacteriocins	Canada/Tunisia	[17]
2007	Defensins	defensins.bii.a-star.edu.sg/	Defensins	Singapore	[16]
2008	RAPD	faculty.ist.unomaha.edu/chen/rapd/index.php	Recombinant AMPs	USA	[20]
2009	PhytAMP	phytamp.pfba-lab-tun.org/main.php	Plant AMPs	Tunisia/Canada	[19]
2010	CAMP	www.bicnirrh.res.in/antimicrobial	AMPs	India	[29]
2012	YADAMD	yadamp.unisa.it/	AMPs	Italy	[30]
2012	DADP	split4.pmfst.hr/dadp/	Amphibian AMPs	Croatia	[22]
2012	THIOBASE	db-mml.sjtu.edu.cn/THIOBASE/	Bacteria thiopeptides	China	[23]
2012	EnzyBase	biotechlab.fudan.edu.cn/database/EnzyBase/home.php	Cleaving enzymes	China	[27]
2013	LAMP	biotechlab.fudan.edu.cn/database/lamp/guide.php	AMPs	China	[31]
2013	MilkAMP	/milkampdb.org	Milk AMPs	Canada	[28]

[a]Adapted from the APD website (http://aps.unmc.edu/AP/links.php) [10, 11]

contains less than 100 amino acids (this has recently been expanded to 200 amino acids so that some important antimicrobial proteins could be collected). Fourth, the peptide originates primarily from natural sources, including bacteria, protozoa, fungi, plants, and animals. Only a small set of synthetic peptides of general interest was collected. Also, the APD emphasizes unique sequences. Therefore, peptides from different species currently occupy the same entry in this database if they share the same amino acid sequence. At present, there are 46 such entries in the APD, which were "found in multiple species" (the quoted phrase can be searched in the additional information field). Since in silico-predicted peptides may not be truly antimicrobial peptides, they are not registered into the APD at this stage. By following the above criteria, the APD database provides a well-defined set of peptides to the research community. Indeed, the APD is a well-recognized resource in the field of AMPs. For example, the web hits were ~15,000 per year prior to 2008 [11]. Since the publication of the second version in 2009, there is a dramatic increase in database use. For example, the web hits reached 86,000 in 2012 alone.

2.2 A Flexible Database Design

The design of any database is to facilitate information search. Users can conduct a simple search by using peptide name and amino acid sequence in single-letter code. Different from other databases in Table 1, the power of the APD search engine can be ascribed to two important features. First, the search engine is composed of a pipeline of search functions. Second, the modular design of the APD enables continued expansion and development. These features greatly facilitate information search at an advanced level. For example, we obtained 268 defensins using the word "defensin" as a search term. The number of defensins rapidly reduced to 19 when the word "monkey" is also used. Only seven peptides were found when a combination of "defensin," "monkey," and "theta" are used.

2.3 Database Search Functions

To make it easier for the APD users, Table 2 lists major search functions, peptide information, and examples. Most of these search functions are self-explanatory. The name field of the APD, however, has been substantially expanded and deserves some description. It consists of the following elements:

Peptide name + family name + peptide source kingdom + post-translational modification + peptide binding molecules.

In the beginning, it gives peptide name, including synonyms and even the outdated names. In the case of human cathelicidin LL-37, the word LL37 is also used in the literature and FALL-39 is an outdated name. To help users to understand the AMP nomenclature, the major methods used to name AMPs are

Table 2
Search functions of the antimicrobial peptide database

Search function	Peptide information	Examples
APD ID	A unique 5-digit number for each database entry	AP00310
Name	Peptide name or synonyms	LL-37 (LL37, FALL-39)
AMP sequence	Amino acid sequence in single-letter code	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES
Name	Life kingdoms	Bacteria, plants, fungi, protists, animals
Name	Life domains	Bacteria, archaea
Name	Classes	Fish, reptiles, amphibians, birds, insects
Name	Peptide family	Defensins, cathelicidins, histatins, cecropins, magainins
Source species	Location where the peptide is found	Neutrophils; *Homo sapiens*
Length	The number of amino acids	37 (for LL-37)
Net charge	at pH 7	+6 (for LL-37)
Hydrophobic%	Sum of L, I, V. M, A, F, W, C divided by peptide length	35 % (for LL-37)
Name	Chemical modification type	*See* Table 5
Structure	1. Known 3D (α, β, αβ, non-αβ) 2. Partial known (bridged, rich) 3. Unknown	Helix for LL-37
Structural method	X-ray; NMR; CD	NMR (for LL-37)
PDB ID	Self explained	2K6O (for LL-37)
Activity	Known antimicrobial activity	Gram+/Gram−; Gram+; Gram−; viruses; HIV-1; fungi
Name	Binding target	*See* Table 6
Additional info	Mechanism of action	Magainin: forming pores
Additional info	Synergy	LL-37 and lysozyme
Additional info	Animal model	Mouse
Author or Pub year	Search author or publication year separately	Any

Table 3
Select antimicrobial peptide families in the APD[a]

Peptide family	Count	Peptide family	Count
Defensins	268	Aureins	12
Cathelicidins	78	Maximins	30
Histatins	12	Brevinins	185
Neuropeptides	20	Temporins	105
Chemokines	26	Ranatuerins	49
Ribonucleases	6	Dermaseptins	55
		Caerins	29
Cyclotides	151	Maculatins	7
		Uperins	12
Lantibiotics	51	Magainins	5
Microcins	13	Cecropins	24

[a]Peptide counts in this and subsequent tables were obtained from the APD on November 30, 2013

summarized in the APD website (aps.unmc.edu/AP/naming.php). These include the peptide property-based method, the source-based method, and a third method that uses both peptide features and source information. For examples, please visit the APD website.

After the peptide name, the peptide family name is also given in the NAME field. Selected AMP families are tabulated in Table 3. Using the peptide family name, one can obtain a list of AMPs from the same family. For example, there are 268 defensins from a variety of sources and 185 brevinins from amphibians.

Following the family name, the peptide is further annotated in the NAME field based on the source domains or kingdoms. The five kingdoms of life are bacteria, protists (protozoa + algae), fungi, plants, and animals [35], while the three domains of life are bacteria, archaea, and eukaryotes [36]. The peptide counts in each kingdom are listed in Table 4. Selected classes in each life domain are also given in the NAME field, allowing users to focus only on the AMPs of their interest.

The importance of post-translational modifications (PTMs) is only secondary to the peptide sequence itself [37]. Because PTMs could influence both structure and function of the peptide, it is necessary to annotate sequence modification information in the same location. Table 5 contains 23 types of PTMs in the APD. To our knowledge, the APD is the only AMP database that contains extensive information on peptide chemical modifications. In addition, the effect of chemical modification on a peptide net charge is considered in the APD.

Table 4
Antimicrobial peptides from the three domains and five kingdoms of life[a]

Domain	*Peptide count*	*Class*	*Peptide count*
Bacteria	209	Insects	216
Archaea	2	Spiders	33
Eukaryota	2,082	Molluscs	27
		Worms	14
Kingdom	*Peptide count*	Crustaceans	32
Bacteria	209	Birds	36
Protists	7	Reptiles	10
Fungi	12	Fish	79
Plants	301	Amphibians	929
Animals	1,761	Ruminants	44
		Humans	102

How AMPs kill pathogens is an important question to ask. The information for binding targets of AMPs is also annotated in the APD (Table 6). In addition to membranes, AMPs can bind to DNA, heat shock proteins, carbohydrates, and lipid II [1–6].

3 Classification of AMPs Based on Peptide Activity, 3D Structure, and Chain Bonding Pattern

There are a variety of approaches for classifying AMPs. Some of these methods are summarized on the classification page of the APD website (aps.unmc.edu/AP/class.php). For example, the peptides may be classified based on the biosynthesis machinery. Some peptides are synthesized by a multiple enzyme system, while the majority of AMPs are gene-coded. The expression and degradation of gene-coded AMPs are elegantly regulated because either over or under expression of AMPs could cause problems [1–5]. AMPs can also be classified based on molecular targets (e.g., membrane targeting and cell-penetrating peptides) [6]. In the following, we first describe structure and activity-based classification schemes in the APD and then introduce a universal classification scheme for antimicrobial peptides.

3.1 Antimicrobial Activity

As key effector molecules of innate immunity, AMPs are able to control invading pathogenic microbes, including bacteria, viruses, fungi, and parasites [1–4]. It is natural to classify these host defense peptides based on their functions, including antibacterial, antiviral, antifungal, insecticidal, and spermicidal activities. In addition,

Table 5
Post-translational modifications of natural antimicrobial peptides

Search key	Post-translational modification	Peptide count
XXA	Amidation	448
XXB	Chromophore/ion-binding moieties	4
XXC	Backbone cyclization	176
XXD	d-Amino acids	17
XXE	Acetylation	11
XXF	Carboxylic-acid-containing unit	8
XXG	Glycosylation	12
XXH	Halogenation (Cl, Br)	8
XXJ	Sidechain–backbone cyclization	15
XXK	Hydroxylation	9
XXL	Lipidation	9
XXM	Methylation	3
XXN	Nitrolation	0
XXO	Oxidation	10
XXP	Phosphorylation	3
XXQ	N-terminal cyclic glutamate	15
XXR	Reduction	2
XXS	Sulfation	1
XXT	Thioether bridge	46
XXU	Rana Box via a single S–S bond	269
XXW	Dehydration	21
XXY	Citrullination	1
Structure search[a]	Disulfide bridges	551

[a]This number was obtained by searching for disulfide bond-containing AMPs classified as "Bridge," "β structure," and "αβ structure" families, respectively. The "bridged" AMPs are known to have disulfide bonds but unknown 3D structure. Beta structures without disulfide bonds were excluded by including "c" as a sequence search term. For the αβ structures, only the AMPs with a packed 3D fold were counted

some AMPs also possess other functional roles such as anticancer, wound healing, and immune modulation [4]. The APD database has annotated 17 types of peptide activities or functions (Table 7). Several newly annotated activity types are unique in this database, making the APD most comprehensive in terms of activity annotation.

Table 6

Binding targets of antimicrobial peptides

Search key[a]	Binding target	Count
BBBh2o	Self aggregation in water	15
BBBm	Oligomers in membranes	4
BBII	Ions	16
BBW	Lipid II	17
BBL	LPS	54
BBr	Receptors	3
BBMm	Membranes	81
BBN	Nucleic acids	11
BBS	Sugars/carbohydrates	44

[a]Search by entering the code into the name field of the APD [10, 11]

Table 7

Biological activities of host defense antimicrobial peptides

Year created	Activity[a]	Count
2003	Antibacterial (G+/G–)	1,909
2003	Antifungal	850
2003	Antiviral	138
2003	Anticancer	158
2003	Hemolytic	284
2008	Anti-HIV	92
2009	Anti-G+	360
2009	Anti-G–	172
2009	Antiparasitic	59
2009	Insecticidal	22
2009	Spermicidal	9
2011	Chemotactic	47
2012	Anti-protist	4
2013	Antioxidant	10
2013	Anti-inflammatory	2
2013	Wound healing	7
2013	Enzyme inhibitor	5

[a]Some newly defined search functions can be searched in the "additional information" field of the APD by entering the words in the table. These include antioxidant, anti-inflammatory, and wound healing, and enzyme inhibitor

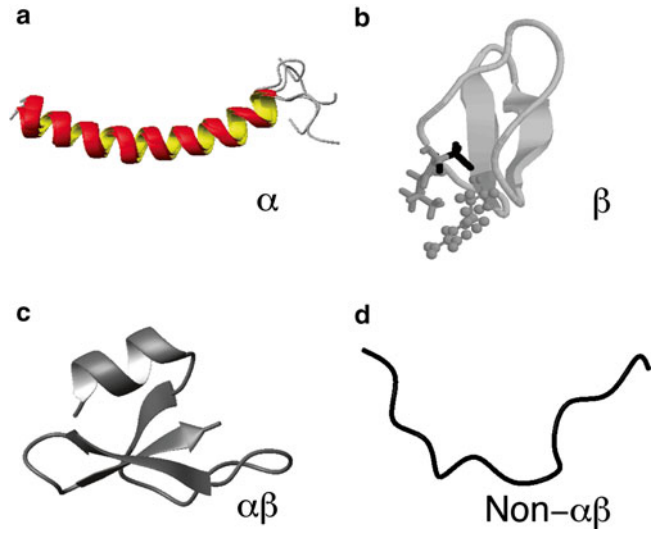


Fig. 1 Classification of the 3D structures of antimicrobial peptides into four families [6]. Shown are representatives from each family: (**a**) α-*helical* structure of human cathelicidin LL-37 (PDB entry: 2K60) [55]; (**b**) the β-*sheet* structure of plant kalata B1 (PDB entry: 1JNU) [56]; (**c**) the αβ *structure* of human β-defensin-1 (HBD-1) (PDB entry: 1IJV) [57]; and (**d**) the *non*-αβ *structure* of cattle indolicidin (PDB entry: 1G89) [58]

3.2 Three-Dimensional Structure of AMPs

According to the APD, only a small population of AMPs (13 %) has a known 3D structure, primarily determined by solution nuclear magnetic resonance (NMR) spectroscopy [10]. In addition, X-ray diffraction was also used to solve the structures of some AMPs with a folded structure in water. The structural information is well annotated in the APD database, including structural class, method for structural determination, structural regions, key residues, and membrane-mimetic models for structural determination. In addition, users can directly view the 3D structure via the link to the PDB [33]. The AMP structures are usually classified into α-helical, β-sheet, and extended structures [4, 38]. A more general classification approach has been proposed recently [6]. In this approach, the AMP structures are classified into four families: α, β, αβ, and non-αβ based on the types of secondary structures. Peptides in the α family contain α-helical structure (Fig. 1a) as the major secondary structure. In contrast, AMPs in the β family are characterized by at least a pair of two β-strands in the structure (Fig. 1b). The αβ family contains both α and β structures (Fig. 1c), whereas the non-αβ family has neither α nor β structure (Fig. 1d). This structural classification scheme is now executed in the APD. Typical examples and peptide counts from different families are provided in Table 8. While the α-helical family is the largest with 328 entries, the non-αβ family is the smallest with merely 9 entries. Table 8 also shows that the lysine/arginine (K/R) ratios in these structural

Table 8
Classification of 3D structures of antimicrobial peptides

Structure[a]	*K*/*R* ratio	Peptide count	Examples
α	13.65/5.26 = 2.59	329	Cecropin, dermcidin, LL-37, magainin
β	5.63/10.7 = 0.53	97	Human alpha defensins (HNP-1, HNP-4, and HD-5), plant kalata B1
αβ	8.47/7.05 = 1.2	81	Drosomycin, Human beta defensins (HBD-1, HBD-4), PhD1
Non-αβ	4.85/10.19 = 0.48	9	Indolicidin, tritrpticin, drosocin, nisin A

[a]For AMPs without 3D structures, additional annotations were made in the APD: (1) unknown, no 3D structure; (2) bridge, disulfide-linked, usually beta-structure; (3) rich, rich in certain amino acids

families differ. While lysines are dominant in the α-helical family, arginines are preferred in the β-family as well as the non-αβ family. Not surprisingly, AMPs with both α and β structures have a moderate *K*/*R* ratio of ~1.2. These ratios might become useful as indicators for classifying a newly discovered peptide into a particular structural family.

3.3 A Universal Classification of AMPs Based on Peptide Bonding Patterns

Because only a small number of AMPs has a 3D structure, we herein propose a systematic classification approach that is independent of 3D structure, peptide source, or activity. This classification is framed based on the connection mode of polypeptide chains. Class I includes linear AMPs (Fig. 2a), which may be chemically modified (amidation, sulfate, phosphate, bromide, or glycosylation) at side chains or even backbones. However, such modifications (Table 5) for class I AMPs do not lead to chain connections between different amino acids. Class II covers all AMPs with chemical bonds between different peptide side chains (Fig. 2b). These include lantibiotics (thioether rings) and the defensin family (disulfide bonds). Broadly, it can be any type of chemical connections between two amino acids. When two or more peptides work together, they belong to this class as long as any of the polypeptide chain contains a side-chain–sidechain connection. Class III AMPs must possess a chemical bond between peptide side chain and backbone (Fig. 2c). The typical members are lassos where the carboxyl group of residue E8 or D9 is covalently linked to the N-terminal amine group. It can be any type of chemical bonding between the side chain of one amino acid and the backbone of another amino acid (*see* Table 9). Lastly, class IV is composed of circular peptides where a peptide bond is formed between the amino and carboxylic ends of the peptide backbone (Fig. 2d). These circular peptides may (or may not) contain additional modifications such as disulfide bonds. Examples are enterocin AS-48 from bacteria, cyclotides from plants, and θ-defensins from primates [37].

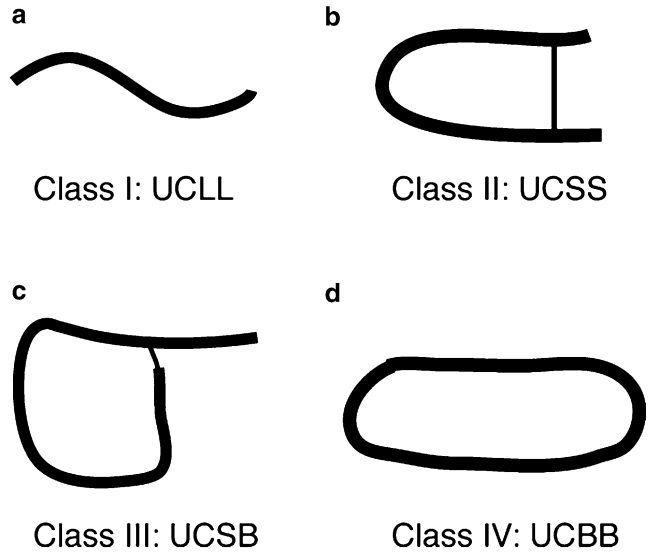


Fig. 2 Classification of antimicrobial peptides based on the connection patterns of the polypeptide chain: (**a**) linear polypeptide chains (e.g., LL-37 and magainins); (**b**) sidechain-linked peptides such as defensins and lantibiotics; (**c**) polypeptide chains with side chain to backbone connection (e.g., lassos); and (**d**) circular peptides with a seamless backbone (e.g., cyclotides)

Each class of AMPs can be further classified. For class I peptides, they can be classified into two subclasses based on the number of polypeptide chains (Table 9). Single-chain linear AMPs are further classified based on chemical modifications. Unmodified AMPs include "amino acid rich" and "not amino acid rich" families. Modified peptides are further divided into two types based on modification sites (side chain or backbone). These systematic classifications for class 1 AMPs are summarized in Table 10 with examples. Likewise, class II AMPs with connections between side chains can be further classified based on the number of polypeptide chains as well as the type of chemical bonds (Table 9). A further classification of the single-chain disulfide-bonded AMPs (e.g., defensins or defensin-like) based on the number of S–S bonds is provided in Table 11. It is also possible to further classify single-chain lantibiotics based on the number of thioether bonds (Table 12). A new type of sidechain–sidechain connection will constitute a new subclass. In the same vein, class III AMPs can be further separated into different types based on the bond type (Table 9). This chemical bond-based classification is also extended to class IV. Circular AMPs are classified based on the additional types and number of chemical bonds in the polypeptide chain (Table 13). This systematic classification system covers all AMPs and should complement with the existing classification systems proposed for AMPs from different life domains [39–43].

Table 9

A universal classification of antimicrobial peptides

Class	Chain linkage	Subclass	Link type	Class symbol	Examples
I	Linear and open chains[a]	1. One chain	None	UCLL1[a]	LL-37, magainins
		2. Two chains	None	UCLL2	Enterocin L50
II	Sidechain–sidechain	1. One chain	Cβ–S–S–Cβ (Disulfide-bond)	UCSS1a[a]	Defensin-like
			Cβ–S–Cβ (thioether)	UCSS1b[a]	lantibiotics
		2. Two chains	Interchain Cβ–S–S–Cβ	UCSS2a	Distinctin, halocidin, centrocin
			Intra-chain Cβ–S–Cβ	UCSS2b	Lacticin-3147, Smb
III	Sidechain–backbone	One chain	CO–NH amide	UCSB1a	Microcin J25, Lariatins
			CO–O ester	UCSB1b	Fusaricidin A
			Cβ–S–Cα	UCSB1c	Thuricin CD
IV	Backbone–backbone	One chain	CO–NH amide	UCBB1a[a]	AS-48, subtilosin A, cyclotides, θ-defenins

[a]Further classifications are provided in Tables 10, 11, 12, and 13

Table 10

Classification of class 1 linear antimicrobial peptides (UCLL1)

Subclass	Modification site[a]	Modification type	Subtype	Peptide examples
UCLL1A	None	None	Not-AA-rich[b]	LL-37
			AA-Rich (25 %)	Pro-rich; Arg-rich PR-39
UCLL1B	Sidechain	Group attachment	Hydroxylation; halogenation; phosphorylation; glycosylation; lipidation; sulfation	Piscidin 4 (hydroxylated Trp); datucin, MccC7
		Sidechain cyclization	Cyclic glutamate	Heliocin
UCLL1C	Backbone	End capping	Amidation; acetylation; other attachments	Aurein 1.2; temproin A
		Configuration change	d-Amino acids	Gramicidin; bombinin H4
		Backbone transformed	Dehydrated	Cypemycin (Linaridins)
			Heterocyclic rings	Thiopeptides in ThioBase

[a]Post-translational modification (PTM) is a broad concept that includes all types of functional groups attached to the peptide chain via covalent bond formation. A detailed list of PTMs is provided in Table 5. Some common examples are N-terminal acetylation, C-terminal amidation, phosphorylation, glycosylation, aromatic halogenation, and sulfation. In the extreme case, even the peptide backbone is modified, leading to dehydrated or heterocycles. However, all these modifications are limited to a single amino acid and do not lead to a polypeptide chain connection between different amino acids as observed in the other three major classes of AMPs (Table 9)

[b]AA = Amino acids

Table 11

Sidechain–sidechain connected antimicrobial peptides: further classification of single-chain peptides containing disulfide bonds (UCSS1a)

Type	S–S bond count	Subtype[a]	Examples
I	1	A B C	Brevinin, esculentin (Rana box) Thanatin Bactenecin
II	2	A B	Ec-AMP1, lasiocepsin, Glycocin F Protegrin, polyphemusin, CXCL1, LEAP-2
III	3	A B	NK-lysin, caenopore-5 HNP-1, HBD-1, big defensins
IV	4	B	ASABF, NaD1, drosomycin
V	5	B	PhD1, WAMP-1a, Ec-CBP
VI	6	B	Cospin

[a]The peptides can further be classified into subtypes based on 3D structure (A: α-helical; B: β-sheet-containing (β and αβ families); C: non-αβ; D: unclassified due to an unknown 3D structure)

Table 12

Sidechain–sidechain connected antimicrobial peptides: further classification of single-chain lantibiotics containing thioether bonds (UCSS1b)

Type	Number of linkage	Examples
I	1	Not found
II	2	Bovicin HJ50
III	3	Epilancin 15X, Lacticin 481
IV	4	Cinnamycin, Actagardine A
V	5	Nisin, Microbisporicin, Subtilin, Ericin A, Paenibacillin
VI	6	Paenicidin A
VII	7	Geobacillin I

Table 13

Classification of circular antimicrobial peptides (UCBB1a)

Type	Additional links	Examples
A	None	Bacterial enterocin AS-48
B	Sidechain–sidechain (Cβ–S–S–Cβ)	Plant cyclotides, primate θ-defensins
C	Sidechain–backbone (Cβ–S–Cα)	Bacterial subtilosin

4 Peptide Prediction

Based on the information content used in the prediction programs, the prediction methods of AMPs have been classified into five types [6]. The first type uses only mature peptide sequences, while the second method involves only the precursor sequences. The third prediction type considers both mature and precursor sequences. The fourth method employs the sequence similarity of the modifying enzymes. Finally, the fifth prediction uses genomic information. It is possible that each prediction above can be achieved in different ways. For example, based on the mature AMP sequences in the APD [10, 11], numerous prediction methods have been developed. In the Lata method [44], two data sets were utilized: antimicrobial and non-antimicrobial. While it is easy to download the positive data set from the APD, it is difficult to get a true negative data set because the activities of the sequences in the negative data set have not been validated by experiments. Yet the program is set up with a good predictive ability. A recent prediction method iAMP-2L [45] considers multiple functions of AMPs annotated in the APD. Different from all other prediction protocols (reviewed in ref. 6), a unique prediction method is programmed in the APD. This method does not require a negative data set, but is coupled with the database. In the following, we describe an upgraded version of this APD method.

The original prediction method in the APD made predictions based on some known rules [10]. Hence, the method was referred to as knowledge-based prediction. For example, AMPs are usually cationic. A peptide with a negative net charge was predicted as "less likely to be an antibacterial peptide." This simple prediction has its limitations because the database does contain anionic AMPs. To overcome this shortcoming, we have updated the prediction interface based on the parameter space defined by the whole peptide set in the APD. The parameters for antimicrobial peptides are better defined due to a fourfold increase in peptide number from the original 525 to the current 2,329. Peptide parameters such as length, net charge, hydrophobic percentage, and amino acid composition can all be calculated. *These parameters constitute the parameter space of natural AMPs.*

In terms of net charge, the known AMPs occupy a very broad range. The AMP with the most negative net charge is chrombacin (net charge −12). Two AMPs, sheep cathelicidin OaBac11 and fish histone-derived Oncorhyncin II, possess the highest net charge of +30. Thus, the boundary conditions for net charge are defined as

$$-12 < \text{net charge} < +30.$$

The above boundary condition can be incorporated into the APD program to make database-based predictions. This expansion

enables the prediction of a broader range of peptide sequences. Because the majority of the AMPs (97.4 %) have a net charge between −5 and +10, it may be useful to define this range as the core region. The small number of AMPs outside the core region may be called the minor region. This core region may be used as an alternative condition for prediction.

The hydrophobic content (i.e., the sum of hydrophobic amino acids divided by the total number of amino acids in a peptide) is another important parameter that determines peptide properties. In the APD, hydrophobic amino acids include alanines (Ala), valines (Val), leucines (Leu), isoleucines (Ile), methionines (Met), phenylalanines (Phe), tryptophans (Trp), and cysteines (Cys) [10]. Based on the database sorting function, we identified the AMPs with the lowest and highest hydrophobic contents. Sheep anionic peptide SAAP (sequence: DDDDDD) contains no hydrophobic residues in the sequence, leading to a hydrophobic content of 0 %, while gramicidins have the highest hydrophobic content of 93 %. Thus, the boundary conditions for peptide hydrophobic contents are defined as

$$0\ \% \leq \text{hydrophobic content} < 93\ \%.$$

The peak of this hydrophobic distribution is located between 40 and 50 % [46]. This leads to another set of boundary conditions for our database-based prediction. We can also define the core region based on the hydrophobic content. The AMPs in the core region (98.6 %) possess a hydrophobic content between 10 and 80 %.

The length of the peptides in the current APD ranges from 5 to 174. The lower limit is real, while the upper limit is arbitrary since it is defined by the scope of peptides collected into the database (<200 amino acids). However, the majority of AMPs (92.9 %) are less than 60 amino acids in length, leading to a definition of the core length region of 5–60. We can anticipate that these boundary conditions will be fully determined when a sufficient number of representative natural AMPs have been identified and registered into the APD.

During this study, we have executed these new database-derived boundary conditions in the prediction interface of the APD (Fig. 3). This interface makes predictions based on sequence similarity. In the first step, the prediction program will calculate the peptide parameters based on the input sequence. The calculated peptide parameters will then be compared with the APD parameter space. If one or more calculated parameters fall outside the database-defined parameter space, the users will be informed that "your input is less likely to be an antibacterial peptide." If all the parameters fall within the defined parameter space, the database will conduct a second tier of prediction by broadly classifying input peptides into several classes: rich in amino acids (>25 % for any

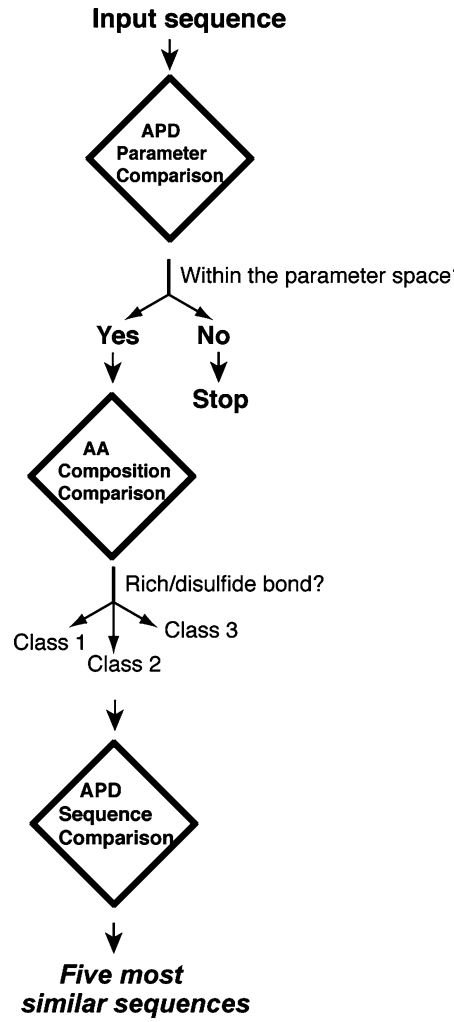


Fig. 3 Prediction of antimicrobial peptides based on the antimicrobial peptide database. The prediction consists of three steps. As the first step, the program will determine whether the input sequence is in the database-defined parameter space (such as charge and hydrophobic content). If identical, the users will be informed. If one or more calculated parameters of the input peptide are out of the boundaries, it is predicted as "your sequence is less likely to be an antibacterial peptide." Second, the input sequence will be classified into three families: rich in amino acids such as histatins and tryptophans, disulfide-linked peptides, and linear. Third, sequence alignments will be conducted to find five peptides that are most similar to the query sequence

amino acid), helical, and disulfide-linked. With the execution of the universal classification proposed in Table 9, a more accurate prediction will be realized. As the third tier of our prediction, the database compares the input sequence with all the peptides in the database by performing sequence alignment. Five peptides with

most similar sequences will be provided in the output. Because we use database-derived parameters for prediction, we refer to this upgraded method as the APD-based prediction (the November 2013 version). Compared to the original prediction [10], the upgraded version is able to handle a broader range of peptide sequences. In addition, the chance of identifying the most similar sequences in the APD also increases substantially as a consequence of a fourfold increase in natural compounds.

The identification of most similar AMPs is a useful feature. For example, O'Shea did not find similar sequences by searching the BLAST database [47], but were able to do so using the APD. Based on the sequence similarity of a novel bacteriocin with plant Ib-AMP3, these authors named the new bacteriocin as bactofensin. The similarity also inspired the authors to test possible antimicrobial activities listed for Ib-AMP3. In addition, the authors can also check whether the new peptide has a similar 3D structure. Thus, the output from the APD prediction programs can guide users to design new experiments to test the structure and activity of the newly identified peptide based on the knowledge annotated for the most similar candidates in the database. Such a prediction of sequence, structure, and activity at multiple levels requires careful annotation of AMP information in the APD.

5 Peptide Design

The APD [10, 11] also provides a useful platform for identification of useful antimicrobials to combat difficult-to-kill pathogens such as human immune-deficiency virus (HIV) and methicillin-resistant *Staphylococcus aureus* (MRSA) [46]. Both database screening and database-guided design have been conducted. By screening a representative set of AMPs selected from the APD, we found several potent anti-HIV or anti-MRSA peptides [48, 49]. New peptides were also obtained by modifying, shuffling, or hybriding natural sequences. Mathematically, a known peptide sequence can be shuffled into multiple sequences. Experimentally, we found that sequence shuffling could lead to all the possibilities: less active, equally active, and more potent sequences [49]. An MIT group developed a large-scale hybrid approach by combining sequence segments of ten residues (i.e., grammars). This grammar approach can generate new sequences, which may, or may not, be bactericidal [50]. A complete different approach in the form of combinatorial libraries can also be pursued [51]. In principle, the amino acid at each position of the peptide sequence can be changed into other amino acids. In practice, it is necessary to bias the choice of amino acids in order to obtain active peptides [52]. This is because the amino acid use in natural AMPs is biased. The APD enabled us to identify the frequently occurring amino

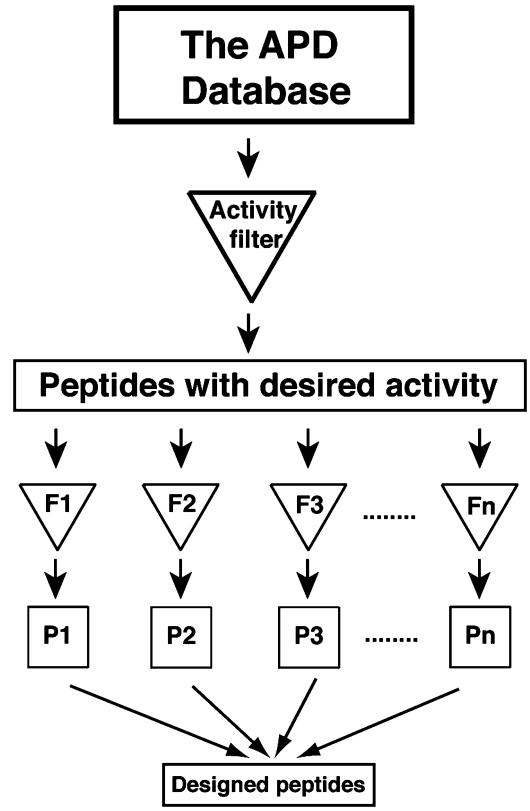


Fig. 4 Ab initio peptide design based on the database filtering technology (DFT). The DFT tech developed recently [54] is composed of two layers of filters. The first layer filter enables the identification of a set of antimicrobial peptides with the desired activity from the antimicrobial peptide database (*see* Table 7). This set of peptides is then used as templates to extract useful parameters for peptide design by utilizing the second layer filters (F1, F2, F3, …, Fn). These peptide parameters (P1, P2, P3, …, Pn) are combined to generate a single or limited number of peptides

acids for AMPs from a variety of life domains [10, 53]. For example, the frequently occurring residues (≥8.5 %) are leucines, glycines, and lysines based on the average percentages of all the 2,329 peptides in the current APD. We demonstrated previously that these three amino acids contained sufficient information for designing antibacterial peptides [11].

Another important approach is de novo design (reviewed in ref. 6). We have recently developed a novel database approach [54]. A flowchart for this approach is provided as Fig. 4. This flowchart contains two major tiers of information filters. The first tier consists of an activity filter that enables one to obtain a set of peptides with desired activity. Table 3 lists 17 types of peptide activities, each of which contains a set of model peptides. In our design,

we selected a group of peptides with activity against gram-positive bacteria. This set of peptides formed the templates for extracting useful parameters for designing anti-MRSA peptides. The second tier contains numerous filters (F1, F2, …, Fn), each defines one parameter for the peptide (P1, P2, …, Pn). In determining these parameters, we followed the most probable principle, which projected the maximum for each parameter. Because the most probable parameters were used, the peptides assembled in this manner had a good chance to be antimicrobial. This is indeed the case. The designed peptide DFTamP1 rapidly killed MRSA USA300, a community-associated staphylococcal pathogen. It also showed some bacterial selectivity since DFTamP1 did not kill gram-negative bacteria *E. coli*, *P. aeruginosa*, or gram-positive *B. subtilis*. This success opens a new avenue to designing peptides with various types of activities (Table 7). Because this new method differs from all existing de novo approaches, it was referred to as ab initio design [54].

6 Concluding Remarks and Future Studies

The antimicrobial peptide database was constructed 10 years ago. It is an original construction in terms of both database design and peptide entries. Each peptide entry in the APD was manually collected from the literature using the public search engines such as PubMed, PDB, and Swiss-Prot. By following a set of rules for data registration, the APD presents a well-defined set of natural AMPs. To achieve a more complete sampling of natural AMPs, the database is extensively annotated and regularly updated. In addition, the pipeline design led to a powerful search engine. This unique database, therefore, constitutes the basis for developing new methods for peptide classification, prediction and design. The APD is the first to adopt both five-kingdom and three-domain classifications, allowing users to search the AMP information from any kingdom (bacteria, protists, fungi, plants, and animals) or classes (e.g., insects, spiders, molluscs, crustaceans, reptiles, amphibians, fish, and birds) (Table 4). Once a domain is defined in the NAME field, the APD behaves like a specialized database (e.g., plant AMPs, bacteriocins, and amphibian peptides). The APD also executed a new structure classification scheme based on the types of secondary structures (α, β, αβ, and non-αβ) in a variety of 3D structures of AMPs (Fig. 1) [6]. Needless to say, the structures in each family can be further grouped based on the number of secondary structures (e.g., α-helix and β-strand). Due to a limited number of known 3D structures, we have proposed a universal classification scheme here based on peptide chain bonding patterns (Fig. 2). Since the information on peptide source, activity, and 3D structure is not required, this systematic classification (Tables 9,

10, 11, 12, and 13) complements to the existing classification methods for AMPs in a defined life kingdom such as bacteria and plants [39–43]. It also offers an approach to unifying the classification of antimicrobial peptides. This classification is general and can be applied to other biologically active peptides.

There are various prediction methods for AMPs (reviewed in ref. 6). The APD is unique in that the prediction is highly coupled with the database. The upgraded version of the APD makes predictions in three steps by following the similarity principle. Each step deals with a specific question. The first tier asks whether the peptide parameters of the input sequence fall within the database parameter space. Based on the amino acid composition analysis, the second tier asks which peptide class the input sequence belongs to. The third tier determines five most similar sequences based on sequence alignment with all the peptides in the database. It is clear why we have been strict in following a set of rules in registering AMPs. Our practice allows us to more accurately map the parameter space for natural AMPs. When a large number of predicted or artificial sequences are included, such parameters could deviate from nature's parameters, thereby influencing the prediction quality. In addition, users can get an idea of the structural type and functional space of the input sequence by viewing the similar sequences already in the APD. For example, the input sequence is most likely to form a helix-bundle structure stabilized by three disulfide bonds if the best match is a saposin-like protein. If the sequence matches human cathelicidin LL-37, it is likely to have multiple functions, ranging from antimicrobial, wound healing, to immune modulation. Like LL-37, the peptide may also have a broad-spectrum activity to kill bacteria, fungi, viruses, and parasites. This information will guide the users to validate both structure and activity of a new peptide.

Finally and importantly, the construction of this well-annotated database also enabled us to develop novel approaches for designing peptides with desired properties. Based on the database, we have tested two general approaches: peptide screening [48, 49] and database-guided design [46, 54]. In particular, we demonstrated the first ab initio design based on the database by developing the database filtering technology [54]. This approach is not limited to the development of anti-MRSA peptides and can be applied to the design of peptides with other types of activities (Table 7) as well. It is also desirable that the designed peptides only kill a specific species. Our detailed annotations of AMP targeting organisms into the database set the stage for this effort. In addition, other database filters such as peptide selectivity and stability to proteases can be created as well. Taken together, the APD is a powerful engine for research and education in the field of innate immunity and drug discovery.

Acknowledgements

This study was supported by grants from the NIH (1R01AI105147-01A1, 1R56AI105147-01) and the State of Nebraska. The author thanks Zhe Wang for programming the original database and Biswajit Mishra for conducting the ab initio design of novel antimicrobials.

References

1. Zasloff M (2002) Antimicrobial peptides of multicellullar organisms. Nature 415:359–365
2. Lehrer RI (2007) Multispecific myeloid defensins. Curr Opin Hematol 14:16–21
3. Boman HG (2003) Antibacterial peptides: basic facts and emerging concepts. J Intern Med 254:197–215
4. Hancock RE, Sahl HG (2006) Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat Biotechnol 24:1551–1557
5. Lai Y, Gallo RL (2009) AMPed up immunity: how antimicrobial peptides have multiple roles in immune defense. Trends Immunol 30: 131–141
6. Wang G (ed) (2010) Antimicrobial peptides: discovery, design and novel therapeutic strategies. CABI, Oxfordshire, UK
7. Steiner H, Hultmark D, Engström Å, Bennich H, Boman HG (1981) Sequence and specificity of two antibacterial proteins involved in insect immunity. Nature 292:246–248
8. Selsted ME, Harwig SS, Ganz T, Schilling JW, Lehrer RI (1985) Primary structures of three human neutrophil defensins. J Clin Invest 76: 1436–1439
9. Zasloff M (1987) Magainins, a class of antimicrobial peptides from Xenopus skin: isolation, characterization of two active forms, and partial cDNA sequence of a precursor. Proc Natl Acad Sci U S A 84:5449–5453
10. Wang Z, Wang G (2004) APD: the antimicrobial peptide database. Nucleic Acids Res 32:D590–D592
11. Wang G, Li X, Wang Z (2009) The updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res 37:D933–D937
12. Tossi A, Sandri L (2002) Molecular diversity in gene-coded, cationic antimicrobial polypeptides. Curr Pharm Des 8:743–761
13. Brahmachary M, Krishnan SP, Koh JL, Khan AM, Seah SH, Tan TW, Brusic V, Bajic VB (2004) ANTIMIC: a database of antimicrobial sequences. Nucleic Acids Res 32:D586–D589
14. Seshadri Sundararajan V, Gabere MN, Pretorius A, Adam S, Christoffels A, Lehväslaiho M, Archer JA, Bajic VB (2012) DAMPD: a manually curated antimicrobial peptide database. Nucleic Acids Res 40:D1108–D1112
15. Gueguen Y, Garnier J, Robert L, Lefranc MP, Mougenot I, de Lorgeril J, Janech M, Gross PS, Warr GW, Cuthbertson B, Barracco MA, Bulet P, Aumelas A, Yang Y, Bo D, Xiang J, Tassanakajon A, Piquemal D, Bachère E (2006) PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. Dev Comp Immunol 30:283–288
16. Seebah S, Suresh A, Zhou S, Choong YH, Chua H, Chuon D, Beuerman R, Verma C (2007) Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. Nucleic Acids Res 35: D265–D268
17. Hammami R, Zouhir A, Ben Hamida J, Fliss I (2007) BACTIBASE: a new web-accessible database for bacteriocin characterization. BMC Microbiol 7:89
18. Wang CK, Kaas Q, Chiche L, Craik DJ (2008) Cybase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. Nucleic Acids Res 36:D206–D210
19. Hammami R, Ben Hamida J, Vergoten G, Fliss I (2009) PhytAMP: a database dedicated to antimicrobial plant peptides. Nucleic Acids Res 37:D963–D968
20. Li Y, Chen Z (2008) RAPD: a database of recombinantly-produced antimicrobial peptides. FEMS Microbiol Lett 289:126–129
21. de Jong A, van Heel AJ, Kok J, Kuipers OP (2010) BAGEL2: mining for bacteriocins in genomic data. Nucleic Acids Res 38: W647–W651

22. Novković M, Simunić J, Bojović V, Tossi A, Juretić D (2012) DADP: the database of anuran defense peptides. Bioinformatics 28: 1406–1407
23. Li J, Qu X, He X, Duan L, Wu G, Bi D, Deng Z, Liu W, Ou HY (2012) ThioFinder: a web-based tool for the identification of thiopeptide gene clusters in DNA sequences. PLoS One 7:e45878
24. Fjell CD, Hancock RE, Cherkasov A (2007) AMPer: a database and an automated discovery tool for antimicrobial peptides. Bioinformatics 23:1148–1155
25. Wade D, Englund J (2002) Synthetic antibiotic peptides database. Protein Pept Lett 9:53–57
26. Whitmore L, Wallace BA (2004) The Peptaibol Database: a database for sequences and structures of naturally occurring peptaibols. Nucleic Acids Res 32:D593–D594
27. Wu H, Lu H, Huang J, Li G, Huang Q (2012) EnzyBase: a novel database for enzybiotic studies. BMC Microbiol 12:54
28. Theolier J, Fliss I, Jean J, Hammami R (2013) MilkAMP: a comprehensive database of antimicrobial peptides of dairy origin. Dairy Sci Technol 94:181–193
29. Thomas S, Karnik S, Barai RS, Jayaraman VK, Idicula-Thomas S (2010) CAMP: a useful resource for research on antimicrobial peptides. Nucleic Acids Res 38:D774–D780
30. Piotto SP, Sessa L, Concilio S, Iannelli P (2012) YADAMP: yet another database of antimicrobial peptides. Int J Antimicrob Agents 39:346–351
31. Zhao X, Wu H, Lu H, Li G, Huang Q (2013) LAMP: a database linking antimicrobial peptides. PLoS One 8:e66557
32. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34: D187–D191
33. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlic A, Quesada M, Quinn GB, Ramos AG, Westbrook JD, Young J, Zardecki C, Berman HM, Bourne PE (2013) The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res 41:D475–D482
34. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA (2002) Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res 30:13–16
35. Whittaker RH (1969) New concepts of kingdoms of organisms. Science 163:150–160
36. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–5090
37. Wang G (2012) Chemical modifications of natural antimicrobial peptides and strategies for peptide engineering. Curr Biotechnol 1:72–79
38. Epand RM, Vogel HJ (1999) Diversity of antimicrobial peptides and their mechanisms of action. Biochim Biophys Acta 1462:11–28
39. Klaenhammer TR (1993) Genetics of bacteriocins produced by lactic acid bacteria. FEMS Microbiol Rev 12:39–85
40. Duquesne S, Destoumieux-Garzón D, Peduzzi J, Rebuffat S (2007) Microcins, gene-encoded antibacterial peptides from enterobacteria. Nat Prod Rep 24:708–734
41. Egorov TA, Odintsova TI, Pukhalsky VA, Grishin EV (2005) Diversity of wheat antimicrobial peptides. Peptides 26:2064–2073
42. Bulet P, Stocklin R (2005) Insect antimicrobial peptides: structures, properties and gene regulation. Protein Pept Lett 12:3–11
43. Conlon JM (2008) Reflections on a systematic nomenclature for antimicrobial peptides from the skins of frogs of the family Ranidae. Peptides 29:1815–1819
44. Lata S, Sharma BK, Raghava GP (2007) Analysis and prediction of antibacterial peptides. BMC Bioinformatics 8:263
45. Xiao X, Wang P, Lin WZ, Jia JH, Chou KC (2013) iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem 436: 168–177
46. Wang G (2013) Database-guided discovery of potent peptides to combat HIV-1 or superbugs. Pharmaceuticals 6:728–758
47. O'Shea EF, O'Connor PM, O'Sullivan O, Cotter PD, Ross RP, Hill C (2013) Bactofencin a, a new type of cationic bacteriocin with unusual immunity. MBio 4:e00498–13
48. Menousek J, Mishra B, Hanke ML, Heim CE, Kielian T, Wang G (2012) Database screening and in vivo efficacy of antimicrobial peptides against methicillin-resistant *Staphylococcus aureus* USA300. Int J Antimicrob Agents 39:402–406
49. Wang G, Watson KM, Peterkofsky A, Buckheit RW Jr (2010) Identification of novel human

immunodeficiency virus type 1 inhibitory peptides based on the antimicrobial peptide database. Antimicrob Agents Chemother 54: 1343–1346

50. Loose C, Jensen K, Rigoutsos I, Stephanopoulos G (2006) A linguistic model for the rational design of antimicrobial peptides. Nature 443:867–869
51. Lam KS, Salmon SE, Hersh EM, Hruby VJ, Kazmierski WM, Knapp RJ (1991) A new type of synthetic peptide library for identifying ligand-binding activity. Nature 354:82–84
52. Cherkasov A, Hilpert K, Jenssen H, Fjell CD, Waldbrook M, Mullaly SC, Volkmer R, Hancock RE (2008) Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. ACS Chem Biol 4:65–74
53. Mishra B, Wang G (2012) The importance of amino acid composition in natural AMPs: an evolutional, structural, and functional perspective. Front Immunol 3:221
54. Mishra B, Wang G (2012) *Ab initio* design of potent anti-MRSA peptides based on database filtering technology. J Am Chem Soc 134: 12426–12429
55. Wang G (2008) Structures of human host defense cathelicidin LL-37 and its smallest antimicrobial peptide KR-12 in lipid micelles. J Biol Chem 283:32637–32643
56. Saether O, Craik DJ, Campbell ID, Sletten K, Juul J, Norman DG (1995) Elucidation of the primary and three-dimensional structure of the uterotonic polypeptide kalata B1. Biochemistry 34:4147–4158
57. Bauer F, Schweimer K, Klüver E, Conejo-Garcia JR, Forssmann WG, Rösch P, Adermann K, Sticht H (2001) Structure determination of human and murine beta-defensins reveals structural conservation in the absence of significant sequence similarity. Protein Sci 10:2470–2479
58. Rozek A, Friedrich CL, Hancock RE (2000) Structure of the bovine antimicrobial peptide indolicidin bound to dodecylphosphocholine and sodium dodecyl sulfate micelles. Biochemistry 39:15765–15774

Chapter 4

Building MHC Class II Epitope Predictor Using Machine Learning Approaches

Loan Ping Eng, Tin Wee Tan, and Joo Chuan Tong

Abstract

Identification of T-cell epitopes binding to MHC class II molecules is an important step in epitope-based vaccine development. This process has since been accelerated with the use of bioinformatics tools to aid in the prediction of peptide binding to MHC class II molecules and also to systematically scan for candidate peptides in antigenic proteins. There have been many prediction software developed over the years using various methods and algorithms and they are becoming increasingly sophisticated. Here, we illustrate the use of machine learning algorithms to train on MHC class II peptide data represented by feature vectors describing their amino acid physicochemical properties. The developed prediction model can then be used to predict new peptide data.

Key words MHC, Antigens/peptides/epitopes, CTD, Machine learning

1 Introduction

Computational prediction of T-cell epitope binding has come a long way since over a decade ago. Many software employing various methods and algorithms are now available, ranging from sequence-based prediction methods using simple sequence motifs and binding matrices, to more complex structure-based prediction methods [1–3]. The prediction of epitope binding to major histocompatibility complex (MHC) class II is of particular interest due to longer, varying length of the peptides and the open peptide-binding cleft of MHC class II molecule. The epitopes are highly promiscuous in terms of length and composition, resulting in a larger repertoire of peptides. Therefore, this poses a big challenge for building an accurate predictor for MHC class II peptide binding.

The computational field of machine learning has gained a strong foothold in many bioinformatics applications, including T-cell epitope prediction. With the accumulation of biological data, the use of machine learning techniques to learn from experience

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_4,

and continually improve prediction has been immensely useful. This chapter outlines the steps in converting variable-length peptide sequences to fixed-length feature vectors using amino acid physicochemical properties for the training of machine learning algorithms, to finally building a prediction model for MHC class II peptide binding.

2 Materials

2.1 Data

MHC class II peptide binding data can be obtained from own experiments or existing public databases such as MHCBN [4] and Immune Epitope Database (IEDB) [5]. Only protein sequences with peptide binding affinity to HLA alleles are required.

2.2 Software

The various machine learning algorithms can be applied from WEKA (http://www.cs.waikato.ac.nz/ml/weka/), an open-source data mining platform integrating various machine learning applications [6].

3 Methods

3.1 Transforming Peptide Sequence into Feature Vectors

Variable-length peptide sequence can be encoded into fixed-length feature vectors based on amino acid physicochemical properties. The 20 standard amino acids can be categorized into groups of three based on physicochemical properties of amino acids such as polarity, hydrophobicity, and normalized van der Waals volume [7–9]. The property groups are extracted from AAindex database by Tomii and Kanehisa (http://www.genome.jp/aaindex/) [10]. The amino acid groupings based on the four properties mentioned are shown in Table 1.

The first step involves transforming peptide sequence into feature vectors based on the selected amino acid properties. Here, composition-transition-distribution (CTD) method is used to describe the global composition of each amino acid property for every peptide sequence, using three global descriptors in CTD, composition (C), transition (T), and distribution (D) [7]. The method proceeds as follows:

1. Composition measures the percentage frequency of a particular amino acid property group in the peptide sequence. C can be calculated by dividing the number of amino acids in each property group by the length of the peptide sequence:

$$C = \left(\frac{n_1 \times 100}{N}, \frac{n_2 \times 100}{N}, \frac{n_3 \times 100}{N} \right),$$

where $N = \sum_{i=1}^{m} n_i$, $m = 3$ represents the number of categories for each amino acid property, and n_i is the number of amino acids in the ith group [9, 11].

Table 1
Amino acid properties divided into three divisions

Property	Divisions		
	Group 1	**Group 2**	**Group 3**
Polarity	*Polarity value 4.9–6.2* L I F W C M V Y	*Polarity value 8.0–9.2* P A T G S	*Polarity value 10.4–13.0* H Q R K N E D
Normalized van der Waals volume	*Volume range 0–2.78* G A S T D C P	*Volume range 2.95–4.0* N V E Q I L	*Volume range 4.03–8.08* M H K F R Y W
Hydrophobicity	*Polar* R K E D Q N	*Neutral* G A S T P H Y	*Hydrophobic* C V L I M F W

All 20 standard amino acids can be divided into groups of three for each amino acid property based on amino acid indices by Tomii and Kanehisa [7–10]

2. Transition can be characterized by the percent frequency of amino acids of a particular property to be followed by amino acids of another property. In other words, the features represent the percentage frequency of property i followed by property j, or j followed by i, for $i, j \in \{n_1, n_2, n_3\}$ [8]. T can be calculated as such:

$$T = \left(\frac{T_{G_1G_2} \times 100}{N-1}, \frac{T_{G_1G_3} \times 100}{N-1}, \frac{T_{G_2G_3} \times 100}{N-1} \right),$$

where $T_{G_iG_j}$ represents the occurrence of amino acid of property i followed by property j, or j followed by i, and $N - 1$ describes the total number of transitions within the peptide sequence [9, 11].

3. Distribution describes the fractions of the entire peptide sequence where the first, 25, 50, 75, and 100 % of amino acids of a particular property is placed within the peptide sequence, respectively:

$$D = (D_1, D_2, D_3),$$

$$D_i = \left(\frac{P_{i0} \times 100}{N}, \frac{P_{i25} \times 100}{N}, \frac{P_{i50} \times 100}{N}, \frac{P_{i75} \times 100}{N}, \frac{P_{i100} \times 100}{N} \right),$$

where P_{ij} (j=0, 25, 50, 75, 100) is the chain length where j % of the amino acids of property i is contained [9, 11].

From the CTD method described above, 21 descriptors can be calculated for each of the three amino acid property groups. This will yield a total of 63 feature vectors to describe each peptide sequence.

3.2 Defining Binders and Non-binders

MHC class II peptide binding data should be classified into binders and non-binders according to their experimentally measured binding affinities. Peptides are classified as binders with binding affinity of less than 500 nM, and non-binders otherwise [8].

3.3 Building a Prediction Model

At this stage, each peptide sequence would be represented by its translated feature vectors. The input data file would have to be formatted according to ARFF file format supported by WEKA as shown in Fig. 1.

Various machine learning classifiers can be used to classify epitopes into binders or non-binders, accessible via the Classify tab

```
% 1. Title: IEDB MHC Class II Allele
%
% 2. Sources: Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I,
% Salimi N, Damle R, Sette A, Peters B. The immune epitope database
% 2.0. Nucleic Acids Res. 2010 Jan;38(Database issue):D854-62. Epub
% 2009 Nov 11.
%
% 3. Allele: HLA-DPA1_01-DPB1_0401
%
@RELATION HLA-DPA1_01-DPB1_0401

@ATTRIBUTE 'polarc1' REAL
@ATTRIBUTE 'polarc2' REAL
@ATTRIBUTE 'polarc3' REAL
.
.
.
@ATTRIBUTE 'hydrod15' REAL
@ATTRIBUTE class {0, 1}

@DATA
26.67,53.33,20.0,28.57,14.29,28.57,13.33,13.33,20.0,60.0,73.33,6.667
,26.67,46.67,80.0,100.0,33.33,33.33,66.67,66.67,86.67,46.67,20.0,33.
33,28.57,14.29,14.29,6.667,40.0,53.33,80.0,100.0,60.0,60.0,73.33,73.
33,86.67,13.33,13.33,26.67,33.33,66.67,26.67,46.67,26.67,21.43,21.43
,21.43,26.67,26.67,33.33,66.67,86.67,6.667,40.0,53.33,80.0,100.0,13.
33,13.33,20.0,60.0,73.33,40.0,26.67,33.33,14.29,14.29,14.29,6.667,40
.0,46.67,93.33,100.0,60.0,60.0,73.33,80.0,86.67,13.33,13.33,26.67,33
.33,66.67,1
20.0,53.33,26.67,21.43,21.43,21.43,26.67,26.67,53.33,53.33,73.33,6.6
67,33.33,46.67,86.67,100.0,13.33,13.33,20.0,60.0,80.0,46.67,33.33,20
.0,28.57,14.29,14.29,6.667,26.67,40.0,46.67,100.0,20.0,20.0,60.0,73.
33,80.0,13.33,13.33,86.67,86.67,93.33,40.0,40.0,20.0,21.43,21.43,21.
43,13.33,20.0,60.0,86.67,93.33,6.667,33.33,40.0,66.67,100.0,26.67,26
.67,53.33,53.33,73.33,46.67,33.33,20.0,28.57,14.29,14.29,6.667,26.67
,40.0,46.67,100.0,20.0,20.0,60.0,73.33,80.0,13.33,13.33,86.67,86.67,
93.33,0
```

Fig. 1 Example of ARFF file format supported by WEKA

in Weka Explorer. Some popular algorithms with their corresponding classifiers in WEKA include (*see* **Note 1**):

1. Decision tree—classifiers\trees\J48
2. Random forest—classifiers\trees\RandomForest
3. Support vector machine—classifiers\functions\SMO
4. Artificial neural network—classifiers\functions\MultilayerPerceptron

In the default setting, a tenfold cross-validation scheme is used to train the dataset. In tenfold cross-validation, the dataset is divided randomly into ten different subsets. Of the ten subsets, one will be retained as testing data for validating the model while the other nine will be used to train the model. The process will be repeated ten times, with each subset used as testing data exactly once. Cross-validation is useful to estimate whether the algorithm is able to generalize beyond training data (*see* **Note 2**). This method is commonly used in machine learning to overcome the problem of overfitting where the model memorizes the training data perfectly rather than being able to generalize to real-world data.

The success of the prediction model depends on the optimization of the classifier. The parameters for each of the algorithms can be changed accordingly by clicking on the classifier, for example, the number of hidden layers and learning rate for artificial neural network and choice of kernel for support vector machine.

3.4 Model Evaluation

The performance of each prediction model can be evaluated from the classifier output using a number of different methods. These include prediction accuracy (ACC), sensitivity (SN), specificity (SP), precision (PR), correlation coefficient (CC), and area under receiving operating characteristic curve (AUC).

1. Prediction accuracy gives a measure of the overall accuracy of the classifier by calculating the number of correctly classified binders and non-binders over the total number of peptides:

$$\mathrm{ACC} = \left(\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \right)$$

2. Sensitivity and specificity summarize the accuracies of positive and negative predictions, respectively. SN is the ratio of binders correctly predicted among all true binders in the dataset. In contrast, SP describes the ratio of non-binders correctly predicted among all non-binders in the dataset:

$$\mathrm{SN} = \left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \right)$$

$$\mathrm{SP} = \left(\frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \right)$$

3. Precision describes the proportion of correctly predicted peptides in the total number of peptides predicted in the class. In this study, precision is the ratio of true binders in all the binders predicted by the classifier:

$$\mathrm{PR} = \left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \right)$$

4. Correlation coefficient measures the correlation between predicted and experimental data [12]. The measure has a value ranging from −1 to +1, with values closer to +1 indicating a better classifier:

$$\mathrm{CC} = \left(\frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})}} \right)$$

5. The area under ROC curve (AUC) summarizes the comparison between two ROC curves, which in this study are the two classes of binders and non-binders. AUC is defined as the probability that a randomly chosen positive example will be ranked higher than a randomly chosen negative example [8]. AUC is commonly used in machine learning for model comparison, where it describes the classifier performance over all possible thresholds for a binary classifier system.

3.5 Predicting New Data

The prediction model with the best performance chosen can finally be used for prediction of new data. The following steps can be applied for prediction of new data:

1. Transform peptide sequence into feature vectors using CTD method.
2. Convert test file into ARFF format, with the class set as "?."
3. Load prediction model into WEKA.
4. Set the new test file as supplied test set.
5. Under "More options," check the "Output predictions" box.
6. Right-click the model and select "Re-evaluate model on current test set."
7. The predicted results can be found in "Predictions on test set" in the classifier output.

4 Notes

1. In general, different prediction models can be constructed using various machine learning algorithms. The performance of each prediction model can be compared and further improved by fine-tuning the parameters.

2. The number of folds for cross-validation can be adjusted in test options depending on the sample size of the dataset. If the dataset is small, it would be more advisable to use a fivefold cross-validation.

References

1. Lafuente EM, Reche PA (2009) Prediction of MHC-peptide binding: a systematic and comprehensive overview. Curr Pharm Des 15(28): 3209–3220
2. Tong JC, Tan TW, Ranganathan S (2007) Methods and protocols for prediction of immunogenic epitopes. Brief Bioinform 8(2): 96–108
3. Wang P, Sidney J, Dow C et al (2008) A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol 4(4): e1000048
4. Lata S, Bhasin M, Raghava GP (2009) MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes. BMC Res Notes 2:61
5. Vita R, Zarebski L, Greenbaum JA et al (2010) The immune epitope database 2.0. Nucleic Acids Res 38(Database issue):D854–D862
6. Hall M, Frank E, Holmes G et al (2009) The WEKA data mining software: an update. ACM SIGKDD Explorations Newslett 11(1):10–18
7. Dubchak I, Muchnik I, Mayor C et al (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. Proteins 35(4): 401–407
8. El-Manzalawy Y, Dobbs D, Honavar V (2008) On evaluating MHC-II binding peptide prediction methods. PLoS One 3(9):e3268
9. Li ZR, Lin HH, Han LY et al (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 34(Web Server issue): W32–W37
10. Tomii K, Kanehisa M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 9(1):27–36
11. Cui J, Han LY, Lin HH et al (2007) Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. Mol Immunol 44(5): 866–877
12. Gowthaman U, Agrewala JN (2008) In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion. J Proteome Res 7(1):154–163

Chapter 5

Brownian Dynamics Simulation of Peptides with the University of Houston Brownian Dynamics (UHBD) Program

Tongye Shen and Chung F. Wong

Abstract

This chapter provides the background theory and a practical protocol for performing Brownian dynamics simulation of peptides. Brownian dynamics simulation represents a complementary approach to Monte Carlo and molecular dynamics methods. Unlike Monte Carlo methods, it could provide dynamical information in a timescale longer than the momentum relaxation time. On the other hand, it is faster than molecular dynamics by approximating the solvent by a continuum and by operating in the over-damped limit. This chapter introduces the use of the University of Houston Brownian Dynamics (UHBD) program [1, 2] to perform Brownian dynamics simulation on peptides.

Key words Brownian dynamics simulation, UHBD program, Helix-capping motifs, Conformational distribution of peptides

1 Introduction

Multiple computational methods for studying the conformations of peptides are available. This chapter focuses on a less known method, Brownian dynamics simulation, which provides a useful complement to other methods such as molecular dynamics and Monte Carlo. Unlike typical Monte Carlo simulations, Brownian dynamics simulations can provide important dynamical information in a timescale longer than the momentum-relaxation time. On the other hand, it is cheaper to use than classical molecular dynamics simulation by not explicitly including solvent molecules and ambient ions and by ignoring extremely short-time dynamics.

Brownian dynamics simulation operates in the over-damped limit. The Ermak-McCammon algorithm [3], which was derived from the Langevin equation [4], provides a numerical recipe for

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_5,

propagating the dynamics of a peptide in this limit. The stochastic equation of motion describing this process reads

$$F_i^{\text{solute}} + F_i^{\text{frict.}} + F_i^{\text{stoc.}} = m_i a_i \tag{1}$$

in which F_i^{solute} is the solvent-mediated systematic force acting on atom i of a peptide from all the other atoms in the peptide, $F_i^{\text{frict.}}$ is the frictional force introduced by the solvent, $F_i^{\text{stoc.}}$ represents the random force acting on atom i due to the random collisions of the solvent molecules with the atom, and m_i and a_i are the mass and acceleration of the atom, respectively.

The frictional force is related to the atomic velocity by

$$F_i^{\text{frict.}} = -\gamma \vec{v}_i \tag{2}$$

where γ is the frictional coefficient of the solvent and $\vec{v}_i$ the velocity of atom i.

Ermak and McCammon derived the following solution for the Langevin equation in the over-damped limit [3]:

$$\vec{r}\left(t+\delta t\right) = \vec{r}\left(t\right) + \left(RT\right)^{-1} \underline{\underline{D}} \vec{F} \delta t + \vec{R} \tag{3}$$

where $\vec{r}\left(t\right)$ represents a vector containing the atomic coordinates at time t, $\vec{F}$ a vector of systematic forces acting on the atoms, and δt the time step. $\vec{R}$ is usually chosen as a vector of random numbers satisfying Gaussian distribution with the first and second moments being the conditions

$$< \vec{R} >= \vec{0} \tag{4}$$

$$< \vec{R}\vec{R}^T >= 2\underline{\underline{D}}\delta t \tag{5}$$

Here, the diffusion tensor D is often approximated by a diagonal matrix with every nonzero diagonal element set to

$$D_i = k_{\text{B}} T / 6\pi\eta b_i \tag{6}$$

in which η is the solvent viscosity, and b_i is the hydrodynamic radius of atom i.

1.1 Constraining Bond Lengths with LINCS [5] to Reduce Simulation Time by Using Larger Time Steps

The time step δt in the Ermak-McCammon equation for propagating the dynamics of a system needs to be small enough so that the systematic force $\vec{F}$ is approximately constant during the evolution of each time step. Thus, the time step is often determined by the highest frequency motion in the system. For peptide systems, this motion is associated with the rapid bond vibration. However, as the amplitude of bond vibration is small in amplitude and does not contribute to the important conformational dynamics of peptides at

physiological temperatures, the bonds can be considered to be rigid. Therefore, one can reduce simulation time of a peptide by not following the rapid vibration of the bonds so that larger time steps can be used. However, the bonds need to be properly constrained to their equilibrium values during a simulation. Different methods have been introduced to constrain bond lengths, with the SHAKE method [6] being one of the first introduced. UHBD [1, 2] employs a newer method called the LINCS [5] that requires less computational time than SHAKE does.

To further decrease computational time, UHBD uses a variable rather than fixed time step. Constraining bond lengths and operating in the over-damped limit make it feasible to use larger time steps in BD simulations, but large time steps could introduce unfavorable atomic clashes occasionally. To avoid this problem without sacrificing the use of large time steps, UHBD uses a large time step, such as 10 fs, most of the time, but switches to a smaller time step whenever the van der Waals energy of a configuration is larger than a user-chosen threshold. The normal time step is split into n smaller ones according to $\delta t' = \delta t/n$. The proper threshold and normal time step are system, model, and temperature dependent. For simulation of short peptides near physiological temperatures, the threshold is of the order of tens of kcal/mol and δt is about 10 fs. The small time steps were used less than 1 % of the time in our earlier simulations of tetrapeptides [7, 8]. This adaptive time step method permits a time step δt five or more times larger than that in a constant time step algorithm.

1.2 Distance-Dependent Dielectric Model

Because Brownian dynamics simulation reduces computational time by not including solvent molecules explicitly but treating the solvent as a continuum, suitable models need to be used to calculate solvent-mediated interactions. The distance-dependent dielectric model is a simple model that is still popular today. In this model, the electrostatic interactions between two atoms are scaled by a dielectric function $\varepsilon(r)$ that depends on the distance r between the two atoms. A commonly used function uses the form $\varepsilon(r) = mr$ in which m varies from 1 to about 5 Å^{-1}. Sigmoidal functions have also been used but have not yet been implemented in UHBD [9, 10].

1.3 Generalized Born Model

A more sophisticated solvent dielectric model is the generalized Born model originally introduced by Still [11]. Different improved variants appeared later on. UHBD has implemented a model by Qiu et al. [12]. It uses the following formula to estimate the electrostatic contributions to the solvation energy:

$$-\frac{1}{2}\left(1-\frac{1}{\varepsilon}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2 + a_{ij}^2 e^{-B}}} \tag{7}$$

where

$$a_{ij}^2 = a_i a_j \tag{8}$$

and

$$B = r_{ij}^2 / \left(2a_{ij}\right)^2 \tag{9}$$

in which a_i is the Born radius of atom i and r_{ij} is the distance between atoms i and j. The generalized Born equation (Eq. 7) includes a Born-like term

$$-\frac{1}{2}\left(1-\frac{1}{\varepsilon}\right)\frac{q_i^2}{a_i} \tag{10}$$

when $i=j$ and a Coulombic-like interaction term when $i\neq j$ except that the distance between two charges is replaced by an effective distance $\sqrt{r_{ij}^2 + a_{ij}^2 e^{-B}}$ that is dependent on the Born radii. Qiu et al. [12] determined the effective Born radius of an atom in a molecule or molecular complex by using the equation

$$G_{\text{pol}} = -166.0\left(1-\frac{1}{\varepsilon}\right)\frac{q_i^2}{a_i} = \left(1-\frac{1}{\varepsilon}\right)\left(\frac{-166.0}{R_{\text{vdw}} + R_{\text{off}} + P_1} + \sum_{j\in\text{stretch}} \frac{P_2 V_j}{r_{ij}^4} + \sum_{j\in\text{bend}} \frac{P_3 V_j}{r_{ij}^4} + \sum_{j\in\text{nonbonded}} \frac{f_{\text{contact}} P_4 V_j}{r_{ij}^4}\right) \tag{11}$$

where R_{vdw} is the van der Waals radius of the ion, R_{off} is a dielectric offset distance, V_j is the volume of a surrounding atom j of atom i in the molecule or molecular complex, and P_1, P_2, P_3, and P_4 are four parameters of the model. This equation relies on the assumption that the presence of another atom j at a distance r_{ij} from atom i displaces a high dielectric region of the solvent with volume V_j and changes the solvation energy by an amount proportional to the volume and inversely proportional to the fourth power of r_{ij} (based on charge-induced dipole arguments). Different scaling parameters P_2, P_3, and P_4 are used for stretching, bending, and nonbonded terms because atomic overlaps differ for these terms. For the nonbonded terms, an extra close contact function f_{contact} is used to reduce the effective volume when atom j gets too close to the atom for which an effective Born radius is sought. Another parameter P_5 is used to fine-tune the contact function f_{contact}. The parameters P_1, P_2, P_3, P_4, and P_5 are determined by requiring Eq. 11 to give the same polarization energy obtained by solving the Poisson equation for a "representative" set of molecules. This analytical equation allows a Born radius to be determined quickly. David et al. [13] found that this generalized Born model did not work as well for proteins as for smaller molecules. However, they suggested that a special set of parameters could be developed for a

specific protein if one intends to study the protein in detail. Methods [14, 15] adding correction terms to the classical charge-induced dipole interactions have also been introduced to further improve the performance of the generalized Born model.

1.4 Choice of Diffusion Constants

The diffusion coefficients of individual atoms can be assigned according to

$$D_i = w_i \frac{k_B T}{6\pi\eta\left(r_i + 1.4\text{Å}\right)} \tag{12}$$

where k_B is the Boltzmann constant and T is the absolute temperature, and η is the viscosity of water: 0.891 cp. The weighting factor w_i can be chosen to be 2.5 for hydrogen and 1 for other atoms.

This is modified from the Stokes-Einstein equation with the hydrodynamic radius of each atom chosen as its van der Waals radius plus 1.4 Å, the approximate radius of a water molecule. Light atoms are given a larger weight to increase their diffusion coefficients. This choice stems from a previous simulation study [16]. However, the choice of diffusion coefficients only affects dynamical properties, and does not affect static properties.

2 Materials

An extension of the University of Houston Brownian Dynamics (UHBD) program [2, 17] is used in the protocol described here. The UHBD program contains many features such as simulating the diffusional encounter between enzyme-substrate pairs, calculating the electrostatic potential around a protein, and computing the binding affinity between two molecules. However, this chapter focuses on using it to perform Brownian dynamics simulation of peptides at atomic resolution.

UHBD runs on different variations of the UNIX/Linux platforms. To install UHBD, one first selects the architecture to which the program is installed. Architecture-specific information is stored in the subdirectory src with filenames M.xxx.src. For example, M.sgi.src contains information for installing UHBD on a Silicon Graphics IRIX workstation, M.gnu.src for linux machines using the gnu compiler, and M.ifc.src for using the Intel Fortran compiler (*see* **Note 1**).

One then issues the following commands:

make ARCH=xxx makefiles

make ARCH=xxx uhbd

(where xxx is the same xxx in M.xxx.src)

make ARCH=xxx programs (also compiles some accessory programs such as Top2.f below)

This will create executables under the subdirectory bin.xxx such as bin.gnu. It is useful to create a symbolic link in a directory that is on every user's path. For example,

ln –s bin.gnu /usr/local/bin

However, this requires permission to write to the directory /usr/local/bin. If one does not have such permission, one can add bin.gnu to one's path.

After this is done, one can invoke UHBD by typing uhbd.

3 Methods

This example demonstrates how to set up a Brownian dynamics simulation of a tetrapeptide with UHBD in a UNIX/Linux operating system.

3.1 Generate a Topology File

Before a dynamical simulation of a peptide can be performed, one first needs to generate a topology file to let UHBD know the starting structure of the peptide and the parameters to use in various energy terms. These energy terms are composed of two groups: bonded and nonbonded. Bonded terms involve two or more atoms connected together by one or more covalent bonds; nonbonded terms describe electrostatic and van der Waals interactions between atoms that are not forming covalent bonds with each other. As described in Subheading 1, the electrostatic terms utilize different models, such as distance-dependent dielectric model and the generalized Born (GB) model, to include solvation effects.

UHBD provides a utility program Top2.f to generate a topology file from a protein structure file (PSF) created by the Chemistry at HARvard Macromolecular Mechanics (CHARMM) program [18, 19]. CHARMM generates a PSF file based on a database of amino acids. A peptide with any sequence can be constructed by linking different amino acids together according to the sequence. The database of amino acids is stored in a residue topology file (RTF). In the RTF file, the topology of each amino acid is described so that a computer program knows the bond connectivity of the amino acid. The atomic partial charge on each atom of an amino acid is also contained in this file. The atomic partial charges are used for calculating electrostatic interactions. Each atom is also assigned an atom type that defines what parameters to use to calculate the Lennard-Jones interactions between the atom and other nonbonded atoms. The atom types also select the proper parameters to use in the bonded terms such as equilibrium bond lengths and the associated force constants. These parameters are stored in other parameter (PRM) files.

To generate a protein structure file (PSF) from the residue topology file (RTF) of amino acids and the parameter files, one first sets up an input file that will be used by the CHARMM program.

An example input file is the following (with comments preceded by an exclamation mark):

```
BOMLEV -2
! determines when the program stops when it encounters errors. It takes a range of
! number from –5 to 5. The program will stop with less severe errors when BOMLEV is
! set to a larger value
OPEN READ UNIT 24 CARD NAME "MASSES.RTF"
! open the file MASSES.RTF for reading on unit 24
! This file contains the masses of different atoms
READ RTF UNIT 24 CARD
! read data in RTF file format from the file MASSES.RTF
CLOSE UNIT 24
! close unit 24 so that it can be used for other I/O
OPEN READ UNIT 24 CARD NAME "AMINO.RTF"
! open the file AMINO.RTF for reading on unit 24
! This file contains the topology, atomic partial charges, and atom types of each amino
! acid
READ RTF UNIT 24 CARD APPE
! read data in RTF format from the file AMINO.RTF
CLOSE UNIT 24
! close unit 24 so that it can be used for other I/O
! The following three lines read the file PARM.PRM containing the
! parameters of different energy terms
OPEN READ UNIT 12 CARD NAME "PARM.PRM"
READ PARA UNIT 12 CARD
CLOSE UNIT 12
! Read the sequence of a tetrapeptide (3 alanines followed by
! an aspartate:
READ SEQUENCE CARD
* AAAD ! title card
* ! title card
4 ! four amino acids are read
ALA ALA ALA ASP ! sequence of the tetrapeptide
! Generate a segment of the protein structure file (psf) with the
```

```
! id PROT using the amino acid sequence just read in
! Make the N terminus a positively charged ammonium group
! Make the C terminus a negatively charged carboxylate group:
GENE PROT SETU FIRST NTER LAST CTER
! change the N terminus from a positively charged ammonium
    group
! to a neutral acetamide group:
PATCH NACT PROT 1 WARN SETU
! the first residue number is 1
! block the C terminus with a N-methyl amide:
PATCH CMAM PROT 4 WARN SETUP
! the last residue number is 4 for this tetrapeptide
! open unit 30, associate it with the file SAAD.psf and write the
! topology of the tetrapeptide SAAD with the associated
    parameters
! of the interaction potential to the file. Close unit 30 after
    writing:
OPEN WRITE UNIT 30 CARD NAME SAAD.psf
WRITE PSF CARD UNIT 30
CLOSE UNIT 30
! signal the end of the input stream:
STOP
```

One can then invoke CHARMM to generate a protein structure file for the tetrapeptide using the input file just created. Suppose the name of this input file is named SAAD_charmm.inp and the executable of the CHARMM program is named charmm, one can issue the command

```
charmm < SAAD_charmm.inp > SAAD.out
```

where SAAD.out is a file that holds various information that the CHARMM program prints out when it runs.

One can then use the utility program Top2.f in UHBD to generate a topology file from the protein structure file SAAD.psf. The resulting topology file can be used by UHBD to perform a Brownian dynamics simulation on the tetrapeptide. To do this, one first creates an input file, say Top2_SAAD.inp, containing the following lines:

```
SAAD.psf
SAAD.top
```

Directory_containing_parameter_files provided by UHBD (MASSES.RTF PARM.ANGLE PARM.ATOM PARM.BOND PARM.DIHEDRAL PARM.IMPROPER)

(*See* **Note 2**.)

Then run the program as

Top2 < Top2_SAAD.inp > Top2.log

where Top2 is the executable after the Fortran utility program Top2.f has been compiled. Top2.log contains information printed out by UHBD during the run.

3.2 Generate an Initial Structure for the Tetrapeptide

Before a simulation can be performed, one needs to provide an initial structure for the tetrapeptide. One can use different commercial or academic programs to generate this structure from the sequence of the tetrapeptide. Swiss PDB viewer [20] provides an example of an academic program that has such capability. It can write the coordinate file in the Protein Data Bank (PDB) format [21] that UHBD can read.

3.3 Performing a Brownian Dynamics Simulation of the Tetrapeptide Using UHBD

After the topology file and an initial structure have been created, a Brownian dynamics simulation can be carried out using an input file directing how the simulation should be performed. Here is the content of an example input file, named SAAD.inp:

```
!
! Read the molecular coordinates in pdb format:
!
read mol 1 file SAAD.pdb pdb end
! molecule 1 now represents the tetrapeptide
! When restarting a run to lengthen a simulation,
! use something like the following instead.
! read
! mol 1 file SAAD_100ps.pdb vpdb
! end
! This will use the file containing the atomic coordinates and velocities
! from an earlier run to continue a Brownian dynamics simulation to a longer time
! In this case, SAAD_100ps.pdb is the last coordinate/velocity file written from the
! previous run
! read the topology file:
!
```

```
rdtop new file SAAD.top mol 1 end
!
! set up a nonbonded pair list using a cutoff distance of 10.0 Å
nblist cut 10.0 end
! see Note 3
! Initialize the Generalized Born model to calculate the solvent mediated
! electrostatics interactions:
gbn init ideal full ndetail mol1 pdie 4. sdie 78. end
! gbn is the keyword for invoking the Generalized Born model
! init: initialize potential parameters
! ideal: use ideal geometry in calculating Born radii
! full: use full GB derivatives rather than the partial approximation that reduces
! computational time
! ndetail: do not print details (otherwise, a lot of information is printed out
! during a Brownian dynamics run
! mol1: operate on molecule 1 (the only molecule in this example)
! pdie: dielectric constant of the peptide
! sdie: dielectric constant of the water solvent
! end: end the input for the Generalized Born driver
! check the energy of the molecule
energy ! invoke the energy driver to calculate energy
all ! include all contributions
ndbf ! exclude the dielectric-boundary energy term
! see Note 4
nqe ! exclude the electrostatic energy term
! see Note 5
!
gbn ! calculate GB energy
end ! end the input deck for the energy command
! Name the file (SAAD100ps in the following example)
! to which the simulation trajectory will be written:
maktrj
unit 1 file SAAD100ps ntrj 10
end
```

```
! ntrj 10 means coordinates and velocities are written out every
    10 steps
!
! Perform Brownian dynamics calculation
!
md rest ! signal an Brownian dynamics run using restart (rest)
    option
bdlin ! perform Brownian dynamics simulation using LINCS
temp 300.0 ! temperature in K
ndbf ! exclude dielectric boundary forces
!
!see Note 6
!
nqe ! do not calculate electrostatic forces using the PB model
!
! see Note 7
!
gamma 10.0 ! set frictional coefficient to 10
surf ! calculate surface area force
! see Note 8
vdw ! calculate van der Waals interactions
eel  ! Calculate Coulombic interactions
coef 0.025 ! multiplicative factor for surface area term
seed 0680865 683764 25574 60865 ! random number generating
    seeds
dt 0.005 ! time step in picoseconds
nste 100000 ! 100000 times steps will be performed
prnt 50 ! print results every 10 steps
end
stop
```

To perform a UHBD run using this input file:

```
uhbd < SAAD.inp > SAAD.out
```

where SAAD.out contains information printed out by UHBD during a run.

4 Notes

1. UHBD requires the following programs installed on your machines: awk, sed, and makedepend.
2. SAAD.top contains the topology file required by UHBD to perform a Brownian dynamics simulation.
3. The electrostatic and van der Waals interactions will only be calculated between two atoms in the nonbonded list. Two atoms separated by a distance larger than 10.0 Å are not in the list and thus their nonbonded interactions are not calculated. Therefore, this list needs to be regenerated every certain number of steps as the structure of the peptide changes with time.
4. This is used by the Poisson-Boltzmann, not the generalized Born, model. The Poisson-Boltzmann model was the first continuum electrostatics model implemented into UHBD before the generalized Born model was implemented.
5. Because this is calculated by the generalized Born engine instead of the Poisson-Boltzmann engine.
6. This is used in the Poisson-Boltzmann model, not in the GB model, and is therefore turned off here.
7. Because we shall use the eel flag below that directs UHBD to calculate the Coulombic interactions using analytically rather than solving the Poisson equation numerically.
8. UHBD uses a surface area-dependent term to describe the energy costs for creating a cavity in which a solute molecule is immersed:

$$\Delta G_{\text{cavity}} = \text{coef} \times \text{Surface\_Area\_of\_the\_solute}$$

References

1. Madura J, McCammon JA (1989) Brownian dynamics simulation of diffusional encounters between triose phosphate isomerase and D-glyceraldehyde phosphate. J Phys Chem 93: 7285–7287
2. Madura JD, Briggs JM, Wade RC, Davis ME, Luty BA et al (1995) Electrostatics and diffusion of molecules in solution - simulations with the University of Houston Brownian Dynamics Program. Comput Phys Commun 91:57–95
3. Ermak DL, McCammon JA (1978) Brownian dynamics with hydrodynamic interactions. J Chem Phys 69:1352–1360
4. Langevin P (1908) On the theory of Brownian motion. R Acad Sci (Paris) 146:530
5. Hess B, Bekker H, Berendsen HJC, Fraaije J (1997) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 18: 1463–1472
6. 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 23:327–341
7. Shen T, Wong CF, McCammon JA (2003) Brownian dynamics simulation of helix-capping motifs. Biopolymers 70:252–259
8. Shen TY, Wong CF, McCammon JA (2001) Atomistic Brownian dynamics simulation of peptide phosphorylation. J Am Chem Soc 123: 9107–9111

9. Mehler EL (1990) Comparison of dielectric response models for simulating electrostatic effects in proteins. Protein Eng 3:415–417
10. Mehler EL, Solmajer T (1991) Electrostatic effects in proteins: comparison of dielectric and charge models. Protein Eng 4:903–910
11. 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
12. Qiu D, Shenkin PS, Hollinger FP, Still WC (1997) The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. J Phys Chem A 101:3005–3014
13. David L, Luo R, Gilson MK (2000) Comparison of generalized born and Poisson models: energetics and dynamics of HIV protease. J Comput Chem 21:295–309
14. Ghosh A, Rapp CS, Friesner RA (1998) Generalized born model based on a surface integral formulation. J Phys Chem B 102:10983–10990
15. Lee MS, Salsbury FR, Brooks CL (2002) Novel generalized Born methods. J Chem Phys 116:10606–10614
16. Smart JL, Marrone TJ, McCammon JA (1997) Conformational sampling with Poisson-Boltzmann forces and a stochastic dynamics/Monte Carlo method: application to alanine dipeptide. J Comput Chem 18:1750–1759
17. Davis ME, Madura JD, Luty BA, McCammon JA (1991) Electrostatics and diffusion of molecules in solution - simulations with the University-of-Houston-Brownian Dynamics Program. Comput Phys Commun 62: 187–197
18. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ et al (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30:1545–1614
19. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S et al (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4:187–217
20. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723
21. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD et al (1977) The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542

Chapter 6

Computational Prediction of Short Linear Motifs from Protein Sequences

Richard J. Edwards and Nicolas Palopoli

Abstract

Short Linear Motifs (SLiMs) are functional protein microdomains that typically mediate interactions between a short linear region in one protein and a globular domain in another. SLiMs usually occur in structurally disordered regions and mediate low affinity interactions. Most SLiMs are 3–15 amino acids in length and have 2–5 defined positions, making them highly likely to occur by chance and extremely difficult to identify. Nevertheless, our knowledge of SLiMs and capacity to predict them from protein sequence data using computational methods has advanced dramatically over the past decade. By considering the biological, structural, and evolutionary context of SLiM occurrences, it is possible to differentiate functional instances from chance matches in many cases and to identify new regions of proteins that have the features consistent with a SLiM-mediated interaction. Their simplicity also makes SLiMs evolutionarily labile and prone to independent origins on different sequence backgrounds through convergent evolution, which can be exploited for predicting novel SLiMs in proteins that share a function or interaction partner.

In this review, we explore our current knowledge of SLiMs and how it can be applied to the task of predicting them computationally from protein sequences. Rather than focusing on specific SLiM prediction tools, we provide an overview of the methods available and concentrate on principles that should continue to be paramount even in the light of future developments. We consider the relative merits of using regular expressions or profiles for SLiM discovery and discuss the main considerations for both predicting new instances of known SLiMs, and de novo prediction of novel SLiMs. In particular, we highlight the importance of correctly modelling evolutionary relationships and the probability of false positive predictions.

Key words Short linear motifs, SLiM, Motif discovery, Protein-protein interactions, Posttranslational modifications, Intrinsically disordered proteins, Regular expressions, Sequence profiles, Sequence motifs

Abbreviations

DMI	Domain-motif interaction
ELM	Eukaryotic linear motif
FPR	False positive rate
GO	Gene ontology
HMM	Hidden Markov model
IDP	Intrinsically disordered protein

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_6,

IDR	Intrinsically disordered region
LDMS	(l, d) motif search
MnM	Minimotif miner
MoRF	Molecular recognition feature
MST	Minimum spanning tree
PPI	Protein-protein interaction
PSSM	Position-specific scoring matrix
PTM	Posttranslational modification
Regex	Regular expression
SLiM	Short linear motif
TPR	True positive rate

1 Introduction

Short Linear Motifs (SLiMs) are a set of protein sequence features with specific attributes that the name suggests [1]:

1. *Short*. SLiMs are typically 3–15 amino acids in length, often with fewer than six (and as few as two) residues that are key to the function.
2. *Linear*. SLiMs are found in linear stretches of protein, typically in intrinsically disordered regions (IDR), and their (unbound) three-dimensional structure is not considered crucial for their activity.
3. *Motif*. SLiMs contain specific residues that are important for function and, as such, are amenable to sequence analysis tools and representations.

The functional relevance of short linear peptides has been recognized for decades (e.g., [2, 3]) but it was only in the early twenty-first century that SLiMs were recognized as a discrete class of element worthy of study in its own right [4, 5]. SLiMs are now recognized to be one of the key components in the cell's repertoire of protein-protein interactions (PPI), mediating a specific type [6] that we refer to here as a domain-motif interaction (DMI). Although it is hard to make a good estimate, it has been suggested that something in the order of 15–40 % of the PPI in a cell may be DMI [7]—a number which is likely to be enriched in signalling networks [8]. In the 10 years that the Eukaryotic Linear Motif (ELM) database has been collecting and curating SLiMs, the number of distinct classes has increased from approx. 80 in 2003 [5] to approx. 200 in 2013 [9] and is set to continue to rise. The latest release of Minimotif Miner (MnM) includes 880 consensus SLiMs [10], although this number is somewhat inflated by the way that length variability and redundancy is handled in the database. This suggests that even though progress in the field has led to the accumulation of much data on SLiMs, there is still much room for discovery of new instances of known and novel motifs.

SLiMs are involved in an incredibly diverse range of biological processes, including cell signalling, posttranslational modification, subcellular localisation, gene expression, membrane binding, protein folding, and cell adhesion [1, 8, 9, 11–14]. SLiMs usually bind with low affinity [8], making them ideal components to establish quick or transient responses. However, many motifs (of the same or different type) can co-occur, acting synergistically to give higher binding affinities [6, 8]. SLiMs also play an important role in disease; not only are they involved in core biological processes that can affect health if they go wrong but the evolutionary plasticity of SLiMs makes them ideal targets for exploitation by viruses via convergent evolution [15, 16].

Methods for SLiM prediction are under constant refinement and development and so this review is neither intended as an explicit "how to" guide to SLiM discovery nor an exhaustive list of methods and tools. Instead, we give an overview of the considerations that need to be made during such analyses, with examples from the literature, and some thoughts on future developments. This review highlights a selection of tools that illustrate key aspects of computational SLiM discovery. A particular focus is given to the tools of the SLiMSuite package, which is specifically geared to the analysis of SLiMs, including some tools that have not previously been published (Table 1). Additional SLiM prediction tools can be found in reviews by Diella et al. [8] and Davey et al. [11].

1.1 SLiM Terminology

The terminology related to SLiM analyses can be confusing because it uses a number of different terms from both biology and computing, some of which have developed their own SLiM-specific meanings. The main terms used in this chapter are therefore listed in the glossary (Table 2; *see* also [8]). We have made every effort to be consistent within the chapter but readers should be aware that some of the terms used can have alternative meanings in related disciplines. The term "motif" is particularly widespread and has a number of discipline-specific meanings. Within this review, "motif" refers to a short sequence motif unless otherwise specified.

1.2 SLiM Notation

A standard notation has been suggested to denote SLiMs in the written literature; *see* [4]. This is not universally applied and variation in notation can be found among publications, even when they describe the same motif [11]. Instead, there are two main classes of motif representation that are commonly used for computational analysis of SLiMs: regular expressions and sequence profiles, referred to in later sections simply as "regex" and "profile", respectively. The former are simple human- and machine-readable qualitative representations of which amino acids are tolerated at which positions in the SLiM. The latter, which for the purposes of this review includes position (specific) scoring/weight matrices

Table 1
The main tools of the SLiMSuite bioinformatics package

Tool	Ref.	Description	Web[a]
Compari Motif	[69]	A unique motif-motif comparison tool for identifying similar SLiMs. Used for clustering results of predictions and identifying known motifs.	Y
GABLAM	[77]	BLAST-based protein similarity scoring and clustering. Used for (Q)SLiMFinder and SLiMProb adjustments for evolutionary relationships.	N
GOPHER	[42]	Automated orthologue prediction and alignment algorithm. Used for conservation-based masking ((Q)SLiMFinder/SLiMProb) and prediction (SLiMPrints).	Y
PRESTO	[b]	Forerunner of SLiMSearch (now SLiMProb). A tool for searching predefined SLiMs against a protein dataset. Does not include overrepresentation/underrepresentation statistics but allows mismatches and more flexible SLiM definitions.	N
QSLiM Finder	[41][b]	Query-based variant of SLiMFinder with increased sensitivity and specificity, ideal for SLiM discovery from host-pathogen interactions or where at least one interaction is established experimentally.	N
SLiMBench	[b]	A new tool for creating and assessing de novo SLiM prediction benchmarking datasets.	N
SLiMdb	[63]	Interactive web pages to explore results of interactome-wide de novo SLiM prediction in humans, with links to other SLiMSuite tools and online public resources.	Y[c]
SLiMDisc	[42, 77]	One of the first de novo SLiM prediction tools that corrected for evolutionary relationships. Based on heuristic ranking of overrepresented motifs in unrelated proteins.	Y
SLiMFinder	[41, 100]	The first de novo SLiM prediction based on a statistical model of overrepresented motifs in unrelated proteins. Repeatedly achieves the greatest specificity in benchmarking.	Y
SLiMMaker	[b]	A simple tool for converting aligned peptides or SLiM occurrences into a regular expression motif.	Y
SLiMPred	[129]	Machine Learning de novo SLiM/MoRF prediction in single proteins based on known motif attributes.	Y[c]
SLiMPrints	[97]	Novel de novo SLiM/MoRF prediction in single proteins from statistical clustering of conserved disordered residues.	Y[c]
SLiMProb	[25]	Unique tool providing biological context (disorder and conservation) for searches of predefined SLiMs along with underrepresentation and overrepresentation statistics, correcting for evolutionary relationships. Formerly called SLiMSearch 1.x but renamed to avoid confusion with SLiMSearch2.	Y
SLiM Search2	[68]	Advanced biological context (disorder, conservation, and protein features), and ranking for proteome-wide searches of predefined motifs. Provides simple enrichment statistics for PPI partners and GO terms.	Y[c]

[a]Webserver available at http://bioware.ucd.ie/
[b]Not published at time of press. See citation details at: http://bioware.soton.ac.uk/
[c]Webserver only. Not part of SLiMSuite download

Table 2
Glossary of key terms

Term	Related terms	Description
Convergent evolution	Molecular mimicry	Independent evolutionary origins of the same function or motif on different genetic backgrounds.
Degenerate	Ambiguous	A SLiM position that can have 2+ different amino acids.
Divergent evolution	Conservation	The accumulation of differences over time following shared ancestry. Where such differences are selected against (purifying selection) sequence conservation will be seen.
Domain-Motif Interaction	DMI	PPI mediated by a SLiM in one protein and a SLiM-binding domain in the other.
Intrinsically Disordered Protein/Region	IDP/IDR	A protein/region that lacks a stable three-dimensional structure in the unbound state.
Instance	Occurrence	A single observation of a SLiM in a single protein.
(l, d) Motif Search	LDMS, (l, d) challenge problem, planted motif search	Motif search algorithms that search for recurring motifs of total length l with up to d mismatches in each occurrence.
MoRF	MoRE	Molecular Recognition Feature/Element. Short to medium-length, intrinsically disordered protein regions that mediate PPI via disorder-to-order transitions.
Pattern	Motif definition	The regular expression that defines a motif.
Posttranslational Modification	PTM	A chemical modification of an amino acid that alters its properties, such as phosphorylation of serine, threonine or tyrosine.
Profile	PSSM, PSWM, PWM, HMM	An extended representation of a sequence where each position accounts for variability between elements of the alphabet. Also known as a Position (Specific) Scoring/Weight Matrix (PSSM/PSWM/PWM). For the purposes of this review, hidden Markov models are also referred to under the "profile" umbrella.
Protein-Protein Interaction	PPI	A physical interaction between two proteins.
Regular Expression	Regex, PROSITE pattern	A common programming notation for string (text) patterns. For the purposes of this review, variants on the standard regular expression notation are included under the "regex" umbrella.
Short Linear Motif	SLiM, Linear Motif, LM, Minimotif	A short (typically <15 aa) linear stretch of protein sequence with specific residues important for function. Within this review, "motif" refers to a SLiM unless otherwise specified.
Support	UP Support	The number of different proteins that contain a given SLiM. "UP" indicates that this is the number of *unrelated* proteins.
Wildcard		A position in a SLiM that can be any amino acid.

(PSSM/PSWM/PWM) and hidden Markov models (HMM), expand on the simplicity of regular expressions by adding a quantitative dimension.

1.2.1 Regex Representations of SLiMs

The main elements of regular expressions are provided in Table 3. Evidence for SLiMs and the contribution of individual residues to their function comes from a variety of sources but is essentially either positive (specific residues are critical for function) or negative (presence of specific residues ablates function). At the two extremes, presence of a single specific amino acid side chain can be necessary for function or sufficient to block binding. Where a SLiM

Table 3
SLiM regex elements

Regex	PROSITE	MnM	SLiMSuite	Description
A	-A-	A	A	A single fixed amino acid, A using standard IUPAC letters.
[ILV]	-[ILV]-	[ILV]	[ILV]	Either I, L or V. Can have any number of possible amino acids.
[^P] or [^DE]	-{P}- or -{DE}-		[^P] or [^DE]	Exclude one or more amino acids.
.	-x-	X	X or .	Wildcard. Any amino acid.
.{n}	-x(n)-		.{n} or X{n}	A repeat of *n* wildcard positions.
.{m,n}	-x(m,n)-		.{m,n} or X{m,n}	A repeat of at least *m* and at most *n* wildcard positions. (*m* can be zero.)
^	<	<	^	N-terminus of protein.
$	>	>	$	C-terminus of protein.
(p1|p2)			(p1|p2)	Either regex pattern p1 or p2.
r{n}	r(m)		r{n}	*n* repetitions of *r*, where *r* is one of the above regex elements.
r{m,n}	r(m,n)		r{m,n}	At least *m* and up to *n* repetitions of *r*, where *r* is one of the above regex elements.
			<r:n:m>	At least *m* of a stretch of *n* residues must match *r*, where *r* is one of the above regex elements (single amino acid, ambiguity or exclusion list).
			<r:n:m:b>	Exactly *m* of a stretch of *n* residues must match *r* and the rest must match *b*, where *r* and *b* are each one of the above regex elements.
			(ABC)	A, B and C in any order.

forms a helical structure upon binding, for example, the presence of a proline may disrupt this. In between these extremes, a number of different amino acids may be tolerated at a given position and still give rise to a functional SLiM instance. Such positions are referred to as "degenerate" or "ambiguous" and consist of sets of amino acids with certain common properties, such as positive charge. Fully degenerate positions that can tolerate any amino acid are referred to as "wildcards" and typically represented with the symbols "." or "X". Sometimes these can also be referred to as "gaps" in a motif, which can be confusing to the unwary and have nothing to do with gaps (insertion/deletion events) in sequence alignments. When regular expressions are derived from sequence alignments, indels in the latter are generally represented by flexible-length wildcards (Table 3).

Regular expressions are purely qualitative, which makes them easy to model and amenable to fairly simple but effective statistics [17]. Another advantage of regular expressions is that they already form part of numerous programming and scripting languages and can therefore be used for simple computational SLiM discovery with minimal overheads. It should be noted that the PROSITE [18] and MnM [10] SLiM repositories have their own variants of regex notation (Table 3), which may need converting to standard regex patterns prior to analysis with other tools or servers. SLiMSuite tools can make this conversion if required. Some tools have expanded the standard regex patterns, as discussed below.

1.2.2 Profile Representations of SLiMs

Most profile-based methods represent SLiM-like sequence signatures as matrices that are derived from input data containing a set of sequences assumed to carry the motif of choice. These can be user-specified after careful inspection of interesting data or extracted from larger datasets using computational methods. Profiles are typically derived from a frequency table of $20 \times N$ fields (with N being the length of the motif), which is computed from the site-specific amino acid counts and normalized by the number of input sequences and inherent biases in amino acid composition. The latter is usually taken from an empirical background distribution or collected from randomized sequences. Building a profile from a restricted set of known sequences can omit valid occurrences of amino acids at positions where they were not observed. To avoid this it is customary to use "pseudocount" observations, which are added into the frequency table even though they were not actually observed. Since the contribution of pseudocounts is small and continues to diminish as more observed data is added, they will have little relevance to the final profile but are crucial mathematically to avoid the issues of null values in log-odds profile representations. The resulting profile should be an overrepresented sequence signature as observed in the data, from which a putative motif could be extracted [19]. Such a profile can be considered as

a special, limited case of the profile hidden Markov model (pHMM) [20, 21]. The added versatility of pHMM comes from their capacity to not only assign different frequencies to residues but also to allow for insertions and deletions of variable length between sites.

On face value, a profile is superior to a regular expression in many ways because the frequency data allows quantitative scoring of a motif instance. Whereas a regular expression might have [ILV] for a position, a profile could encode the information that 90 % of instances have an isoleucine (I) and only a minority have leucine (L) or valine (V) and weight observations accordingly. The drawback is the requirement for sufficient data to make the profile accurate. SLiM instances are generally few in number and so there is a big danger of over-fitting the profile model, especially given that there are 20 possible states for each position in the motif. Rather than modelling the true constraints of the motif, profiles could simply be representing any early bias in discoveries. For this reason, profiles tend to be popular for DNA motifs (where the number of instances is often high and there are few possible states per position) but are of much more limited use for SLiMs and other alignment-free protein motifs. Exceptions are posttranslational modifications (PTM), such as phosphorylation, which often have many occurrences and recognition motifs based on large screenings of peptide libraries (well-exploited in methods such as Scansite [22]). Where sufficient data exists, profiles can be very powerful because of their ability to quantitatively assess deviations for core SLiM consensus definitions. This can help when identifying previously unseen variants of known motifs and could prove essential to effectively mine large data in the search for novel SLiM instances.

1.2.3 Limitations with Current Motif Definition Schema

The common SLiM formats do have some limitations in the nature of information that they can encode. There is currently no good way to represent interdependencies between sites, for example, where the constraints on one position are determined by the amino acid at another. For profiles, context-sensitive HMM [23] may help to model non-contiguous relationships but are yet to be widely applied in bioinformatics. For regex motifs, some effort has been made in this direction with the 3of5 webserver [24], which recognizes "*n* of *m*" stretches where *n* residues in a window of length *m* are of a given type. This was extended further by PRESTO (Table 1) and its successor in the SLiMSuite package, SLiMProb (formerly SLiMSearch 1.x [25]), to allow more complex either/or stretches in the form "<*r*:*n*:*m*:*b*>", where *r* and *b* can be any single or ambiguous regex elements (Table 3) of which *n* residues in a window of *m* positions must match *r* and the remainder must match *b*. If *b* is a wildcard then a simpler "<*r*:*n*:*m*>" notation can be used, which corresponds to the original "*n* of *m*" pattern element. This notation allows very efficient encoding of complex regex patterns, although these are actually exploded by SLiMProb into

different sub-variants for searching. For more complex scenarios, multiple versions of a motif are defined, such as the Class I and Class II SH3 domain motifs in ELM, [RKY]xxPxxP and PxxPx[KR], respectively [9]. Current SLiM definitions also do not encode the secondary/tertiary structural constraints [26], even if some of the SLiM databases do store and utilize such information for specific motif entries, as described in later sections.

1.3 SLiM Evolution

There are two key principles underlying the evolution of SLiMs in protein sequences: conservation of individual SLiM occurrences (divergent evolution) and independent evolution of SLiM occurrences in unrelated proteins (convergent evolution) [1]. The functional constraints of SLiMs mean that they are subject to purifying selection and will generally show a higher level of conservation than the surrounding residues in disordered regions [27]. The evolutionary plasticity of SLiMs is generally higher than residues that are both functionally and structurally constrained, with single point mutations often sufficient to destroy a motif occurrence or even create a functional SLiM from previously inactive protein sequence. Such plasticity may be harnessed by positive selection to rapidly rewire PPI networks, particularly considering that the low affinity nature of SLiM-mediated DMI probably confer an extra tolerance of SLiM gains and losses in the network [12].

There is no one-size-fits-all solution to SLiM discovery and one must carefully consider the nature of the data before selecting the evolutionary models that should be applied. Where occurrences are likely to be functionally relevant and there is reason to suspect that this function would be found in ancestors, e.g., it encodes a function seen across all mammals/vertebrates, it makes sense to look for signals of evolutionary conservation on a background of divergence. If, on the other hand, a SLiM occurrence is speculated to be new (in evolutionary terms) or even non-physiological (obtained from experiments such as peptide library screens or yeast two-hybrid data) then evolutionary conservation will be misleading at best and counter-productive at worst. The nature and distribution of the SLiM occurrences must be considered before invoking a model of convergent evolution. Phage display is essentially convergent evolution in the laboratory, whilst random peptide libraries have the sequence independence of convergence even if there has been no evolution per se. If, on the other hand, all known occurrences of a SLiM come from the same protein family then conserved function is the most parsimonious explanation and it makes little sense to model convergence. The exception, of course, is where additional evidence points to multiple independent origins of SLiM function.

1.4 SLiMs and Protein Structure

The majority of SLiMs occur in intrinsically disordered regions (IDR) of proteins, at least in their unbound state [1]. The reduced structural constraints of IDR result in reduced evolutionary

constraints and mean that they are generally free to evolve at a faster rate at the sequence level [28], even if they generally conserve their ordered/disordered protein segments [29]. This, in turn, contributes to the previously mentioned evolutionary plasticity of SLiMs, in addition to conferring a degree of structural flexibility on SLiMs that includes potential disorder-to-order transitions linked to protein binding [30]. Indeed, certain SLiMs are known to undergo conformational rearrangements of this type [8, 31], although this is unlikely to be the case for all SLiMs. Following molecular dynamics simulation, Cino et al. have recently proposed that SLiMs tend to adopt conformations typical of their bound state even in the free state [32]. Under this model, "pre-equilibrium" structured SLiM conformations are stabilized later by the interaction, as opposed to an "induced fit" model where binding itself triggers the conformational change. Either way, whilst flanking regions of SLiMs tend to match the composition of IDR [30], the key positions in SLiMs themselves are enriched for hydrophobic and aromatic amino acids more typical of structured regions [1]. Indeed, many SLiMs can be thought of as regions of disorder with a propensity towards order [30]. This flanking disorder may itself be under positive selection to confer protection against peptide aggregation around SLiMs [33]. How much of the enrichment of SLiMs in IDR versus globular regions is due to structural constraints for SLiM-mediated binding, and how much is simply due to the increased evolutionary plasticity of IDR increasing the chance of SLiMs evolving convergently, is yet to be established.

SLiMs include PTM sites, some of which occur on the (structured) surface of globular domains. There are also extracellular SLiMs, which occur in proteins with less intrinsic disorder than intracellular proteins [1]. As a result, approx. 15 % of all known SLiM instances are actually on globular domains. These regions can present extra challenges for SLiM discovery, as they will also contain a number of structural motifs that are constrained in three-dimensional space. Whilst not necessarily linear, many structural motifs will include linear stretches that could be erroneously identified by SLiM predictors. Methods for predicting structural motifs directly do exist (e.g., SiteBinder [34]), but these are not considered in this review.

1.4.1 *Protein Isoforms and SLiMs*

SLiMs are undoubtedly responsible for some of the functional diversity imparted on protein sequences via alternative splicing/promoter use [35–37]. Alternative translation initiation sites can also give rise to different protein products [38, 39] and are likely to similarly alter the SLiM complement of proteins, particularly in terms of N-terminal subcellular targeting motifs. To date, however, most resources for both PPI and SLiM prediction deal predominantly with "canonical" protein sequences and thus protein isoforms are not further considered in this review. All of the

approaches described can potentially be applied to protein isoforms and this flexibility represents one of the benefits of tools that permit analysis of bespoke protein sequences rather than relying on, for example, Uniprot [40] data. It should also be noted that methods such as SLiMFinder [41] that correct for evolutionary relationships within input sequences should also be able to deal with multiple isoforms for each protein, although this has not been formally tested. Note also that GOPHER [42], which is supplied with SLiMSuite, can be used to generate alignments of orthologous splice variants from appropriate source data, such as Ensembl proteomes [43].

1.4.2 SLiMs, MoREs and MoRFs

SLiMs are not the only binding features present in IDR. Regions within IDR that mediate PPI via a disorder-to-order transition upon binding have also been labelled Molecular Recognition Elements (MoREs) [44] (if reasonably short and helical) or Molecular Recognition Features (MoRFs) [45, 46]. There is not a clear delineation between the concepts of SLiMs and MoREs/MoRFs. Some classes of SLiMs probably represent a subset of MoRFs that are short and have specific residues involved in the interaction. Other SLiMs are too short to count as MoRFs (defined as 10–70 aa in length) and/or do not undergo the stipulated structural transition. SLiMs and MoRFs are therefore best considered as complementary and overlapping sets of molecular features. IDR-mediated PPI may include either, both or indeed neither element [6].

1.5 Definition and Databases of Known SLiMs

This review is predominantly concerned with the task of predicting SLiMs from one or more protein sequences. Before examining the primary methods for doing so, it is useful to briefly consider where our current SLiM definitions come from as well as the key databases for storing them. SLiMs are notoriously difficult to define and one must always entertain the notion that definitions found in SLiM databases are incomplete and/or biased by the nature of their discovery. Most known SLiMs were experimentally discovered, although precisely defining the motif often involves bioinformatics, such as a sequence alignment, and manual decisions regarding what comprises the important and/or conserved residues. Often, motifs are simplified to a “canonical” core but also have “non-canonical” instances that deviate from the main definition. This can create some confusion for SLiM rediscovery as it is not always clear what definition(s) of a motif to use. SLiMs are affected by their immediate context, with flanking residues that do not seem to increase affinity directly but are crucial to the specificity of binding [47]. It is therefore highly likely that the flanking sequence could add binding constraints that would render certain residues superfluous. SLiM predictions do not normally tolerate mismatches because the SLiMs themselves already have very low information content and a high probability of occurring by chance.

In situations where non-canonical occurrences are common search tools that incorporate mismatches (e.g., PRESTO) might be required.

When considering the experimental evidence for SLiMs, the nature of the protein–peptide interaction and whether it provides biophysical or biological support is important. In other words, is the experiment providing evidence of what *could* bind or what *does* bind? High-throughput experiments, including screening peptide libraries and similar technologies such as phage display can potentially define binding motifs without any known PPI. This approach can have advantages, in that it can potentially define motifs for "singleton" interactions (e.g., those with only a single occurrence in nature) and can also generate the high numbers of sequence variants required for building profiles, as exemplified for PDZ and SH3 domains [48] and for PTM by Scansite [22]. The high number of variants is also good for identifying amino acids that are not tolerated in particular positions, which otherwise tends to require careful mutation studies. It should be remembered, however, that such SLiMs are not always physiological: peptide-based techniques will be biased towards sequences that have the strongest affinity, whilst SLiMs in nature often have a lower affinity than possible in order to maintain the correct signalling dynamics [1, 8, 15]. This lack of physiological relevance is not necessarily an issue and permits the exploitation of data that might otherwise be ignored. For example, Liu et al. have found evidence from yeast two-hybrid experiments that out-of-frame constructs, which code for short peptides without homology to known proteins and are typically discarded as false positives, may contain novel SLiMs that can be identified computationally [49].

There are now a number of public repositories that are largely or wholly dedicated to collating and curating SLiMs from the literature. These are an excellent source of known motifs and motif instances, which can be used either to interrogate a protein of interest or to assess a potentially new SLiM discovery. An overview of the four main SLiM databases is given below. In addition, a number of targeted motif databases exist for specific classes of SLiM, particularly PTM [50, 51].

1.5.1 PROSITE

PROSITE was one of the earliest collections of linear motif definitions for both SLiMs and longer globular domains [18] although it has largely been superseded by ELM [9] and MnM [10] as a repository for SLiMs. PROSITE motif notation is similar to standard regular expression notation but has some important differences (Table 3). Its regex domain definitions provide a potential source for identifying putative SLiMs that are actually structural motifs or parts of larger regions of homology. For domain searches themselves, it is more usual to use the sequence profiles in PROSITE [52, 53] or HMMs (e.g., SMART [54] and Pfam [55]).

1.5.2 ELM

The Eukaryotic Linear Motif (ELM) database is now over 10 years old [9] and the number of annotated ELMs (as of Jan 2014) has increased to nearly 200 classes and over 2,400 instances in six categories (as denoted by their prefix):

- CLV: Proteolytic cleavage. Sites of posttranslational enzymatic cleavage.
- DOC: Docking. These recruit a modifying enzyme but are not targeted by the active site.
- DEG: Degron. Part of the proteasomal degradation pathway, directing protein polyubiquitination.
- LIG: General ligand binding. Mediating PPI is the primary/sole known function.
- MOD: Posttranslational modification sites, e.g., phosphorylation. (Note that proteolytic cleavage has its own CLV class).
- TRG: Subcellular targeting. Recognized by machinery that directs the parent protein to appropriate cellular localization.

Note that the DOC and DEG categories are recent additions and many studies will have these motifs classed as LIG under the previous classification. The remaining LIG category can best be thought of as SLiMs for which the main, or possibly only, known function is to mediate a PPI. Arguably, all ELMs are protein ligands but it can be useful to consider distinct subsets in case they have different biases in attributes and behavior. Indeed, a recent review using the older four-category classification highlighted some differences between ligands, modifications, and targeting sites [1]. Future releases may extend this classification further.

In addition to the database, ELM hosts a motif search server that includes built-in filters based on evolutionary conservation [56] and structural considerations [57], which are explored in more detail in later sections. Other resources at ELM include the iELM server for exploring SLiM interactions [58], the Phospho.ELM database of experimentally verified phosphorylation sites [59], the switches.ELM "compendium of conditional regulatory interaction interfaces" [13, 60] and a curated set of eukaryotic SLiMs that are the target of molecular mimicry by viral proteins [15]. The ELM conservation scorer is also available at the site to run on user-supplied proteins or alignments. Although MnM has more instances than ELM (even with the 43,000 Phospho.ELM sites), the quality of the curation and availability of the data make ELM the leading SLiM repository.

1.5.3 Minimotif Miner (MnM)

Minimotif Miner (MnM) probably has the largest collection of known SLiMs from the literature, with over 295,000 instances and 880 consensus sequences in MnM 3.0, of which the vast majority are PTM [10]. Some of these are redundant, and so the real

number is likely to be somewhat smaller. Figure 3 of the MnM 3.0 paper, for example, lists both Rx[KR]R and Rx[RK]R as furin proteolysis motifs. MnM is enriched in mammalian motifs but not restricted to eukaryotes by design, with some entries found in bacteria. As with ELM, these are available for searching against an input sequence using an online search tool (*see* next section). Unfortunately, unlike ELM, MnM have not made their SLiM collection available to download and interrogate outside of their webserver, which limits the utility of the service.

1.5.4 Scansite

Many SLiMs are recognition sites for reading, writing or erasing PTMs. Phosphorylation is particularly widespread in signalling systems [61]. Scansite is a leading database for phosphopeptide motifs and the premier profile-based SLiM database and search tool [22]. Scansite3 is its latest version and has profile models for 70 mammalian and 54 yeast protein kinases and phosphopeptide binding domains (e.g., 14–3–3, SH2, SH3, PDZ). The majority of the data in Scansite were generated using "oriented peptide libraries", which fix a central (possibly phosphorylated) serine, threonine or tyrosine residue, and generate random libraries of flanking sequences that are incubated with the domain of interest [62]. Subsequent Edman sequencing of phosphorylated/bound peptides generates the amino acid frequency distribution at each position, which is then converted into a sequence recognition profile. These data are excellent at identifying the optimal (e.g., highest affinity) binding profile for a given phosphopeptide domain. It should be noted, however, that biologically relevant SLiM occurrences are not necessarily optimized for maximum binding affinity [1, 8, 15] and may therefore show a profile different from those generated by a peptide library screen.

1.6 Databases of Predicted SLiMs

There are currently no databases collecting and/or annotating predicted SLiMs but several large-scale SLiM predictions have their data available as supplementary data and/or online (Table 4). Interpreting SLiM predictions is largely a matter of placing them in context. Results from an interactome-wide de novo SLiM prediction in humans [63], for example, have been made available as a series of linked web pages called SLiMdb. This enables predictions to be grouped and studied by SLiM, hub (i.e., proposed PPI partner), parent protein, and GO classification [64]. Entries link out to data in external resources including Ensembl [43], Uniprot [40], GO, OMIM [65], HPRD [66], and Genecards [67]. Further context can be provided by searching the motifs against the human proteome using SLiMSearch2 [68]. Predicted SLiMs have also been compared to each other and/or to databases of known motifs using CompariMotif [69], which is available both as a webserver and a standalone program. This enables clusters of similar motifs to be identified and explored. In future, it is planned to extend

Table 4
Large-scale de novo SLiM Discovery analyses

Method	Data	Source	Species	Data available?	Predictions available?	Reference
FIRE-pro	GO, PPI, sub-cellular localization, half-life	Online databases and curated bibliography	Yeast	Y (with formatted data for other species)	Y	[124]
SLiMFinder	PPI	Online databases	Human	Y	Y	[63]
LMD (DILIMOT)	PPI	Yeast two-hybrid, online databases and curated bibliography	Human, fly, nematode, yeast	N (retrievable from original authors)	Y	[71]
D-STAR	PPI with SH3 domains and in TGFβ signalling pathway	Online databases and curated bibliography	Yeast	N (retrievable from original authors)	Y (partial)	[121]
motif-x	Phosphopeptides	Immunoaffinity and SCX Chromatography	Human	N (retrievable from original authors)	Y (from publication)	[135]
motif-x and scan-x	PPI	Online databases	Human, mouse, fly, yeast	N (retrievable from original authors)	Y	[136]
motif-x	Phosphopeptides	LC/MS-MS	Fly, mouse, yeast	Y	Y	[137–139]
MeMotif	Transmembrane proteins	Online databases	All	N (retrievable from source)	Y	[70]

SLiMdb to improve data querying and include data from other SLiM prediction studies. Another example of a resource in which motif predictions have been made available for interactive exploration is the MeMotif database of consensus linear motifs from alpha-helical transmembrane protein structures [70].

1.7 DNA and Protein Motif Search Tools

Most motif prediction tools developed for DNA or protein sequence motifs can be adapted to the other biopolymer by simply changing the alphabet. There are important differences between DNA and protein sequences, however, and these should not be ignored or overlooked. DNA is simple in comparison, with only four possible base states (ignoring methylation) and DNA sequences analyzed tend to be relatively long, from hundreds to millions of bases. This enables a much more accurate modelling of the background sequence space, even at the di- or tri-nucleotide level; if protein-coding regions are present in the DNA, amino acid and codon usage bias might result in tri-nucleotide sequence biases. DNA motifs also tend to have much higher support in search datasets. In contrast, protein sequences are much shorter (the average unmasked human protein being approx. 500 aa in length) and there are 20 amino acids (excluding PTM), which means that di-amino acid frequencies can rarely be accurately estimated. Protein motifs also tend to have fewer occurrences. As a result, sophisticated methods that can work well for DNA motif prediction are usually either inappropriate or impractical for protein motif applications. For this reason, this review concentrates on tools that have been explicitly designed and/or benchmarked for de novo SLiM discovery. Exceptions include algorithms designed to identify motifs from individual proteins based on alignments of orthologues, for which it is possible to assemble quite large datasets, and the Scansite peptide library approach described above [22, 62].

1.8 SLiM Discovery Benchmarking

A challenging and sometimes overlooked aspect of both the development and appropriate application of SLiM discovery tools is robust benchmarking. New methods require adequate benchmarking data to ascertain their utility and whether they offer an improvement over existing methods. The latter comparison can be particularly hard if the most similar methods have themselves been inadequately benchmarked. The latest releases of SLiMSuite include SLiMBench (unpublished), a tool for generating SLiM discovery benchmarking datasets and assessing performance. SLiMBench and some model benchmark datasets will be made available at the SLiMSuite website. In the meantime, here are some considerations for the benchmarking of SLiM discovery tools:

- *Scale*. Whilst they can be useful exemplars for specific method features, individual observations do not constitute benchmarking and are easily subject to performance bias, whether

deliberate or accidental. Regrettably, a number of the less specialized tools are benchmarked quite well on DNA data but neglect protein applications with a limited number of poorly conceived test datasets. The restricted number of known SLiMs does present a problem and previous methods have been somewhat limited in terms of benchmarking on real data (*see* for example DILIMOT [71] and SLiMFinder [41]) but, at the very least, the ELM database should be used [9]. Simulated data is also useful for getting the numbers up, subject to the considerations below.

- *Bias.* Benchmarking data should ideally be unbiased but, at the very least, its biases must be clear. It is OK to include a benchmark that is biased towards the particular model being tested but this will only tell you whether the algorithm is working computationally, not whether it is useful biologically. Benchmarks should also include data that does not make the same assumptions as the new model being tested. Particular attention should be paid to dataset size and signal:noise ratios in the data as methods generally perform better when these are both large, which is not always realistic.
- *Realism.* Regardless of benchmarking performance on simulated data, there is always the possibility—indeed, likelihood—that real data will have additional biases. Checking performance against real data is therefore crucial. It is often hard to get the same numbers as with simulations and the assessments are often less robust as a result but this cannot really be helped. (Re-benchmarking later is always an option.). The important thing is that benchmarking is not solely on simulated data.
- *Accuracy versus efficiency.* Although computational efficiency is important, accuracy of methods is more important. Whilst a slow method can often be overcome by careful parameter selection and/or finding a faster/bigger computer, rapid results of unknown accuracy are of limited use.
- *False Positive Rates.* Often, methods are only benchmarked in terms of recovery of true positive motifs. A frequent approach is to rank predictions and then demonstrate that the known motif is returned among the top-ranked motifs and/or most of the top-ranked motifs are true positives. The problem with this is that all such test datasets have motifs to be found. In real biological scenarios, it is often not known whether (a) there is a real motif in the data to be found and/or (b) if so, whether there is actually enough signal for said motif to occur more than by chance. For de novo discovery, it is imperative that methods are also benchmarked on datasets that have no real/planted motifs and thus all predictions are false positives. Simulated data is particularly useful for modelling these.

To carry appropriate SLiM predictions forward to laboratory validation, an estimate of the likelihood that a given returned motif is a false positive is essential.

- *Application-focused*. Predictive bioinformatics tools frequently make use of Receiver Operating Characteristic (ROC) curves, particularly for classification problems such as predicting known motifs. ROC curves, which plot true positive rates (TPR, the proportion of positives that are correctly predicted) against false positive rates (FPR, the proportion of negatives that are wrongly predicted) for different thresholds, can be useful exploratory tools but they leave a lot to be desired when it comes to biological benchmarking. Usually, one is operating in a specific part of the ROC space—typically, either minimizing FPR or maximizing TPR. Tools may excel in one area but do poorly in another and this is not captured adequately by "Area Under the (ROC) Curve" (AUC) statistics. When selecting a SLiM discovery tool, it is best to choose one and/or select parameter settings that perform appropriately for the desired application. SLiMFinder, for example, is extremely stringent, making it ideal where false positives are to be avoided [41]. Where false predictions are not an issue, however, the SLiMChance ([17, 41]) significance threshold of SLiMFinder can be relaxed to increase sensitivity.
- *Comparative*. Whenever possible, new methods should be compared to existing methods where they are available (Table 5).

True positives can be identified by comparing predicted SLiMs to databases of known motifs, either manually or using CompariMotif [69]. Although motif comparisons are scored and ranked by CompariMotif, we are not currently aware of a statistical framework for these comparisons nor is there a well-modelled threshold for assigning a match between two motifs. In order to consider a SLiM to be a true positive, SLiMBench uses CompariMotif criteria that matches must: [1] have 2+ positions match; [2] have a normalized information content of at least 1.5 (approx. equivalent to 1 fixed position and one mildly degenerate position, *see* [69]); [3] match at least half of the smallest motif. Although strict application of these criteria will misclassify some motif matches, agreement with our manual classification was good (data not shown) and it has the advantage of being consistent and unsupervised, which is clearly beneficial for comparative benchmarking. For the moment, however, motif matches are largely a matter of individual discretion; caution and discretion should be employed when interpreting claims of "true positive" motif predictions, especially in the absence of data being made available.

Table 5 Computational SLiM Discovery methods

Tool	Description	Ref.	Availability[a]	Prediction[b]	Notation[c]	Input[d]	Scoring[e]
AMS	AutoMotif Server. Machine Learning predictions of PTM	[101]	D	K	O	M	S
ELM	Eukaryotic Linear Motif server. Regex searches of known SLiMs with numerous contextual filters	[9]	W	K	R	S	C D O
iELM	Interactions of Eukaryotic Linear Motif. Predict new instances of known ELM motifs from PPI data	[58]	W	K	R	S M	D O
iSPOT	Infer Sequence Prediction Of Target. Prediction of PDZ, SH3 and WW binding sequences from structural data	[91]	W	K	P	M P	O
MnM	Minimotif Miner. Regex searches of curated literature motifs with numerous contextual filters	[10]	W	K	R	S	C O S
ScanProsite	Perform Regex and Profiles searches of PROSITE patterns or user-defined motifs against user proteins or public databases	[76]	W	K U	R P	S M P	O
Scansite	Profile-based searches of known phophoSLiMs against user sequences. Searches of user-defined regex and profile motifs against public databases	[22]	W	K U	R P	S P	O P
3of5	3of5 regex search tool. Simple protein searches with expanded regex notation	[24]	W	U	R	M	
ANCHOR	Identifies regions with propensity for order within IDR. Can map user-defined regex onto disorder profiles	[125]	W	U D	R	S	D S
FIMO	Find Individual Motif Occurrences (MEME Suite). Search MEME profiles against user proteins or public databases	[73]	W D	U	P	M P	C S
GLAM2SCAN	Scanning with Gapped Motifs (MEME Suite). Search GLAM2 profiles against user proteins or public databases	[74]	W D	U	P	M P	C
MAST	Motif Alignment and Search Tool (MEME Suite). Search with multiple profile motifs in combination for proteins with high combined scores	[75]	W D	U	P	M P	C S

(continued)

Table 5
(continued)

Tool	Description	Ref.	Availability[a]	Prediction[b]	Notation[c]	Input[d]	Scoring[e]
PRESTO	Protein Regular Expression Search Tool (SLiMSuite). Regex search tool of user-defined SLiMs against local protein data with expanded regex notation and tolerance of mismatches	[140]	D	U	R	M	C D O
SLiMProb	Short Linear Motif Probability (SLiMSuite). Formerly SLiMSearch 1.x. Regex search tool of user-defined SLiMs against local protein data with expanded regex notation and numerous contextual masking options. Returns significantly over- and under-representation statistics controlling for homology	[25]	W D	U	R	M	H C D O S
SLiMSearch2	Short Linear Motif Search (SLiMSuite). Proteome screen of regex with contextual filters	[68]	W D	U	R	P	C D O
SLiMScape	Short Linear Motif analysis plugin for Cytoscape (SLiMSuite). Can run SLiMProb or SLiMFinder on proteins selected within Cytoscape	[99]	D	U D	R	M	H C D O S
D-MIST	Domain-Motif Interactions from Structural Topology. Machine Learning predictions of DMI from PDB based on structural context	[131]	D	D	P	P	D O
D-MOTIF	LDMS CMM tool. Identifies correlated motifs in PPI data	[121]	D	D	R	P	
D-STAR	LDMS CMM tool. Identifies correlated motifs in PPI data	[121]	D	D	R	P	
DILIMOT	DIscovery of LInear MOTifs. Formerly LMD. Models convergent evolution/over-representation of TEIRESIAS regex motifs with evolutionary and structural filters	[71, 113]	W	D	R	M	H C D
FIRE-pro	Finding Informative Regulatory Elements in proteins. LDMS CMM tool using Mutual Information to identify motifs that correlate with biological features	[124]	W D	D	R	P	O S

GLAM2	Gapped Local Alignment of Motifs (MEME Suite). Profile-based de novo prediction of over-represented patterns using Gibbs sampling and simulated annealing	[74]	W D	D	P	M	
MEME	Multiple Em for Motif Elicitation (MEME Suite). Profile-based *de novo* prediction of over-represented patterns using expectation maximisation	[114]	W D	D	P	M	
MFSPSSMpred	Masked, Filtered and Smoothed Position-Specific Scoring Matrix-based Predictor. Identifies short regions with propensity for order within IDR based on sequence features and evolutionary conservation	[106]	W D	D	O	S	C D O
MoRFpred	MoRF predictor. Identifies regions with propensity for order within IDR	[128]	W	D	O	S	D
motif-x	Generates fixed position motif from alignment peptides based on over-representation versus background amino acid frequencies	[107]	W	D	R	M P	
MotifCluster	LDMS CMM tool. Identifies correlated motifs in PPI data	[122]	D	D	R	P	O S
MOTIPS	MOTIf analysis Pipeline. De novo profile prediction based on over-representation in short aligned peptides combined with domain-based PPI data	[134]	W D	D	P	M P	C D
NestedMICA	Nested Motif Independent Component Analysis. Identification of enriched motifs versus background reference proteins	[117, 118]	D	D	P	M	
PepSite	Predicts possible DMI from peptides and structural data	[92]	W D	D	R	S	O S
qPMS7	Over-representation of LDMS patterns without correction for homology	[119, 120]	W D	D	R	M	
Pratt	Over-represented regex motif prediction without correction for homology	[110]	W	D	R	M	

(continued)

Table 5
(continued)

Tool	Description	Ref.	Availability[a]	Prediction[b]	Notation[c]	Input[d]	Scoring[e]
QSLiMFinder	Query SLiMFinder (SLiMSuite). Query-based de novo regex SLiM prediction modelling convergent evolution with correction for homology, numerous masking options and statistical support	[41]	W D	D	R	M	H C D O S
SLIDER	LDMS CMM tool. Identifies correlated motifs in PPI data by mapping motifs onto PPI interfaces using structural data	[123]	D	D	R	P	D O
SLiMDisc	Short Linear Motif Discovery (SLiMSuite). Regex de novo SLiM prediction modelling convergent evolution with correction for homology and numerous masking options	[42, 77]	W D	D	R	M	H C D O
SLiMFinder	Short Linear Motif Finder (SLiMSuite). Regex de novo SLiM prediction modelling convergent evolution with correction for homology, numerous masking options and statistical support	[41, 100]	W D	D	R	M	H C D O S
SLiMMaker	Short Linear Motif Maker (SLiMSuite). Simple regex consensus generator from aligned peptide sequences	[140]	W D	D	R	M	
SLiMPred	Short Linear Motif Predictor (SLiMSuite). Artificial Neural Network predictor of SLiMs from sequence features	[129]	W	D	O	S	C D
SLiMPrints	Short Linear Motif fingerprints (SLiMSuite). Prediction of SLiM conservation fingerprints using statistical modelling of RLC	[97]	W	D	R	S	C D S
TEIRESIAS	Simple but efficient text pattern search tool	[112]	D	D	R	M	

[a]Availability of software: *W* Webserver, *D* Download
[b]Prediction type: *K* Known motifs, *U* User-defined, *D* De novo
[c]SLiM notation used: *R* Regex, *P* Profile, *O* Other
[d]Input data for searching: *S* Single Protein, *M* Multiple Proteins (methods accepting multiple proteins can usually be scaled for single proteins or proteomes), *P* Proteome/Database
[e]Scoring and filtering options: *H* Homology correction, *C* SLiM conservation, *D* Structure/disorder filtering/masking, *O* Other filters/scores, *S* Significance estimate

2 Computational Prediction of Known SLiMs

Computational techniques for profile and regex searches are well established and thus finding predefined patterns in sequences is a computationally trivial exercise. Due to their short length, SLiMs are very likely to be found in proteins of typical sizes; the difficult job is distinguishing genuine functional instances of a SLiM from random background occurrences. Choosing the right search tool for known SLiMs is therefore largely governed by what is known about the motifs in question.

Many of the repositories of known motifs also include tools for searching those motifs against a given protein, including the Eukaryotic Linear Motif (ELM) resource [9], Minimotif Miner (MnM) [10] and Scansite [22]. There also exist a number of bespoke tools for searching user-defined motifs against protein datasets (Tables 1 and 5) and no doubt more will be added in the future. ELM and MnM have all-or-nothing matches based on their regular expressions, which are then rated and/or filtered according to contextual information to help the user discriminate true positives from false positives. SLiMProb [25] and SLiMSearch2 [68] allow similar searches for sets of proteins and whole proteomes, respectively, using user-defined regular expressions. Both also provide contextual scoring/ranking options and output that permits users to visually explore predicted instances.

Scansite [22] harnesses the power of probabilistic profile models of known cell signalling interaction motifs to predict new instances in user-defined sequences or various public protein databases. The Scansite3 server can also make SLiM predictions with user-specified matrices of binding affinities per site, enabling users to easily search with their own profiles. Additional flexibility in motif definition is introduced by allowing the specification of an approximate consensus sequence of the motif, which is then used to automatically construct a matrix with similar characteristics. Scansite3 also features an option for searching user-defined peptides or regular expressions against a selection of protein databases. MEME Suite [72] offers a set of scanning tools to allow searching sequence databases with the profile motifs, such as those identified de novo by other tools in the suite. Ungapped motifs found by MEME can be used as input for FIMO [73] to find all motif occurrences in a public protein database, ranked by significance according to their Benjamini–Hochberg corrected *p*-values. GLAM2SCAN offers the same functionality for input gapped alignments provided by GLAM2 [74]. A slightly different approach is taken by MAST [75] as it considers the full set of input motifs as a whole. It first determines the best scores for all matches between pairs of motifs and proteins in the database and then combines these into overall scores between the complete set of motifs and

each protein. The *E*-values calculated by MAST are used to filter out random hits (with a user-defined threshold) and rank the remaining significant proteins. Since MAST results provide a single score for each protein in the database, and information from multiple motifs can be provided as input, the program could be useful to retrieve proteins where different motifs co-occur.

PROSITE patterns can be searched online using Scanprosite [76]. Scanprosite allows proteins to be scanned for PROSITE patterns, or the user can define patterns to be searched against public sequence databases or user-defined protein datasets. Because it does not have any of the filtering tools advised for SLiM discovery (discussed below), Scanprosite is not recommended for SLiM prediction. Users should also note that the Scanprosite default is to "exclude motifs with a high probability of occurrence from the scan," which includes many of the SLiMs in the database. SLiMProb can perform local searches using PROSITE patterns in place of standard regex notation. Likewise, CompariMotif [69] can be used to compare regex motifs with PROSITE patterns.

In general, there are two assessments of SLiM predictions that a user wants to perform: assessing individual predicted occurrences, or assessing enrichment of a dataset for predicted occurrences. These are explored in more detail below.

2.1 Assessing and Ranking Individual SLiM Occurrences

SLiM discovery methods are notorious for over-prediction. To combat this, there are a number of possible considerations that can be very useful for filtering and/or ranking SLiM occurrences. Nevertheless, users should always be mindful that bioinformatics predictions almost invariably need additional validation to be sure of function, even where an estimated confidence in the predictions is returned. False positives can occur purely by chance or they might have a different function from that being sought. Variations of dibasic [KR][KR] motifs, for example, form the core of five different cleavage motifs in ELM as well as several targeting motifs [9]. These latter false positives are particularly hard to identify because they are probably under very similar structural and evolutionary constraints to the motif of interest.

There are essentially three strategies that can be applied to predicted SLiM occurrences. Firstly, contextual data can be simply provided to users, allowing them to weigh different lines of supporting evidence using specialist knowledge and human judgement. Secondly, features can be scored/weighted to produce a final metric for each occurrence, by which they can be ranked. Lastly, scores and/or context features can be used to reject certain occurrences outright. Where sequence/structure context is used, filters are often applied to the input data prior to the motif search, which is more efficient. These are clearly not mutually exclusive approaches and tools will often filter the weakest predictions before ranking and/or reporting context for the remainder. Filtering itself is not

an all-or-nothing affair and should be set to an appropriate level for downstream analysis and the relative tolerance of false positives versus false negatives. Scansite [22], MnM [10] and SLiMSearch2 [68], for example, have different stringency settings depending on how strictly the user wants to filter occurrences.

2.1.1 Sequence Space Considerations

The number of motif predictions returned by any algorithm will rely heavily on the sequence space searched, with longer/more proteins likely to return more hits to the motif regex/profile by chance. On the other hand, real instances will be missed if the sequences containing them are missing or excluded from the search dataset. Indeed, the selection of protein sequences is just as important as the choice of the search algorithm for most applications. This also applies to analyses of individual proteins. SLiMs have been implicated in functional differences associated with splice variation [35–37] and so limiting analysis to canonical sequences could miss potential occurrences. Similarly, SNPs could create or destroy SLiM instances and should be considered where relevant.

When trying to identify and/or rate predicted novel instances of known motifs in a limited number of specific proteins of interest, it is usually safer to err on the side of caution when masking sequence data and/or filtering instances. Because such predictions are tempered by circumstantial data that is not inherently part of the SLiM definition, it is generally a good idea to maximize the sequence space. When searching larger datasets, it is more normal to apply more stringent filters and restrict the number of returned results to a manageable number. If this can be achieved by masking the input data prior to analysis then efficiency can also be improved, which is particularly useful if additional data (sequence alignments, etc.) are created or analyzed for each instance. When looking for enrichment of SLiMs within a dataset, protein sequence selection and sequence space masking are even more important as they will affect any statistical assessment of abundance.

2.1.2 Protein Structure

The majority of SLiMs are found in disordered regions of proteins, at least in the unbound form [1]. Therefore, it frequently makes sense to screen out globular regions prior to motif prediction [41, 63, 71, 77]; although some true positive instances are likely to be erroneously discarded, the hope is that a much higher proportion of false positives will be removed and thus the resulting predictions will be enriched for real SLiMs. Typically, a disorder prediction program (reviewed in [78, 79]) is used to identify and screen out predicted globular regions (e.g., [41, 63]). IUPred [80] is particularly popular for disorder prediction in SLiM discovery because it combines reasonable accuracy with being freely available for academic use. No disorder predictor is completely accurate, so it is generally recommended to err on the side of over-prediction when masking based on disorder. Whereas the default

IUPred disorder cutoff is a score of ≥0.5, for example, cutoffs of ≥0.2 [41, 63] or ≥0.3 [81] are typically used due to the observation that 80–90 % of known ELM occurrences would be retained by such thresholds [1] [80]. We have previously found that the default IUPred disorder cutoff of ≥0.5 correctly classified approx. 95 % of ordered residues in the DisProt database [82] but only approx. 50 % of disordered residues (i.e., it is conservative), whereas a cutoff of ≥0.2 correctly classified approx. 95 % of disordered residues but only approx. 50 % of ordered residues (i.e., is very relaxed) (data not shown). It should be noted that this analysis was performed on a very limited dataset, although similar figures are given for IUPred defaults on CASP data [78]. No systematic analysis of the optimum disorder prediction for SLiM discovery has yet been executed although our own testing of IUPred has indicated that a conservative cutoff of 0.2 gives the best trade-off between specificity and sensitivity for predicting occurrences of known ELMs (data not shown).

An alternative strategy is to mask out domains identified by a domain database, such as Pfam [55]. Whilst this can be effective and was employed by Neduva et al. in their landmark SLiM discovery paper [71], it must be done with caution. Not all domains in Pfam are completely globular and a small but significant proportion are completely disordered [83–85], an observation that is supported by the suggestion that 40 % of the domain folds in the consensus domain dictionary (CDD) [86] are unstable, rather disordered and should not be considered traditional domains [87]. Thus, carefree masking of domains could result in removal of some genuine SLiM-containing regions. As the notion of disordered protein domains as biologically functional and important regions continues to gain widespread acceptance, it is possible that more such domains could end up in domain databases. Combining domain prediction with disorder prediction and/or cross-referencing to a database of disorder domain sequences (e.g., [83]) should help to avoid such errors. ELM, for example, uses a structural filter that combines solvent accessibility and secondary structure [57] to complement disorder predictions by GlobPlot [88] and IUPred [80] and domain predictions from SMART [54] and Pfam [55].

Structural information about PPI is scarce but since it offers direct evidence, it is a valuable resource for SLiM prediction. The 3did database [89] has 462 DMI of known 3D structure (as of January 2014) and rules derived from such data have the potential for predicting new instances and even entirely new classes of DMI [90]. Although biological relevance cannot be established from structure alone it is particularly useful to define the interaction interface with high confidence. An early example of this is iSPOT [91], which uses structural data to estimate the propensity of an input sequence to bind PDZ, SH3, or WW domains. It stores

frequency tables of residue–residue contact pairs calculated from PDB structures of peptide-domain complexes and uses these to score the interaction with each fragment of a defined length in the input sequence. More ambitious is PepSite [92], which models preferred peptide-binding environments for protein surfaces to assess whether a given peptide could bind a given globular domain. This, in principle, could be used to help assess whether a novel instance of a SLiM is able to bind the appropriate domain.

2.1.3 SLiM Conservation

Evolutionary conservation is good for ruling out motif instances that are unlikely to have functional importance; however, there is a major complication to using evolutionary conservation as a discriminator for SLiM discovery. The plasticity of SLiMs and their propensity to occur in IDR means that they are frequently not conserved to the same evolutionary depth as globular domains [1, 12]. Even when the SLiM is conserved, the low conservation of surrounding disordered residues can generate alignment errors [93]. High variability in evolutionary dynamics between different SLiMs and IDR further limits the use of absolute measures of sequence conservation, which are heavily dependent on sequence quality and availability as well as the background evolutionary rate (i.e., functional constraint) of the parent protein. As a result, conservation metrics trained on discovering globular domains tend to overlook SLiMs; SLiM discovery requires its own conservation metrics.

MnM uses a conservation score based on BLAST pairwise alignment scores of HomoloGene clusters [94] and introduced the idea of adjusting conservation scores of predicted SLiM occurrences based on the overall conservation of the full-length proteins [95]. Dinkel and Sticht extended this idea, using weighted percentage identities that were normalized to the global percentage identity of the parent proteins [96]. These adjusted scores were then calculated across increasing numbers of homologues (sorted by relatedness) and the final distribution of scores used to rank predictions. Chica et al. took a different approach and normalized conservation scores by weighting conserved occurrences according to evolutionary distance and then normalizing to the overall tree weighting for each parent protein to allow comparisons between proteins and alignments [56].

These methods still suffer from different proteins (and protein regions) having different distributions of homologues available and/or different functional constraints across the full-length protein, independent of SLiM constraints. The solution, introduced by Davey et al. [27], is to measure the conservation of SLiMs *relative* to a surrounding window of (disordered) residues. This "Relative Local Conservation" (RLC) approach successfully adjusts for both homologue number/distance and alignment quality; if alignment of the SLiM-containing region is poor compared to the

protein as a whole, this should not affect the score. Likewise, RLC effectively normalizes homologue numbers and evolutionary distances. Because individual alignment column scores are used for the RLC normalization, methods that weight conservation according to evolutionary distance can also be incorporated into the method [27]. Furthermore, it is possible to use the distribution of RLC scores to assign a statistical probability to observing a given cluster of high RLC scores at a motif instance, which can be used for ranking predicted occurrences [68] and even directly predicting de novo SLiMs [97].

Notwithstanding the fact that alignment errors will disrupt conservation patterns, the possibility remains that genuinely functional SLiM instances have evolved recently and therefore show little conservation. Likewise, apparent conservation of a given SLiM instance may be a chance occurrence or the consequence of evolutionary constraint on a similar/nearby sequence pattern wholly independently of the SLiM of interest; SLiMs often co-occur [13, 60] and flanking residues can also show correlated evolutionary patterns [98]. Despite these limitations, evolutionary conservation can be a powerful tool when harnessed correctly. ELM [9] incorporates a conservation filter based on the tree-weighting method of Chica et al. [56], whilst SLiMSearch2 [68] uses RLC [27] to help rank and filter results. SLiMProb [25] can mask input data using a number of conservation schemes including RLC.

2.1.4 Use of Other Contextual Information

Where there is sufficient annotation, additional contextual information can be used to rank or filter results. This is exemplified by MnM 3.0 [10], which employs a number of filters that compare predictions to the known target of the motif using a tightly controlled semantic syntax framework [14]. This allows biological data such as gene ontology (GO) and protein/genetic interactions to be combined with homology and structural data to screen out false positives. At the highest stringency, the authors report 39 % retention of validated instances with zero false positives returned. Whether this holds true for all SLiMs is yet to be seen but it demonstrates the importance of contextual information when ranking or scoring motif predictions.

PPI networks are an obvious source of contextual information to help support or reject SLiM predictions. For motifs with known binding partners/domains, these data can be used directly to assess occurrences. iELM [58] predicts new instances of known motifs by mapping instances together with known binding domains onto PPI networks, which can be supplied by the user. There are also tools for interactively exploring SLiMs in the context of PPI networks: SLiMScape [99] is a Cytoscape plug-in that can directly run SLiMSearch predictions of known motifs [68] or SLiMFinder de novo SLiM prediction [100]. For novel motifs, enrichment in such datasets can be indicative of function. This is explored in the next section.

Features of known SLiMs can also be used to build predictors of novel instances using machine learning algorithms. The AutoMotif Service (AMS) [101], for example, trains artificial neural network pattern classifiers for the automatic prediction of PTM sites. Annotated (positive) instances are taken from UniProt and Phospho.ELM, while negative training data is randomly chosen from fragments of sequences with no known PTMs. The disadvantage of machine learning approaches is their tendency to be a "black box", making human understanding and assessment of individual predictions quite difficult.

2.2 Assessing SLiM Occurrences at the Dataset Level

Assessing SLiM occurrences at the dataset level is largely performed for one of two reasons: exploring dataset function through known motifs, or exploring possible motif function through motif distributions. The latter is frequently used to add weight to the de novo SLiM predictions discussed in Subheading 3. In practical terms, the main difference is how many different protein datasets and SLiMs are considered: either a single dataset of interest is searched with one or more SLiMs, or a single SLiM is assessed in multiple datasets that subsequently need to be sorted and ranked (and controlled for multiple testing). The latter exercise is usually the remit of specific tools, such as SLiMSearch2 [68], which will identify GO categories and IntAct PPI partners [102] enriched for a specific SLiM from whole proteome occurrence data. For the purpose of this review, we focus on the general case of assessing a single SLiM in a single dataset. Where multiple SLiMs and/or datasets are used, an additional multiple testing correction will be required.

2.2.1 Overrepresentation Statistics

The most common analysis is to assess a motif for overrepresentation in a particular dataset. The most frequently used statistical approaches for such assessment are the cumulative binomial distribution and the cumulative hypergeometric distribution. In each case, we are interested in the probability of observing k or more successes (motif occurrences) given n trials (positions/sequences in which a motif could occur), each with a probability of success p. The main difference between the two approaches is that the binomial distribution calculates a probability based on n trials *with replacement*, which means that each motif occurrence is deemed to be independent and drawn from an infinite population. This is the norm for SLiM prediction tools. The hypergeometric distribution, in contrast, models n trials *without replacement* given a finite population N. This would be more appropriate for situations in which the total number of motif occurrences in a full dataset was known and enrichment was being assessed for a sub-sample of that dataset, e.g., a single PPI dataset or GO term in a whole proteome. Where N is much larger than n, the binomial is a good approximation and more efficient to calculate.

There are essentially two levels at which motif occurrence probabilities can be considered. At the sequence level, k is the total number of observed occurrences, n is the number of discrete positions at which a SLiM could occur and p is the probability that a motif occurs at any given position. At the dataset level, k is the number of different proteins containing the SLiM, n is the total number of proteins and p is the probability of the SLiM occurring in each protein. In each case, k is normally quite obvious but n and p can vary in the way that they are calculated.

The biggest challenge is correctly modelling the background frequency distribution from which to derive the probability of occurrence per site/protein, p. Due to differences in amino acid composition, one has to consider whether to base expectations on protein-specific or dataset-specific amino acid frequencies. If residues have been masked based on evolutionary or structural information as previously described, there can be big differences between the masked and unmasked data. Alternatively, frequencies might be derived from a different "background" dataset, such as a complete proteome. At one extreme, if sequence biases are not correctly handled then returned motifs will simply represent this bias. SLiMs, for example, tend to occur in disordered regions [1] and so it is common to mask out globular domains and/or predicted ordered regions [41, 63, 71]. Disordered sequences are known to have a different amino acid composition to globular domains [1, 80] and thus if full-length protein sequences are used for the background expectation, there will be a tendency to overpredict motifs because all disordered amino acids are enriched. At the other extreme, if one masks the sequences perfectly so that *only* motif sequences are unmasked, the observed amino acid frequencies will be biased *because of* the motif occurrences, which might make the motifs themselves appear uninteresting. This is especially true for low complexity motifs, which predominantly feature the same amino acid(s) in multiple positions. In general, the sequence space is considerably larger than the motif instances and so the assumption is that the motif does not bias the background.

Once the probability of a given motif occurring at any given position can be calculated from a background frequency, the number of such positions must be taken into consideration. For sequence data, the probability of motif occurrences is clearly related to the size of the sequence space being searched, which in turn determines n. It is here that conservation and disorder masking make a big difference by reducing the number of sites at which a motif can occur by chance. To a first approximation, n is equivalent to the number of unmasked amino acids in the protein.

For datasets of multiple proteins, the probability of occurrence in each protein can be calculated as just described. Alternatively, amino acid frequencies can be bypassed by empirically estimating per-protein probabilities based on occurrences in a background dataset. This approach must be used with caution; there may be

biases in protein composition in addition to any amino acid composition bias. The main problem is the presence of homology, which makes motifs more likely to have extreme distributions than would be expected if all proteins were unrelated, making p hard to estimate. Problems with homology also apply to the dataset of interest, for which the statistical model assumes that the n proteins are independent. Evolutionary relationships between proteins can heavily skew statistics by breaking this assumption of independence, regardless of how p is calculated. To counter this, SLiMProb uses the SLiMChance probability model of SLiMFinder [41], which in turn uses the "Unrelated Protein Clusters" (UPC) correction for evolutionary relationships introduced by SLiMDisc [77]. Under this model, BLAST [103] is used to identify homologous proteins, which are then clustered such that no protein in an UPC has detectable homology with a protein in another UPC. Dataset size (n), motif support (k) and the probability of SLiM occurrence (p) are then calculated using the UPC rather than individual proteins. The importance of this correction cannot be overstated: statistical models that ignore sequence homology cannot be trusted unless the input has similarly been purged of homology. SLiMProb also calculates enrichment for proteins without evolutionary filtering (i.e., assuming evolutionary independence) and for the overall number of occurrences across all sequences (i.e., k is all occurrences, n is the entire sequence space and p is the probability per site), which enables the effect of the correction to be examined.

2.2.2 Underrepresentation Statistics

Although it is less common, it can also be interesting to assess whether a SLiM is *underrepresented* in a given dataset. The statistics for this are essentially the same, except that the main concern is the probability of seeing *k or fewer* occurrences given n and p. It has been observed that false positive occurrences of some motifs are underrepresented in certain datasets [104]. Combining overrepresentation and underrepresentation statistics could therefore prove an interesting way to explore the evolutionary and, by proxy, functional dynamics of a SLiM within a proteome. In particular, one would expect an overrepresentation of conserved instances where the SLiM is functionally important but an underrepresentation of non-conserved instances where they might disrupt signalling and are therefore subject to negative selection. This has clear implications for using overrepresentation in de novo SLiM prediction and is discussed in more detail below.

3 Computational De Novo SLiM Prediction

Many of the considerations for predicting instances of known motifs also apply to the task of predicting SLiMs de novo. Understanding critical features of known SLiMs has allowed the

establishment of a set of rules helpful to find new motifs. The task is clearly more complex when the nature of the motif is not known. It is not simply a question of where functional instances of the SLiM might occur; selecting the right tool for the job depends on the data available, the nature of the motif (length, conservation, induced structure (if any), whether specific amino acid side chains/PTM will be involved, etc.) and the level of confidence that a SLiM is present in the data at all. Where the latter is unknown, estimation of the significance of results is essential. This section will explore these issues and how to make the best use of the available data. Recommendations will be made where possible (*see* also Table 5) but it should be noted that de novo SLiM discovery is still a developing field and there may not (yet) be an obvious "best" approach in all situations.

It is important to remember that de novo SLiM discovery will rarely return the SLiM precisely as it would be defined by in-depth study. This is because the number of instances in nature is frequently insufficient to have fully explored sequence space through evolutionary time. With the exception of very specific motifs, such as the integrin-binding RGD motif, it would be impossible to return the complete motif definition given the sequence data available (Table 6). Even then, it is possible that additional subtle features of the SLiM are yet to be discovered. Because SLiM discovery tools are mining the strongest signals, they will generally return a simpler version of the motifs and/or include some extraneous flanking residues. One should remember this when analyzing the results of any SLiM prediction: the real SLiM (if there is one!) is likely to be a somewhat refined version of the pattern returned by the program. It is also important to remember that not all occurrences in the input data are necessarily going to be functional. Depending on the planned follow up, it might be necessary to rank and/or filter those occurrences using the same techniques as previously discussed for predicting known motifs. Likewise, the data used for SLiM prediction might not include all of the functional instances; it can be useful to perform a large-scale analysis of the distribution of the predicted SLiM, both to identify additional instances and provide insight into whether the motif is genuinely associated with the dataset from which it was predicted.

Tools for de novo motif prediction from sequence data can be broadly classified depending on their goal:

1. Alignment-based algorithms aim to best describe a single motif based on an alignment of motif occurrences.
2. Alignment-free methods aim to interrogate multiple sequences to identify a new common feature.

Alignment-based methods clearly need to use additional information when compared to alignment-free methods in order to

Table 6
SLiMMaker consensus motifs from annotated ELM instances for top 20 ELMs ranked by instances in ELM

ELM	ELM regex definition	Refined SLiMMaker regex[a]	N[b]
LIG_WRPW_1	[WFY]RP[WFY].{0,7}$	[WY]RP[WY]	93/95
LIG_EH_1	.NPF.	NPF	88/88
LIG_AP2alpha_2	DP[FW]	DP[FW]	54/54
LIG_PDZ_Class_1	…[ST].[ACVILF]$	[ST].[LV]$	41/48
MOD_NMyristoyl	^M{0,1}(G)[^EDRKHPFYW].. [STAGCN][^P]	^MG[AGNQS]..[AGS]	38/48
MOD_SUMO	[VILMAFP](K).E	[FILV]K.E	43/45
CLV_C14_Caspase3-7	[DSTE][^P][^DEWHFYC] D[GSAN]	[DST].[LPTV]D[AGS]	25/39
LIG_SUMO_SBM_1	[ILV](.[ILV]|[ILV]|[ILV].)[ILV] [STDE]{1,10}	[ILV][ILV][DIL][DLS] [DST]	27/39
LIG_CtBP_PxDLS_1	(P[LVIPME][DENS][LM] [VASTRG])|(G[LVIPME] [DENS][LM][VASTRG]((K)| (.[KR])))	P[ILM][DN]L[RS]	19/32
LIG_Rb_LxCxE_1	[LI].C.[DE]	L.C.[DE]	31/32
LIG_WW_1	PP.Y	PP[AEP]Y	21/28
MOD_PKA_2	.R.([ST])[^P]..	R.S	27/28
TRG_PEX_1	W…[FY]	W..[DEQ][FY]	23/27
TRG_NLS_MonoExtN_4	(([PKR].{0,1}[^DE])|([PKR])) ((K[RK])|(RK))(([^DE] [KR])|([KR][^DE]))[^DE]	[KPR].[KR].[KR]	18/26
LIG_PTAP_UEV_1	.P[TS]AP.	P[ST]AP[LPQS]	20/25
MOD_PKA_1	[RK][RK].([ST])[^P]..	[KR]R.[ST]	23/25
DEG_SCF_TIR1_1	.[VLIA][VLI]GWPP[VLI]…R.	QIVGWPPVRSYRK	3/24
LIG_NRBOX	[^P]L[^P][^P]LL[^P]	L..LL	24/24
MOD_CMANNOS	(W)..W	W[GS][EPS]W	12/24
MOD_LATS_1	H.[KR]..([ST])[^P]	H.R..[ST]	21/23

[a]Product of iterative SLiMMaker regex construction from annotated ELM instances with default settings: each variant in an ambiguous position must be present in at least 3 sequences; max 5 variants per ambiguous position; during iterations, 75 % sequences must match position to be non-wildcard
[b]The number of annotated ELM instances matching the refined SLiMMaker regex

constrain the motif search. Such methods are more restricted in terms of potential applications but can use approaches that are unsuitable for less constrained alignment-free data.

3.1 Alignment-Based (Divergent Evolution) Methods

Building on the success of protein domain prediction/definition, some de novo SLiM methods attempt to identify SLiMs on the basis of signals of evolutionary conservation among homologous protein regions, i.e., purifying selection acting at functionally important sites during divergent evolution. The challenge is that globular domains dominate this signal and so SLiM-specific models of evolution must be applied. Recently, methods have harnessed the power of the "Relative Local Conservation" (RLC) method discussed in Subheading 2 [27]. SLiMPrints uses a statistical model of clustered RLC-conserved residues to identify evolutionary signatures of SLiM occurrences and return them as regex patterns of fixed and wildcard positions [97]. A similar approach has also been taken using phylogenetic hidden Markov models (phylo-HMM) to search for locally conserved sequences in unstructured regions [105]. The nature of profile-based methods make them particularly suitable to capture the evolutionary constraints of homologous sequences; another example is MFSPSSMpred [106], which incorporates local conservation scores from multiple sequence alignments into a support-vector machine model of MoRF sequence features to predict novel SLiMs/MoRFs. These methods have the advantage of being able to identify singletons (i.e., motifs with a single known instance) but additional data and/or experiments will be required to predict the function of any SLiMs that are discovered.

Alignments can be based on function rather than homology. Motif-x, for example, is an alignment-based method that is actually modelling convergent evolution by aligning otherwise unrelated sequences around key residues such as phosphorylation sites recognized from mass spectrometry data [107]. Fixed position motifs are constructed from a window either side of the aligned residue. SLiMMaker will similarly generate a consensus regex SLiM (with ambiguity) from a set of aligned peptide sequences, whether they are homologous or not (Table 6).

3.2 Alignment-Free (Convergent Evolution) Methods

One of the most common and successful approaches for de novo SLiM prediction is the interrogation of multiple different proteins for shared sequence patterns. Unlike most alignment-based approaches above, these methods are modelling *convergent* evolution, i.e., the independent origin of shared motifs on unrelated protein backgrounds. There are a number of potential sources for such protein sequences [11] but the most common are PPI data [108] and functional classifications such as GO. In each case, methods are generally seeking either (a) the most abundant patterns in the data, or (b) the most enriched patterns versus a background expectation.

The latter are generally more effective due to inherent biases in amino acid frequencies but they are reliant on good background models to calculate the expected motif abundance by chance.

One of the earliest dedicated de novo SLiM discovery tools was Pratt [109, 110], which updated and extended an earlier approach by Neuwald and Green [111]. Although Pratt is an alignment-free method, it was originally designed with divergent sequence motifs in mind, such as PROSITE family descriptors [18], which is reflected in the parameters. Pratt is still useful for returning a ranked list of motifs that include amino acid and wildcard-length degeneracy. It is designed to find patterns that occur in the majority of sequences and does so efficiently. The algorithm can be very slow when searching for patterns present in only a few sequences, particularly when the motifs are small. Output is highly dependent on parameter settings and it does not return a statistical significance for predicted SLiMs. Furthermore, because it was designed with sets of homologous proteins in mind, there is no evolutionary filter to model convergent evolution. As a consequence, although still available at EBI, Pratt is not recommended for general de novo SLiM discovery. A better alternative is SLiMFinder [41] (below), which returns Pratt-like flexible patterns but is optimized for convergent evolution and also provides an estimate of statistical significance for predictions.

Another notable early tool is TEIRESIAS [112], a general text pattern finding tool that has been widely applied to the problem of de novo SLiM discovery and served as the inspiration or basis for several tools that followed. TEIRESIAS can return "degenerate" motifs with site-specific variability through equivalence sets of residues, i.e., sets of amino acids that can co-occur in ambiguous positions. Any user-defined equivalence sets may be used but it is most common to group amino acids that share physicochemical properties. Given a set of protein sequences, a scanning phase takes place to collect all putative motifs of given length, proportion of defined sites and support in the dataset. These are then combined recursively into longer patterns with enough support, keeping the efficiency of the algorithm by discarding patterns that are less specific versions of others, while accounting for ambiguity by treating all residues in the same set as equals. Homologous sequences in the input dataset can inflate support for certain motifs, which can also result in very long run times and massively increase the number of patterns returned. Nevertheless, it is still sometimes used for baseline performance comparisons in methods benchmarking.

3.2.1 Methods Correcting for Evolutionary Relationships

One weakness of the early methods is their failure to consider evolutionary relationships in the data. This is important, otherwise large regions of homology will be returned as motifs in a potentially misleading fashion. The first method to correct for this was DILIMOT [71, 113], which filtered homologous regions and kept

a single representative for analysis. Whilst this kept the underlying TEIRESIAS pattern discovery step efficient, it has the slight disadvantage of removing sequence variants that might better reflect the core SLiM. Another concern is that weakly homologous regions flanking those removed by the filter might remain in the data and bias results. Nevertheless, DILIMOT was a major advance in de novo SLiM discovery and its application to human, fly, nematode and yeast PPI data was a landmark paper in the field.

SLiMDisc (Short Linear Motif Discovery) [42, 77] was developed around the same time as DILIMOT but took a different approach to correcting for evolutionary relationships. Instead of filtering homologous sequences, motifs were given a heuristic score that was based on their homology-corrected support and information content (i.e., length and degeneracy). Three different correction methods were tested. The best performance was achieved by scaling motif support using a "Minimum Spanning Tree" (MST), which would produce a corrected support from 1 to N, where N is the number of proteins in which the motif is found. If all N proteins were identical, MST would scale support to equal 1. If all N were unrelated, support would be N. Like DILIMOT, SLiMDisc used TEIRESIAS for underlying pattern discovery and was essentially an add-on for filtering and ranking TEIRESIAS output. The original SLiMDisc scoring was subsequently modified in the webserver implementation using "SLiM Pickings," which weighted the original SLiMDisc score according to the ratio of observed versus expected support, corrected for evolutionary relationships and amino acid frequencies of the input data. Later releases of the webserver have seen TEIRESIAS pattern finding replaced by the SLiMBuild algorithm of SLiMFinder [41, 100].

There are two main drawbacks of the SLiMDisc/DILIMOT approach. Firstly, scores are not directly comparable between datasets. Secondly, whilst the methods are very good at returning real motifs among the top-ranked patterns in the output, there is no way of assessing how likely it is that a given data had *any* genuinely overrepresented motifs. SLiMFinder (Short Linear Motif Finder) [41, 100] overcomes these two problems by carefully controlling the motif space during motif construction using its own SLiMBuild algorithm in place of TEIRESIAS. This motif space is then used by the SLiMChance algorithm to robustly, if somewhat stringently, estimate the significance of overrepresented motifs. SLiMChance uses the binomial distribution as described for SLiMProb in Subheading 2 with an additional multiple testing correction for motif space. This again uses the cumulative binomial distribution, where k is 1 (a single successful motif), n is the total number of motifs in the motif space, and p is the individual motif's overrepresentation probability. These solutions enabled large-scale analysis of tens of thousands of human protein datasets [63]. SLiMFinder also introduced the capability to return motifs with flexible-length runs of wildcard positions, which are important for some motifs.

It should be noted that correcting for evolutionary relationships is not always possible. Sometimes, numerous overrepresented motifs will be returned from the same (sub)set of input proteins even if there is no BLAST-detectable homology; BLAST can sometimes miss homology in short and/or low complexity proteins. For this reason, it is always advisable to manually visualize the context of significant motifs. This is easier with tools like SLiMFinder that output such alignments for visualization as part of the results. In extreme cases, sequences may have diverged to the point that conserved SLiMs are the only detectable homology. This will be impossible to distinguish from convergent evolution but should not affect the performance of SLiM discovery tools. There can also be problems with very large datasets. The "Unrelated Protein Cluster" (UPC) method employed by SLiMFinder works by clustering proteins via BLAST homology connections such that no proteins in one UPC will have detectable homology with any proteins in a different UPC. This does not necessarily mean that all the proteins in a cluster will share sequence homology: if protein A is homologous to B and B is homologous to C in a different region/domain, A and C will be grouped in the same UPC despite having no direct homology. For large datasets of multi-domain proteins, such as mammalian proteomes, this can result in a substantial proportion of the data (in the order of half the proteome) clumping together into a single giant UPC (data not shown).

3.2.2 Profile-Based Methods

Another popular set of programs for motif discovery is the MEME Suite of motif-based sequence analysis tools [72]. MEME [114] was developed originally as a method for novel motif discovery in DNA sequences. Genomic sequences are still its main focus but it can be used for finding signals in any biological sequence and has been applied to SLiM discovery. MEME uses ungapped PSSMs to represent motifs as extracted from an unaligned set of input sequences. It assumes that each sequence in the starting dataset contains an instance of the motif and employs the expectation-maximization algorithm [115, 116] to find motif patterns with high likelihood. Haslam and Shields [81] have found that MEME cannot perform as well as SLiMFinder, which is based on regular expressions, unless evolutionary weighting and local conservation filtering are applied. In that case both approaches are shown to be complementary, with some motifs being only returned by one of them, suggesting that their joint application can lead to an extended coverage of the results. It should be noted, however, that MEME does not return a significance estimate akin to SLiMChance and therefore this analysis was based on the ranks of positive predictions. Other tools in the MEME Suite extend the reach of the main algorithm by allowing searches, comparisons, and functional predictions for DNA and/or protein motifs. Notably, motif discovery is improved with GLAM2 [74] by incorporating gaps of flexible

length in the definition of motifs. Given a set of sequences believed to share one or more motifs it will perform a gapped alignment of them to find, score and rank all conserved patterns. Like MEME, GLAM2 is optimized for DNA motif discovery.

NestedMICA [117, 118] uses different probabilistic models to represent motif-carrying and non-motif fragments in the input sequences, identified through a nested Sampling strategy. The motifs are represented by a set of profiles extracted from the provided data and a predefined, non-homogeneous model of "uninteresting" background information serves as reference. These are all combined in an HMM model that is updated in each iterative step of nested sampling after discarding a certain fraction of sequence space and until the likelihood of the resulting motif profile is maximized. The output of NestedMICA is a profile for each motif, displayed as a sequence logo. An assessment of its performance against MEME over a purposely built benchmark dataset showed that NestedMICA can retrieve more true positive hits while reducing the false negatives at the same time [118]. However, although the information content values in each column of the logo give a hint on variability and conservation, the program does not provide any significance measure of motif support.

3.2.3 (l, d) Motif Searches

One common motif formulation for general de novo motif discovery is the "(l, d) motif search" (LDMS) (also known as a "planted motif search" or "(l, d) challenge problem"). LDMS algorithms search for all motifs of total length l (including wildcards) with up to d mismatches (i.e., 0-d wildcards). There are many LDMS algorithms and programs; recent examples include qPMS7 [119, 120]. LDMS algorithms are generally developed and tested for DNA motif discovery, but are rarely benchmarked or optimized for protein searches, which have very different constraints and criteria. Developments frequently concentrate on computational performance (i.e., speed) but often overlook important biological considerations, such as evolutionary relationships that can bias results, low support in the input data, or the possibility that there may be *no* real motifs in the data to find, which results in a lack of statistical significance. For these reasons, LDMS tools are not generally recommended for de novo SLiM discovery.

This is not to say that no LDMS algorithms are useful, but it is inadvisable to apply an algorithm to protein motif prediction if it has only been benchmarked on DNA data. Without this estimation of statistical significance, applying an LDMS algorithm to a dataset that may not contain a motif (of the "(l, d)" nature being sought) is almost guaranteed to generate false positive predictions. In principle, LDMS motif space could be modelled for a statistical assessment of motif support, although it is often over-constrained by the need to fix the l and d parameters prior to searching. Another feature of LMDS algorithms that can be problematic is that they

do not generally return a motif in the way defined in Subheading 1. Instead, the output is a consensus sequence and a list of variants with mismatches. Because these mismatches can occur in different positions in each motif instance, it can be difficult to generate a regular expression that captures the variability, although they could conceivably be coupled to an alignment-based algorithm to construct a sequence profile. If the natural incorporation of mismatches from the consensus "motif" could be correctly incorporated into a robust statistical framework like SLiMChance [17, 41], LDMS algorithms could yet prove useful for de novo SLiM discovery in real biological data.

3.2.4 Co-occurrence Methodologies

Correlated Motif Mining (CMM) methods suppose that motifs, being short and flexible, may be directly involved in interactions between larger domains. D-MOTIF and D-STAR [121] are the exact and the approximate versions of an algorithm built on the LDMS model to find instances of two motifs that are correlated in the same interaction in a PPI network. Leung et al. raised adequacy and scalability issues in D-MOTIF and D-STAR and proposed an alternative model to find motif pairs based on fast clustering heuristics that they implemented as MotifCluster [122]. Another CMM method, SLIDER [123], incorporates structural data and maps correlated LDMS occurrences onto PPI interfaces. In addition to the underlying LDMS drawbacks, the main issue with CMM is that there are no clear examples of SLiM-SLiM PPI and it is highly likely that either or both motifs returned are actually structural/family signature motifs of domain-based interactions. Another weakness of this method is that it cannot identify motifs that all interact with the same single partner. In addition, whole interactome data is required for the prediction, which limits application.

FIRE-pro [124] is another CMM method that uses mutual information (MI) to discover motifs whose presence/absence correlates with a biological feature of the proteins in question. FIRE-pro first builds gapped *k*-mers (fixed position motifs with *k* defined positions separated by runs of 0–3 wildcard "gap" positions) and calculates their mutual information with the biological classification of the proteins in the dataset. Motifs that tend to be present in proteins that are positive for the biological feature of interest (e.g., GO category or PPI partner) and absent in negative proteins have a high MI score. This is then compared to randomized feature classification and only those motifs with higher MI than 10,000 randomizations are retained. More informative descriptions of the significant motifs are ultimately informed by subjecting those to a greedy search of variants that increases their degeneracy. FIRE-pro does not use protein disorder information and might therefore return structural motifs; a possible future improvement would be to couple FIRE-pro with disorder and RLC masking. Although FIRE-pro does filter evolutionary relationships, the BLAST

E-value used is extremely stringent (1e-50) and therefore it is highly likely that homology will be influencing some of the MI associations. One also needs to be very careful of complex PPI relationships and PPI-GO correlations that could give rise to false associations, particularly in multi-domain proteins. With those caveats in mind, the high efficiency of FIRE-pro makes it suitable to analyzing proteome-scale data to discover, rediscover, and make an initial functional prediction of SLiMs.

3.3 Sequence Property/Feature Methods

Not all de novo SLiM discovery tools make predictions based on sequence specificity. Whilst not the focus of this review, it is useful to briefly highlight a few of these other methods. Frequently, these approaches will complement a sequence-based approach and can provide useful corroborating evidence regarding the nature/importance of a given site. Alternatively, they might be useful to preprocess data for sequence-based predictions and or rank/filter results as explored in Subheading 2. ANCHOR [125, 126], α-MoRF-PredII [44, 127], MoRFPred [128], and MFSPSSMpred [106] use signs of propensity for structure within an IDR to predict potential SLiM- or MoRF-containing regions. SLiMPred [129] takes a more flexible machine learning approach and uses annotated ELM instances and structural, biophysical, and biochemical attributes predicted from the primary sequence to build a bidirectional recurrent neural network that generates a per-residue probability of being part of a SLiM. Output can then be scanned for clusters of SLiM-like residues. Although these methods do not generate motifs as such, they have the advantage of being able to identify interaction sites that lack the sequence specificity of SLiMs and/or where additional data (e.g., homologues, structures, or PPI) are unavailable. Where homologues are available, alignment-based tools can then be used to generate a motif consensus or profile for the identified region.

Efforts have also been made to predict novel SLiMs directly from structural data in the Protein Data Bank (PDB) repository [130]. D-MIST [131] is a profile-based method that uses structure-derived binding profiles to interrogate sequence databases for novel PPI. It first extracts motifs known to bind the same domain from structural complexes where the latter is present. The motifs are then used to seed a Gibbs sampling search of similar sequences from empirical binary interactions, from which PSSMs can be constructed and used to find other proteins with a similar interface. Similarly, SLiMDIet [132] has been developed to identify SLiMs in the binding interface of solved structures of PPI complexes in PDB. Pfam domain binding interfaces are clustered by structural similarity and the residues belonging to the domain face and the partner face in each cluster are then aligned. Based on the contacts of the interaction, SLiMs are extracted as flexible, gapped PSSMs, and their statistical significance assessed through PPI data.

Stein and Aloy [90] took a more focused approach and specifically modelled the features of known SLiMs from solved DMI structures in PDB [89], identifying a signature stretched and elongated structure that was characteristic of DMI peptides. Using machine learning and contextual filters, they then predicted novel DMI from PDB PPI, deriving consensus patterns using SLiMFinder [41] where possible. As with SLiMDIet, significance of predicted motifs was assessed using overrepresentation in PPI data. These methods show a lot of promise and are likely to become increasingly useful as the number of solved DMI continues to increase.

3.4 Statistics for De Novo SLiM Discovery

When it comes to motif prediction (and benchmarking of SLiM discovery algorithms) an indication of significance through testing on data without a genuine signal is crucial. This is because one of the primary challenges for de novo SLiM discovery is determining whether there is a motif to be found at all. There are good discussions of motif statistics for both regex [17] and profile [19] approaches elsewhere. Instead, this review concentrates on some of the biological and practical considerations that are likely to be pertinent, whatever the specific statistical model employed.

3.4.1 Sequence Space Considerations

The considerations for sequence space when searching for de novo motifs are much the same as previously discussed for making dataset-level overrepresentation assessments for known SLiMs. The main difference is that the additional multiple testing corrections for motif space in de novo searches dramatically reduce significance levels and thus any loss of signal by erroneously removing real instances (either by masking them out or excluding the parent sequence from the dataset) can have much stronger consequences than for the prediction of known motifs. As a result, whereas known motif prediction has a tendency to err towards stringent filtering, de novo prediction generally needs to maximize the available signal, even if that comes at the cost of increased noise. For a discussion of some approaches in dataset construction that can influence the signal–noise ratio, *see* [11].

Conserved Versus Non-conserved Motif Occurrences

Motif enrichment is potentially confounded by two opposing trends occurring in the dataset: enrichment for functional motifs [63, 71] and depletion of non-functional motifs [104]. SLiM-mediated PPI are frequently cooperative and/or competitive [8, 133] and therefore having competing sites of interaction in the wrong place at the wrong time could upset the delicate balance of signalling. Random occurrences of a SLiM in proteins that (could) interact with a given SLiM-binding domain are probably under negative selection [104]. These conflicting signals could obviously hamper SLiM prediction as the true random background would be smaller than that modelled. In reality, the number of motifs expected to occur by chance is usually quite small and therefore the loss of these

from the actual signal is hopefully not too detrimental. The fact that SLiM prediction by overrepresentation works for many known examples [41, 63] supports this notion. Nonetheless, the observation that current SLiM prediction methods seem to be on the cusp of successful SLiM discovery in many cases implies that slight increases in signal or decreases in noise could be the difference between a SLiM being significantly overrepresented or not. Correct modelling of negative selection could potentially push some of these datasets into the detectable signal–noise range.

3.4.2 Motif Space Considerations

The main difference between searching for known SLiMs and de novo discovery is the large multiple-testing correction that needs to be done when the SLiM is unknown. In essence, any possible motif that could have been constructed by the de novo method could be overrepresented in the data. One of the major advances of SLiMFinder [41] was the SLiMBuild algorithm that tightly controlled the building of the motif space and thus enabled an exact calculation of the number of possible motifs. TEIRESIAS [112], which underpinned earlier algorithms such as DILIMOT [71, 113] and SLiMDisc [42, 77], does not control the motif space in the same way, which makes it hard to estimate how many different motifs are actually being tested. There is clearly a trade-off in the selection of SLiMBuild parameters, such as the maximum number of wildcards between defined positions (set to 2 by default): increasing the number of motifs increases the chance of the correct motif being within that motif space but also increases the size of the significance correction that is required. It should be noted that there might be ways of improving performance by altering the motif-building rules. Many motifs predicted by Neduva et al. [71, 113], for example, have three consecutive wildcards and would thus be missed by SLiMFinder defaults. It could be that by restricting the total number of wildcards akin to LDMS algorithms, rather than constraining the wildcards between pairs of defined positions, a more appropriate motif space could be constructed. Other approaches could reduce the motif space to focus on specific motif types. SLiMFinder, for example, includes an "alpha helix" mode that considers the helix periodicity and only searches for motifs in positions i, $i+3/4$, $i+7$, although this is yet to be benchmarked.

Motif Independence and Clouding

The statistics for SLiMs with different numbers of defined positions are generally kept separate as clearly they are not independent. PxxP, for example, is a sub-motif of PxxPx[KR] and their frequencies will clearly be related. Unfortunately, when such overlapping motifs are returned, it is currently impossible to tell whether the shorter is enriched because of enrichment of the larger or vice versa. Early attempts with SLiMFinder to incorporate the frequency of shorter versions of the returned pattern to assess significance were not very successful (data not shown) but it is a

potential improvement to the statistical model that could be considered in future. Instead, SLiMFinder groups overlapping motifs (based on occurrences in proteins, not pattern definitions) into "clouds," allowing the user to rapidly identify different variants of the same general motif prediction [41]. Similarities between different SLiMs can also be identified a posteriori using CompariMotif [69]. Aligned occurrences of SLiMFinder clouds could be subsequently passed through an alignment-based tool, such as SLiMMaker or MEME, to give a more complete definition of the cloud.

A particular challenge to the statistical assumption of motif independence is the treatment of ambiguity. Ambiguous motifs are clearly not independent from the variants used to build them, thus increasing the motif space by all possible ambiguous motifs would unfairly and dramatically inflate the multiple testing correction. The current implementation of SLiMChance therefore ignores ambiguity when calculating and correcting for the size of the SLiMBuild motif space. For a complete motif space and limited equivalencies, this does not seem to affect the model too badly. The affects have not been well modelled, however. If too many equivalence sets are used it could result in inflation of significance for ambiguous motifs. In general, while ambiguous motifs are often more informative than pure fixed position motifs, confidence that they represent reliable predictions can be substantially increased if there is also a pure fixed position motif returned in the same cloud.

Altered Alphabets and Specified Amino Acids

One way to reduce the size of the motif space being searched is to reduce the alphabet. This can be achieved by combining certain amino acids with similar properties. The utility of this approach is not clear, however. There is an obvious trade-off between reducing the motif space and increasing the probability of given patterns occurring by chance by increasing the corresponding frequency of the new characters. This is seen in the difference between protein and DNA motifs: the latter need to be longer and/or more abundant to achieve significance. A second problem with this idea is that the biological justification is not clear. Whilst one could imagine combining lysine and arginine as positively charged amino acids, for example, they are not equally used by known motifs [1].

A more powerful way (in principle) to reduce motif space is to mask out certain amino acids that are unlikely to be of interest. Alanine and glycine, for example, are generally considered to be quite boring. Because this reduces the sequence search space as well as the motif space, it gets around the problem of motifs becoming more likely to occur by chance. That said, it should be noted that all 20 amino acids are found in the defined positions of at least one known motif, so such filtering should be applied with caution. SLiMFinder includes an option to mask out specific amino acids but this has not been benchmarked.

An alternative that is less extreme, and often easier to justify biologically, is to focus on motifs that contain a specific amino acid, such as a tyrosine if tyrosine phosphorylation is known to be important. This does not reduce the motif space as dramatically but can ease interpretation. An obvious application for such reductions would be the prediction of PTM sites. Indeed, in this instance it is possible to *expand* the alphabet by encoding modified residues with a twenty-first letter (e.g., Z) and then (optionally) specifying that motifs should have this letter. Although this is an option in SLiMFinder, we are not aware of any published work exploring or using this feature. A possible exception is the successful discovery of terminal motifs by SLiMFinder, which adds N-terminus (^) and C-terminus ($) characters to each protein before expanding the protein alphabet to include them [41, 63].

Controlling Motif Space with Defined Queries

The other way to reduce the motif space is to build it on a specific sequence (or set of sequences) rather than looking at all possible motifs. This is the basis of QSLiMFinder ("Query SLiMFinder" [140], Table 1), in which the motif space is built on a specific "query" protein sequence. Query motifs are then assessed for enrichment in another set of proteins that, for example, share a common PPI partner using the basic approach of SLiMFinder. Clearly the query sequence(s) used for building the motif space cannot be included in the search space itself as this would artificially inflate the support for those motifs in the data and thus there is a trade-off between reducing the motif space multiple testing and loss of signal in the data. QSLiMFinder can substantially improve search sensitivity over SLiMFinder where the query protein/region is quite small, e.g., a short binding region has been identified (data not shown). One caveat is that ambiguity cannot be usefully included in the statistical model: in order to be valuable, motif variants outside of the query must be included but these inflate the motif search space by an unknown amount. One solution is to return ambiguous motifs but only pay heed to those that also have a significant fixed-position pattern returned in the same cloud.

3.5 Low Complexity Motifs

Low complexity motifs are motifs that are dominated by a small number of amino acids. Examples include proline-rich motifs, serine-rich motifs, RG and RS repeats, and patches of positive or negative charge. Low complexity regions have a tendency to return a lot of similar motifs, especially if variable-length wildcards are used. Masking low complexity sequences can avoid this but at the risk of missing genuine low complexity motifs. As some low-complexity motifs are certainly functional, this trade-off is largely going to be determined by the scale of the analysis being performed. Large-scale analyses will probably want to mask low complexity regions more stringently, as the probability of throwing

together some proteins that share low complexity regions by chance will be high. Focused small-scale studies, on the other hand, should be more cautious. When such motifs are returned, the question must be asked: does the low complexity motif simply reflect a sequence bias, or does any such sequence bias reflect a high frequency of functional low complexity motifs in the dataset? As with most bioinformatics predictions, SLiM discovery tools cannot themselves answer this and additional evidence must be considered.

Large-scale application of DILIMOT to PPI data showed a marked tendency to recover proline- and serine-rich motifs [63, 71]. This could reflect a genuine bias towards these residues in SLiMs or might be a reflection of their occurrence in low complexity proline- and serine-rich regions of proteins, which are conserved (at the level of amino acid enrichment) between species. A large-scale analysis of human interactome data using SLiMFinder did not find the same degree of bias [63]. As this analysis masked out very low complexity regions and used motif conservation (as opposed to rediscovery), the implication is that the enrichment in the Neduva et al. study is largely due to low complexity regions. Of course, such low complexity regions are presumably functional and are genuinely enriched in the data: the question is whether they are enriched because they mediate the PPI of interest, or whether they are a different recurring feature of proteins that some methods are particularly sensitive to finding.

3.6 Predicting SLiMs from Short Peptide Data

The importance of correcting for evolutionary relationships has been stressed in the preceding sections. Sometimes, such correction is neither appropriate nor necessary because the input data does not have evolutionary relationships to worry about. Examples of this are phage display and peptide libraries, which sample a large sequence space and select for short peptide regions that can bind a desired partner. These techniques can be very useful for SLiM discovery as they can substantially increase the number of motif occurrences. Indeed, they can potentially be used to identify motifs in singleton interactions where enrichment in true PPI is not possible. The proportion of input sequences assumed to contain the motif is also high, which makes profile methods such as those in the MEME Suite [72] popular for such applications. SLiMFinder can be used to predict significantly enriched regex motifs from peptide data (*see* example 3 in the original paper [41]) but the background amino acid frequencies will need to be corrected to represent the pre-selection peptide sequences. The evolutionary filtering and sequence masking should also be switched off but redundancy in the peptides should be removed prior to analysis, unless it represents true independent enrichment.

Domain binding and phosphorylation targets obtained with these methods can also serve as input for specialized SLiM discovery tools.

MOTIPS [134] converts the given data into a sequence profile (after a normalization step to ensure consistent scoring among evidence from different sources), which is used in turn for whole-proteome scans that produce a list of potential domain targets. The score for each putative motif is combined with feature assessments based on residue-specific, precomputed values of conservation, solvent accessibility and disorder bias and then compared with a validated sequence set. The final output of MOTIPS is a ranked list of motif hits according to the likelihood of interacting with the domain of interest. In this fashion, the ability to independently recover the domain used for the original experiment can be used to assess the success or failure in motif prediction.

3.7 Challenges to Interpretation of De Novo SLiM Predictions

Edwards et al. [63] provide a fairly detailed discussion of the challenges in interpreting de novo SLiM predictions. Fundamentally, there are two connected questions:

1. Is the motif genuinely enriched? (i.e., Is the statistical model good?)
2. Is the enrichment for the reasons postulated when the dataset was constructed?

Both questions are arguably impossible to answer by bioinformatics alone, although robust benchmarking and data exploration can get a good handle on the former. Assuming that the SLiM prediction program functions as intended, the question then becomes whether the assumptions of the statistical model are valid or whether violations of that model could generate false positive enrichment. Again, the big consideration here is one of underlying protein sequence bias versus motif-specific sequence bias. Trying different ways of masking the data and/or generating the background model can help get a handle on this. It is also important to check thoroughly for evolutionary relationships between sequences that may have escaped detection.

The second question is usually more important but harder to get a handle on: given that a set of target sequences S were selected for a reason (e.g., common PPI partner or subcellular location), and the analysis show they are enriched for motif M, what is the causal relationship between M and S? Essentially three explanations are feasible:

1. M is (at least in part) responsible for S. This is usually the desired outcome, and therefore the default explanation, but it should be concluded with caution without additional supporting data and/or follow-up experiments because the alternative explanations are also possible.
2. M is correlated with S but not causal. Non-independence of biological data makes it challenging to differentiate causation from correlation. Suppose, for example, all proteins in S bind a

protein A. The interactome of A is likely to be enriched for proteins targeted to a specific subcellular component C and/or share interactions with another protein B. Does motif M interact with protein A or protein B or target proteins to C?

3. M and S are unrelated. M is an enriched feature of the parent sequence dataset (e.g., the whole proteome) and its enrichment in S is purely chance.

There is currently no good way to distinguish between these explanations from an analysis of a single protein dataset but hints can be achieved by additional analyses. For example, specifically looking for enrichment in other PPI or GO datasets could give hints regarding non-causal correlations, whilst analyzing randomly assembled datasets of real proteins from the same source can give insights into nonspecific enrichment. (*See* [11] and [63] for further discussion of these issues.) Correlating occurrences of predicted SLiMs with different biological features in a similar vein to FIRE-pro [124] might prove to be very helpful in this endeavor.

Flanking regions of SLiMs have been shown to be important for both function and specificity of binding [47, 98]. It is therefore not surprising that patterns returned from de novo SLiM prediction of known motifs (i.e., true positive benchmarks) frequently include flanking residues beyond the database definition. There could be several reasons for this. Chance enrichment of particular flanking residues could result in the longer SLiM being significantly enriched due to enrichment of the shorter versions. Alternatively, the flanking residues could belong to a second co-occurring motif as part of a "switch" in which two neighboring or overlapping SLiMs mediate mutually exclusive binding [13, 60], possibly by the steric hindrance introduced by the bound globular domain [8]. Finally, of course, there remains the possibility that literature/database definition of the SLiM is incomplete, and that the enriched flanking position is actually part of the SLiM or a sub-class thereof.

4 Concluding Remarks

Computational SLiM prediction is a blossoming field with new methods being developed on a regular basis. Whilst welcome, this can be confusing for the uninitiated, who may struggle to choose from the various tools available. There is no universal best solution to all SLiM prediction problems and so the nature of the input data as well as any potential follow up must be taken into consideration. Is the task motif instance prediction, or de novo discovery? Is the target of the search a single protein of interest, an alignment, a small dataset of multiple proteins, or a whole proteome/interactome? Are motifs likely to be shared by family members and thus

have arisen by divergent evolution across homologues, or are they independent convergently evolved instances in unrelated proteins? Modelling the latter is the most common approach for de novo prediction but it is crucial to correct for evolutionary relationships in the data or else the former will be identified without realizing it. Despite the advances made by DILIMOT [71, 113], SLiMDisc [77] and SLiMFinder [41] in this area, a surprising number of de novo discovery tools are still published that overlook this fundamental discovery bias.

Performance benchmarking is often overlooked but this is vital if one is to understand the strengths, weaknesses and biases of the predictions produced. We have developed SLiMBench and made it part of the SLiMSuite package, which we hope will make this exercise easier in future. Not only can this enable direct comparisons of method performance, it can also help optimize parameter settings for SLiM discovery. The importance of an estimate of statistical significance for de novo predictions cannot be overstated. Is there really a motif to be found in the data? If there may not be, what is the False Positive Rate of the method being applied and what implications does this have given the scale of the analysis and planned experimental follow up? This is particularly crucial for large-scale analyses. Statistical significance is not only important for estimated False Discovery Rates; it is the only metric that is comparable between datasets with different sequence numbers, lengths, and/or composition.

Computational SLiM discovery has made a lot of progress over the last decade in successfully identifying overrepresented motifs. Nevertheless, Davey et al. point out that "Computational approaches, which should lead and focus experimental discovery, are in many ways lagging behind the advances of the experimentalists… [and] have yet to reveal the expected multitude of novel motif classes and instances." [1] Methods that differentiate between causal and coincidental enrichment are the key to the future success of bioinformatics approaches to this challenging yet important biological problem.

References

1. Davey NE, Van Roey K, Weatheritt RJ et al (2012) Attributes of short linear motifs. Mol Biosyst 8(1):268–281
2. Pawson T (1995) Protein modules and signalling networks. Nature 373(6515):573–580
3. Davis BD, Tai PC (1980) The mechanism of protein secretion across membranes. Nature 283(5746):433–438
4. Aasland R, Abrams C, Ampe C et al (2002) Normalization of nomenclature for peptide motifs as ligands of modular protein domains. FEBS Lett 513(1):141–144
5. Puntervoll P, Linding R, Gemund C et al (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31(13):3625–3630
6. Pancsa R, Fuxreiter M (2012) Interactions via intrinsically disordered regions: what kind of motifs? IUBMB Life 64(6):513–520

7. Neduva V, Russell RB (2006) Peptides mediating interaction networks: new leads at last. Curr Opin Biotechnol 17(5):465–471
8. Diella F, Haslam N, Chica C et al (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13:6580–6603
9. Dinkel H, Van Roey K, Michael S et al (2014) The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res 42(1):D259–D266
10. Mi T, Merlin JC, Deverasetty S et al (2012) Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. Nucleic Acids Res 40(Database issue):D252–D260
11. Davey NE, Edwards RJ, Shields DC (2010) Computational identification and analysis of protein short linear motifs. Front Biosci (Landmark Ed) 15:801–825
12. Neduva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. FEBS Lett 579(15):3342–3345
13. Van Roey K, Gibson TJ, Davey NE (2012) Motif switches: decision-making in cell regulation. Curr Opin Struct Biol 22(3):378–385
14. Vyas J, Nowling RJ, Maciejewski MW et al (2009) A proposed syntax for Minimotif Semantics, version 1. BMC Genomics 10:360
15. Davey NE, Trave G, Gibson TJ (2011) How viruses hijack cell regulation. Trends Biochem Sci 36(3):159–169
16. Garamszegi S, Franzosa EA, Xia Y (2013) Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks. PLoS Pathog 9(12):e1003778
17. Davey NE, Edwards RJ, Shields DC (2010) Estimation and efficient computation of the true probability of recurrence of short linear protein sequence motifs in unrelated proteins. BMC Bioinform 11:14
18. Sigrist CJ, Cerutti L, Hulo N et al (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3(3):265–274
19. Xia X (2012) Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction. Scientifica (Cairo) 2012:917540
20. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9):755–763
21. Krogh A, Brown M, Mian IS et al (1994) Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235(5):1501–1531
22. Obenauer JC, Cantley LC, Yaffe MB (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31(13):3635–3641
23. Yoon BJ (2009) Hidden Markov models and their applications in biological sequence analysis. Curr Genomics 10(6):402–415
24. Seiler M, Mehrle A, Poustka A et al (2006) The 3of5 web application for complex and comprehensive pattern matching in protein sequences. BMC Bioinform 7:144
25. Davey NE, Haslam NJ, Shields DC et al (2010) SLiMSearch: a webserver for finding novel occurrences of short linear motifs in proteins, incorporating sequence context. Lect Notes Bioinform 6282:50–61
26. Meszaros B, Dosztanyi Z, Simon I (2012) Disordered binding regions and linear motifs–bridging the gap between two models of molecular recognition. PLoS One 7(10): e46829
27. Davey NE, Shields DC, Edwards RJ (2009) Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics 25(4):443–450
28. Brown CJ, Takayama S, Campen AM et al (2002) Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol 55(1):104–110
29. Tóth-Petróczy A, Mészáros B, Simon I et al (2008) Assessing conservation of disordered regions in proteins. Open Proteom J 1: 46–53
30. Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. Bioinformatics 23(8):950–956
31. Remaut H, Waksman G (2006) Protein-protein interaction through beta-strand addition. Trends Biochem Sci 31(8):436–444
32. Cino EA, Choy WY, Karttunen M (2013) Conformational biases of linear motifs. J Phys Chem B 117(50):15943–15957
33. Abeln S, Frenkel D (2008) Disordered flanks prevent peptide aggregation. PLoS Comput Biol 4(12):e1000241
34. Sehnal D, Varekova RS, Huber HJ et al (2012) SiteBinder: an improved approach for comparing multiple protein structural motifs. J Chem Inf Model 52(2):343–359
35. Buljan M, Chalancon G, Eustermann S et al (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol Cell 46(6):871–883
36. Weatheritt RJ, Davey NE, Gibson TJ (2012) Linear motifs confer functional diversity onto

splice variants. Nucleic Acids Res 40(15): 7123–7131
37. Weatheritt RJ, Gibson TJ (2012) Linear motifs: lost in (pre)translation. Trends Biochem Sci 37(8):333–341
38. Wan J, Qian SB (2014) TISdb: a database for alternative translation initiation in mammalian cells. Nucleic Acids Res 42(1):D845–D850
39. Kochetov AV (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. Bioessays 30(7):683–691
40. UniProt C (2014) Activities at the universal protein resource (UniProt). Nucleic Acids Res 42(1):D191–D198
41. Edwards RJ, Davey NE, Shields DC (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PLoS One 2(10):e967
42. Davey NE, Edwards RJ, Shields DC (2007) The SLiMDisc server: short, linear motif discovery in proteins. Nucleic Acids Res 35(Web Server issue):W455–W459
43. Flicek P, Amode MR, Barrell D et al (2014) Ensembl 2014. Nucleic Acids Res 42(1):D749–D755
44. Oldfield CJ, Cheng Y, Cortese MS et al (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry 44(37): 12454–12470
45. Mohan A, Oldfield CJ, Radivojac P et al (2006) Analysis of molecular recognition features (MoRFs). J Mol Biol 362(5):1043–1059
46. Vacic V, Oldfield CJ, Mohan A et al (2007) Characterization of molecular recognition features, MoRFs, and their binding partners. J Proteome Res 6(6):2351–2366
47. Stein A, Aloy P (2008) Contextual specificity in peptide-mediated protein interactions. PLoS One 3(7):e2524
48. Teyra J, Sidhu SS, Kim PM (2012) Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains. FEBS Lett 586(17):2631–2637
49. Liu Y, Woods NT, Kim D et al (2011) Yeast two-hybrid junk sequences contain selected linear motifs. Nucleic Acids Res 39(19):e128
50. Eisenhaber B, Eisenhaber F (2010) Prediction of posttranslational modification of proteins from their amino acid sequence. Methods Mol Biol 609:365–384
51. Trost B, Kusalik A (2011) Computational prediction of eukaryotic phosphorylation sites. Bioinformatics 27(21):2927–2935
52. Sigrist CJ, De Castro E, Langendijk-Genevaux PS et al (2005) ProRule: a new database containing functional and structural information on PROSITE profiles. Bioinformatics 21(21): 4060–4066
53. Sigrist CJ, de Castro E, Cerutti L et al (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41(Database issue):D344–D347
54. Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res 40(Database issue):D302–D305
55. Punta M, Coggill PC, Eberhardt RY et al (2012) The Pfam protein families database. Nucleic Acids Res 40(Database issue): D290–D301
56. Chica C, Labarga A, Gould CM et al (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. BMC Bioinform 9:229
57. Via A, Gould CM, Gemund C et al (2009) A structure filter for the Eukaryotic Linear Motif Resource. BMC Bioinform 10:351
58. Weatheritt RJ, Jehl P, Dinkel H et al (2012) iELM–a web server to explore short linear motif-mediated interactions. Nucleic Acids Res 40(Web Server issue):W364–W369
59. Dinkel H, Chica C, Via A et al (2011) Phospho.ELM: a database of phosphorylation sites–update 2011. Nucleic Acids Res 39(Database issue):D261–D267
60. Van Roey K, Dinkel H, Weatheritt RJ et al (2013) The switches.ELM resource: a compendium of conditional regulatory interaction interfaces. Sci Signal 6(269):rs7
61. Jin J, Pawson T (2012) Modular evolution of phosphorylation-based signalling systems. Philos Trans R Soc Lond B Biol Sci 367(1602):2540–2555
62. Songyang Z, Blechner S, Hoagland N et al (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases. Curr Biol 4(11):973–982
63. Edwards RJ, Davey NE, O'Brien K et al (2012) Interactome-wide prediction of short, disordered protein interaction motifs in humans. Mol Biosyst 8(1):282–295
64. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29
65. Hamosh A, Scott AF, Amberger J et al (2000) Online mendelian inheritance in man (OMIM). Hum Mutat 15(1):57–61
66. Goel R, Harsha HC, Pandey A et al (2012) Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. Mol Biosyst 8(2):453–463

67. Safran M, Dalah I, Alexander J et al (2010) GeneCards Version 3: the human gene integrator. Database (Oxford) 2010:baq020
68. Davey NE, Haslam NJ, Shields DC et al (2011) SLiMSearch 2.0: biological context for short linear motifs in proteins. Nucleic Acids Res 39(Web Server issue):W56–W60
69. Edwards RJ, Davey NE, Shields DC (2008) CompariMotif: quick and easy comparisons of sequence motifs. Bioinformatics 24(10): 1307–1309
70. Marsico A, Scheubert K, Tuukkanen A et al (2010) MeMotif: a database of linear motifs in alpha-helical transmembrane proteins. Nucleic Acids Res 38(Database issue):D181–D189
71. Neduva V, Linding R, Su-Angrand I et al (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3(12):e405
72. Bailey TL, Boden M, Buske FA et al (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37(Web Server issue):W202–W208
73. Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics 27(7):1017–1018
74. Frith MC, Saunders NF, Kobe B et al (2008) Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput Biol 4(4):e1000071
75. Bailey TL, Gribskov M (1997) Score distributions for simultaneous matching to multiple motifs. J Comput Biol 4(1):45–59
76. de Castro E, Sigrist CJ, Gattiker A et al (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34(Web Server issue):W362–W365
77. Davey NE, Shields DC, Edwards RJ (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. Nucleic Acids Res 34(12):3546–3554
78. Peng ZL, Kurgan L (2012) Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci 13(1):6–18
79. Deng X, Eickholt J, Cheng J (2012) A comprehensive overview of computational protein disorder prediction methods. Mol Biosyst 8(1):114–121
80. Dosztanyi Z, Csizmok V, Tompa P et al (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21(16):3433–3434
81. Haslam NJ, Shields DC (2012) Profile-based short linear protein motif discovery. BMC Bioinform 13:104
82. Sickmeier M, Hamilton JA, LeGall T et al (2007) DisProt: the database of disordered proteins. Nucleic Acids Res 35(Database issue):D786–D793
83. Chen JW, Romero P, Uversky VN et al (2006) Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. J Proteome Res 5(4):879–887
84. Tompa P, Fuxreiter M, Oldfield CJ et al (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. Bioessays 31(3):328–335
85. Williams RW, Xue B, Uversky VN et al (2013) Distribution and cluster analysis of predicted intrinsically disordered protein Pfam domains. Intrins Disord Prot 1:e25724
86. Schaeffer RD, Jonsson AL, Simms AM et al (2011) Generation of a consensus protein domain dictionary. Bioinformatics 27(1):46–54
87. Towse CL, Daggett V (2012) When a domain is not a domain, and why it is important to properly filter proteins in databases: conflicting definitions and fold classification systems for structural domains make filtering of such databases imperative. Bioessays 34(12):1060–1069
88. Linding R, Russell RB, Neduva V et al (2003) GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res 31(13):3701–3708
89. Mosca R, Ceol A, Stein A et al (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res 42(1):D374–D379
90. Stein A, Aloy P (2010) Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures. PLoS Comput Biol 6(5):e1000789
91. Brannetti B, Helmer-Citterich M (2003) iSPOT: a web tool to infer the interaction specificity of families of protein modules. Nucleic Acids Res 31(13):3709–3711
92. Trabuco LG, Lise S, Petsalaki E et al (2012) PepSite: prediction of peptide-binding sites from protein surfaces. Nucleic Acids Res 40(Web Server issue):W423–W427
93. Perrodou E, Chica C, Poch O et al (2008) A new protein linear motif benchmark for multiple sequence alignment software. BMC Bioinform 9:213
94. Sayers EW, Barrett T, Benson DA et al (2011) Database resources of the National Center for

Biotechnology Information. Nucleic Acids Res 39(Database issue):D38–D51
95. Balla S, Thapar V, Verma S et al (2006) Minimotif Miner: a tool for investigating protein function. Nat Methods 3(3):175–177
96. Dinkel H, Sticht H (2007) A computational strategy for the prediction of functional linear peptide motifs in proteins. Bioinformatics 23(24):3297–3303
97. Davey NE, Cowan JL, Shields DC et al (2012) SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. Nucleic Acids Res 40(21):10628–10641
98. Chica C, Diella F, Gibson TJ (2009) Evidence for the concerted evolution between short linear protein motifs and their flanking regions. PLoS One 4(7):e6052
99. O'Brien KT, Haslam NJ, Shields DC (2013) SLiMScape: a protein short linear motif analysis plugin for Cytoscape. BMC Bioinform 14:224
100. Davey NE, Haslam NJ, Shields DC et al (2010) SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. Nucleic Acids Res 38(Web Server issue):W534–W539
101. Plewczynski D, Basu S, Saha I (2012) AMS 4.0: consensus prediction of post-translational modifications in protein sequences. Amino Acids 43(2):573–582
102. Kerrien S, Aranda B, Breuza L et al (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40(Database issue):D841–D846
103. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
104. Via A, Gherardini PF, Ferraro E et al (2007) False occurrences of functional motifs in protein sequences highlight evolutionary constraints. BMC Bioinform 8:68
105. Nguyen Ba AN, Yeh BJ, van Dyk D et al (2012) Proteome-wide discovery of evolutionary conserved sequences in disordered regions. Sci Signal 5(215):rs1
106. Fang C, Noguchi T, Tominaga D et al (2013) MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinform 14:300
107. Chou MF, Schwartz D (2011) Biological sequence motif discovery using motif-x. *Curr Protoc Bioinform* Chapter 13, Unit 13 15–24
108. Orchard S (2012) Molecular interaction databases. Proteomics 12(10):1656–1662
109. Jonassen I (1997) Efficient discovery of conserved patterns using a pattern graph. Comput Appl Biosci 13(5):509–522
110. Jonassen I, Collins JF, Higgins DG (1995) Finding flexible patterns in unaligned protein sequences. Protein Sci 4(8):1587–1595
111. Neuwald AF, Green P (1994) Detecting patterns in protein sequences. J Mol Biol 239(5):698–712
112. Rigoutsos I, Floratos A (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics 14(1):55–67
113. Neduva V, Russell RB (2006) DILIMOT: discovery of linear motifs in proteins. Nucleic Acids Res 34(Web Server issue): W350–W355
114. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28–36
115. Lawrence CE, Reilly AA (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins 7(1):41–51
116. Do CB, Batzoglou S (2008) What is the expectation maximization algorithm? Nat Biotechnol 26(8):897–899
117. Down TA, Hubbard TJ (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Res 33(5):1445–1453
118. Dogruel M, Down TA, Hubbard TJ (2008) NestedMICA as an ab initio protein motif discovery tool. BMC Bioinform 9:19
119. Dinh H, Rajasekaran S (2013) PMS: a panoptic motif search tool. PLoS One 8(12):e80660
120. Dinh H, Rajasekaran S, Davila J (2012) qPMS7: a fast algorithm for finding (l, d)-motifs in DNA and protein sequences. PLoS One 7(7):e41425
121. Tan SH, Hugo W, Sung WK et al (2006) A correlated motif approach for finding short linear motifs from protein interaction networks. BMC Bioinform 7:502
122. Leung HC, Siu MH, Yiu SM et al (2009) Clustering-based approach for predicting motif pairs from protein interaction data. J Bioinform Comput Biol 7(4):701–716
123. Boyen P, Van Dyck D, Neven F et al (2011) SLIDER: a generic metaheuristic for the discovery of correlated motifs in protein-protein interaction networks. IEEE/ACM Trans Comput Biol Bioinform 8(5):1344–1357
124. Lieber DS, Elemento O, Tavazoie S (2010) Large-scale discovery and characterization of

protein regulatory motifs in eukaryotes. PLoS One 5(12):e14444

125. Dosztanyi Z, Meszaros B, Simon I (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 25(20):2745–2746
126. Meszaros B, Simon I, Dosztanyi Z (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 5(5):e1000376
127. Cheng Y, Oldfield CJ, Meng J et al (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry 46(47): 13468–13477
128. Disfani FM, Hsu WL, Mizianty MJ et al (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28(12):i75–i83
129. Mooney C, Pollastri G, Shields DC et al (2012) Prediction of short linear protein binding regions. J Mol Biol 415(1):193–204
130. Rose PW, Bi C, Bluhm WF et al (2013) The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res 41(Database issue):D475–D482
131. Betel D, Breitkreuz KE, Isserlin R et al (2007) Structure-templated predictions of novel protein interactions from sequence information. PLoS Comput Biol 3(9):1783–1789
132. Hugo W, Sung WK, Ng SK (2013) Discovering interacting domains and motifs in protein-protein interactions. Methods Mol Biol 939:9–20
133. Gibson TJ (2009) Cell regulation: determined to signal discrete cooperation. Trends Biochem Sci 34(10):471–482
134. Lam HY, Kim PM, Mok J et al (2010) MOTIPS: automated motif analysis for predicting targets of modular protein domains. BMC Bioinform 11:243
135. Schwartz D, Gygi SP (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat Biotechnol 23(11): 1391–1398
136. Schwartz D, Chou MF, Church GM (2009) Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. Mol Cell Proteomics 8(2): 365–379
137. Villen J, Beausoleil SA, Gerber SA et al (2007) Large-scale phosphorylation analysis of mouse liver. Proc Natl Acad Sci U S A 104(5): 1488–1493
138. Wilson-Grady JT, Villen J, Gygi SP (2008) Phosphoproteome analysis of fission yeast. J Proteome Res 7(3):1088–1097
139. Zhai B, Villen J, Beausoleil SA et al (2008) Phosphoproteome analysis of Drosophila melanogaster embryos. J Proteome Res 7(4):1675–1682
140. Edwards RJ. SLiMSuite software package. 2013 [cited 25/1/14]; Available from: http://www.southampton.ac.uk/~re1u06/software/packages/slimsuite/

Chapter 7

Peptide Toxicity Prediction

Sudheer Gupta, Pallavi Kapoor, Kumardeep Chaudhary, Ankur Gautam, Rahul Kumar, and Gajendra P.S. Raghava

Abstract

Last decade has witnessed the revival of interest in peptides as potential therapeutics candidates. However, one of the bottlenecks in the success of therapeutic peptides in clinics is their toxicity towards eukaryotic cells. Therefore, considerable efforts have been made over the years both in wet and dry lab to overcome this limitation. With the advances in peptide synthesis, now it is possible to fine-tune the physicochemical properties of peptides by incorporating several chemical modifications and thus to optimize the peptide functionality in order to minimize the toxicity without compromising their therapeutic activity. Also various in silico tools for peptide toxicity prediction and peptide designing have been developed, which facilitates designing of therapeutic peptides with desired toxicity. In this chapter, we have discussed both wet lab and dry lab approaches used to optimize peptide toxicity. More emphasis has been given to describe the in silico method, ToxinPred, to predict the toxicity of peptide and about how to design a peptide or protein with desired toxicity by mutating minimum number of amino acids.

Key words ToxinPred, Peptide, Toxicity, Webserver

1 Introduction

Over the last decade, peptides have been emerged as potential therapeutics to fight various diseases, including cancer, diabetes, and cardiovascular diseases [1]. An unprecedented interest has been seen among the scientific community for exploiting peptides as therapeutics. Due to this, peptides have paved their way to clinical applications and proven to be of considerable importance with six peptide drugs gaining global sales of more than US $750 million in 2008 [2]. However, several promising candidates could soon be identified with the advent of high-throughput screening and advances in peptide synthesis. High specificity, high tissue penetration, and comparable low production cost make peptides a preferred choice of therapeutics over small molecules and antibod-

* Authors contributed equally.

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_7,

ies [3]. In addition, peptides can be easily modified compared to small molecules. However, immunogenicity, toxicity, and stability are the major concerns for peptide/protein-based therapeutics [4]. Therefore, proper care should be taken while designing a therapeutic peptide so that these limitations can be overcome. Among different types of toxicities, immunotoxicity has been addressed by several groups [5, 6]. There are numbers of in silico tools available, which can predict whether peptides would be immunogenic or not [7–10]. Peptide stability can be enhanced by several ways like cyclization, changing backbone chemistry, incorporation of nonnatural and modified residues, etc. [11]. But no attention has been paid so far for predicting toxicity of peptide or protein, which is essential for a peptide to become a potential therapeutic.

2 Techniques for Measuring Toxicity of Peptide

One of the major concerns while developing peptide-based therapeutics is their undesired toxicity towards eukaryotic cells. Various toxicity assays are used widely in drug discovery research to determine the toxicity of peptides. There are many ways to measure toxicity of peptides in vitro; however; assessing cell membrane integrity is one of the most common methods to measure cell viability and cytotoxic effects. Peptides that have cytotoxic effects often compromise cell membrane integrity. A brief description of few commonly used assays is given below:

2.1 LDH Leakage Assay

Lactate dehydrogenase (LDH) is a cytosolic enzyme present in almost all types of cells. When the integrity of plasma membrane is compromised, LDH is released from the cytoplasm into cell culture media [12]. This extracellular released LDH can be quantified in vitro enzymatic reaction. First, LDH catalyzes the conversion of lactate to pyruvate via reduction of NAD^+ to NADH. Second, diaphorase uses NADH to reduce a tetrazolium salt (INT) to a red formazan product. Therefore, the amount of formazan formed is directly proportional to cytotoxicity.

2.2 MTT Assay

This is one of the simplest and most reliable methods to measure cytotoxicity of cells. Cytotoxicity can also be monitored using the 3-(4, 5-dimethyl-2-thiazolyl)-2, 5-diphenyl-2H-tetrazolium bromide (MTT) assay [13]. In this assay, yellow MTT substrate is reduced in metabolically active cells by the action of dehydrogenase enzymes and generates intracellular purple formazan product, which can be solubilized by DMSO and easily quantified calorimetrically. The amount of formazan is inversely proportionate to cytotoxicity.

2.3 ATP-Based Assay

ATP is an outstanding marker for cell viability because of its presence in all metabolically active cells. ATP concentration decreases rapidly when cells undergo necrosis or apoptosis.

Therefore, monitoring ATP is a good indicator of cytotoxic effects. Diverse assays measuring ATP levels have been developed by various companies, which can be used with different analysis platforms including colorimetry, fluorimetry, and bioluminescence [14].

2.4 Hemolytic Assay

Toxicity of therapeutic peptides against normal eukaryotic cells is usually first checked by testing their hemolytic activity against red blood cells (RBCs). This assay is based on the quantitative measurements of released hemoglobin from membrane-compromised RBCs [15]. In this assay, peptides are incubated with freshly isolated RBCs for various time periods. After this incubation, released hemoglobin is quantified spectrometrically. The amount of released hemoglobin is directly proportional to hemolytic potency of peptide.

3 Ways to Improve Therapeutic Index

The therapeutic index of a peptide is a comparison of the amount of a therapeutic effect to the amount of toxicity. Peptides "Hits" with high therapeutic index in general are ideal candidates for further lead optimization and drug development. The major barrier to the use of peptides as potential drugs is their toxicity or ability to lyse eukaryotic cells particularly erythrocytes. Therefore, hemolytic activity needs to be minimized in order to convert potential lead molecules into useful drugs. The ability of a peptide to be hemolytic or toxic more or less depends on its physicochemical properties like length, charge, amino acid sequence, amphipathicity, helicity, and hydrophobicity [16]. Changing these properties will help to modify the activity of peptides or it may minimize the toxicity as well. There are various ways/modifications by which one can alter the physicochemical properties of peptides, which can reduce the hemolytic potency, or toxicity of peptides. But at the same time these modifications may also alter the therapeutic activity of peptides. Therefore, one has to fine-tune the physicochemical properties of peptides (by various modifications) in such a way that therapeutic activity can be enhanced whereas toxicity can be minimized (e.g., high therapeutic index). In order to reduce the toxicity or increase the therapeutic index of peptide leads, one of the most common approaches used in the wet lab is to develop various chemically modified peptide analogs followed by their in vitro evaluation of toxicity using above-described assays. Therefore, in the past, a great deal of attention has been focused on introducing chemical modifications in the native peptides to delineate features resulting in high therapeutic activity with low toxicity. Following is the brief description of most common modifications introduced in the peptides:

1. *Modification of peptide by changing amino acid content*
   This is one of the most common strategies used to alter the physicochemical properties of peptides. Different amino acids have different physicochemical properties; therefore alteration in amino acid content results in alteration in physicochemical properties of peptides and thus in biological activity as well. Certain combination of amino acid is toxic for cell, which can be avoided while designing therapeutic peptides. In a study, Nell et al. [17] have altered the primary sequence of LL37, a human AMP, by replacing neutral amino acid (Asn and Gln) with two positively charged residues (Arg). The modified peptide was less toxic compared to parent peptide [17].
2. *Modification by inserting* d*-amino acids*
   Natural peptides and proteins are composed of l-amino acids. Therefore, replacing l-amino acids with their enantiomers, i.e., d-amino acids, could be a promising strategy to improve the therapeutic index of peptides. Recently, Albada et al. have shown that systemic l-to-d exchange of amino acid could decrease the hemolytic potency of peptides without compromising their therapeutic activity [18].
3. *Retro-inverso peptides*
   Similarly, peptides consisting of d-amino acids can be synthesized in reverse sequence order, which results in a similar side chain topology to the parent peptide. By this way, properties of peptides can also be altered and may be useful for reducing toxicity.
4. *Modification by inserting modified amino acids*
   Compared to the 20 natural (proteinogenic) amino acids, unnatural/modified amino acids are not encoded by the Universal Genetic Code. The physicochemical properties of native peptides can be improved significantly by the substitution of natural amino acids with unnatural ones.
5. *Cyclization*
   Cyclization of peptide is an important modification. Cyclic peptides mimic natural peptide structures but are constrained and have less conformational freedom compared to linear peptides. Therefore, properties of cyclic peptides are different from linear peptides and this could reduce the toxicity of peptide [16].
6. *End modifications*
   Ends of peptides are playing an important role in function of many peptides like antimicrobial peptides and therefore, modifications at ends could affect the properties of these peptides. Most common end modifications are C-amidation, and N-acetylation, which has great potential to improve the therapeutic index of peptides.

7. *Lipidation*
   Lipidation is an important modification, which can be done to increase the hydrophobicity of peptide. The lipid is attached either directly to amino acid or through a linker. In the past, various lipid moieties that could advance cellular association with the hydrophobic cell membranes have been incorporated onto various peptides in order to improve the therapeutic index.
8. *PEGylation*
   PEGylation of peptides can elicit steric hindrance to reduce interactions with the cell membrane [19] and thus can be used to reduce the cytotoxicity of peptides. In addition, highly hydrophilic PEG spacers may improve the solubility of some peptides and thus are of low toxic. In a study by Fox et al., they have shown that PEGylation of antimicrobial peptide (AMP) CaLL significantly reduced the hemolytic potency without compromising the antibacterial activity of peptide [20]. Similarly in another study, Morris et al. have demonstrated that PEGylation of AMP minimizes lung tissue toxicity while maintaining antimicrobial activity [21].

4 Computational Resources

Toxins have been studied broadly in the past decade and a lot of knowledge is available in the literature. This information has been stored, annotated, and made easy to retrieve by several databases to serve the scientific community. The following databases compile data, present the information which user can access, and provide a broader view of all possible information available for the toxins.

1. *ATDB* (Animal Toxin Database): It is a metadatabase integrating the information from all other databases. The database [22] contains more than 3,235 animal toxins and also deals with toxin ontology to standardize the toxin annotations. All the information is available at http://protchem.hunnu.edu.cn/toxin.
2. *VFDB* (Virulence factor database): The database is a collection of virulence factors of various medically significant pathogens. The database [23] provides in-depth knowledge of mechanisms used by pathogenic bacteria in the bacterial diseases. The updated version provides the pathogenomic composition in terms of virulence. The new version is available at http://www.mgc.ac.cn/VFs/.
3. *DBETH* (Database of Bacterial Exotoxin for Human): It is a database [24] of sequences, structures, and interaction networks of 229 exotoxins from 26 different human pathogenic bacterial genera available at the link http://www.hpppi.iicb.res.in/

btox/. This database provides user-friendly platform to perform several sequence and structure-based analyses of bacterial exotoxins.

4. *ArachnoServer*: It is a manually curated database providing information on the sequence, structure, and biological activity of protein toxins from spider venom. The database provides information about mature peptides. The user can access database at www.arachnoserver.org.
5. *ConoServer*: It is a database [25], specializing in the sequence and structures of conopeptides, which are peptides expressed by carnivorous marine cone snail. The database provides detailed description of cone snail species and their biological activity. The updated version includes three-dimensional structures, endoplasmic reticulum signal sequence conservation trends, and transcriptomics and proteomics data. The link is available at http://www.conoserver.org/.
6. *UnitProtKB*: (www.uniprot.org) It is a general protein sequence database [26], which is broadly divided into two sections
   (a) UnitProtKB/TrEMBL: Here the sequences are automatically annotated; the sequences from a single organism, which show 100 % identity over the entire length, are taken as a single entry.
   (b) UnitProtKB/Swissprot: In this section the information is manually curated and documented and the single entry contains isoforms, fragments of the protein sequence.

In these databases search tools are provided to retrieve the toxins of interest. Both experimentally validated and predicted proteins are stored in the database.

5 Webservers

For initial screening of the peptides for their toxicity values and designing of therapeutic peptides, a number of user-friendly webservers are available; here we are describing each of them.

1. *ClanTox*: This is a classifier for small animal toxins available at the link (http://www.clantox.cs.huji.ac.il/). The machine learning is done on the data manually curated for ion channel toxin inhibitor [27]. The classifier is also based on the frequency and distribution of cysteine residues within the primary sequences, which are important structural factors for toxin stability. Additional information for the protein is also provided like presence of signal peptide, number of cysteine residues, and associated functional annotations.

2. *BTXpred*: It is an in silico method for predicting bacterial toxins whether it is an endotoxin or exotoxin [28]. Support vector machine-based models were used for predicting bacterial toxins and further discriminating between endotoxin and exotoxin. The webserver also provides additional module for classifying exotoxins using hidden Markov models (HMM), PSI-BLAST, and combination of two approaches. User-friendly webserver is available at http://www.imtech.res.in/raghava/btxpred/.
3. *NTXpred*: It is a method developed for predicting neurotoxins and classification based on their function and origin [29]. The SVM model developed in prediction is based on amino acid- and dipeptide-based composition. Another approach used in this classification is a combination of PSI-BLAST and SVM module achieving overall accuracy of 95.11 %. A user-friendly webserver is available at www.imtech.res.in/raghava/ntxpred/.
4. *VICMpred*: It is an SVM-based method [30] for predicting gram-negative bacterial toxins using amino acid-based patterns and composition. The classification was also based on the tetrapeptides as the input features for predicting the function of a protein. The webserver is available at http://www.imtech.res.in/raghava/vicmpred/.
5. *ToxinPred*: It is an in silico method [31] for predicting toxicity of peptides. The user can design and predict peptides with desired toxicity values. A user-friendly webserver is available at http://crdd.osdd.net/raghava/toxinpred/ with different types of modules like designing of the peptides with all possible single mutation and predicting the toxicity of the mutants generated. Further user can also scan whole protein for toxic region and find out least as well as the most toxic region (Fig. 1).
6. *DBETH server*: It is a part of DBETH database [24], where identification is based upon establishing homology with known toxin sequences/domains or by using machine learning techniques to classify human pathogenic exotoxin.

6 In Silico Models for Toxicity Prediction

Toxicity is a major hurdle in the process of developing drugs from lead molecules. Unless a peptide is nontoxic, it cannot be used in clinical therapeutic applications. Therefore, testing toxicity of any therapeutic peptide lead molecule is an essential part of peptide-based drug discovery. However, it is a very labor-intensive and expensive task. In order to overcome these limitations, computational approaches, which can predict toxicity of any peptide/protein at very early stages of drug development, are highly demanding. To reduce or remove toxicity of peptides, various efforts have been

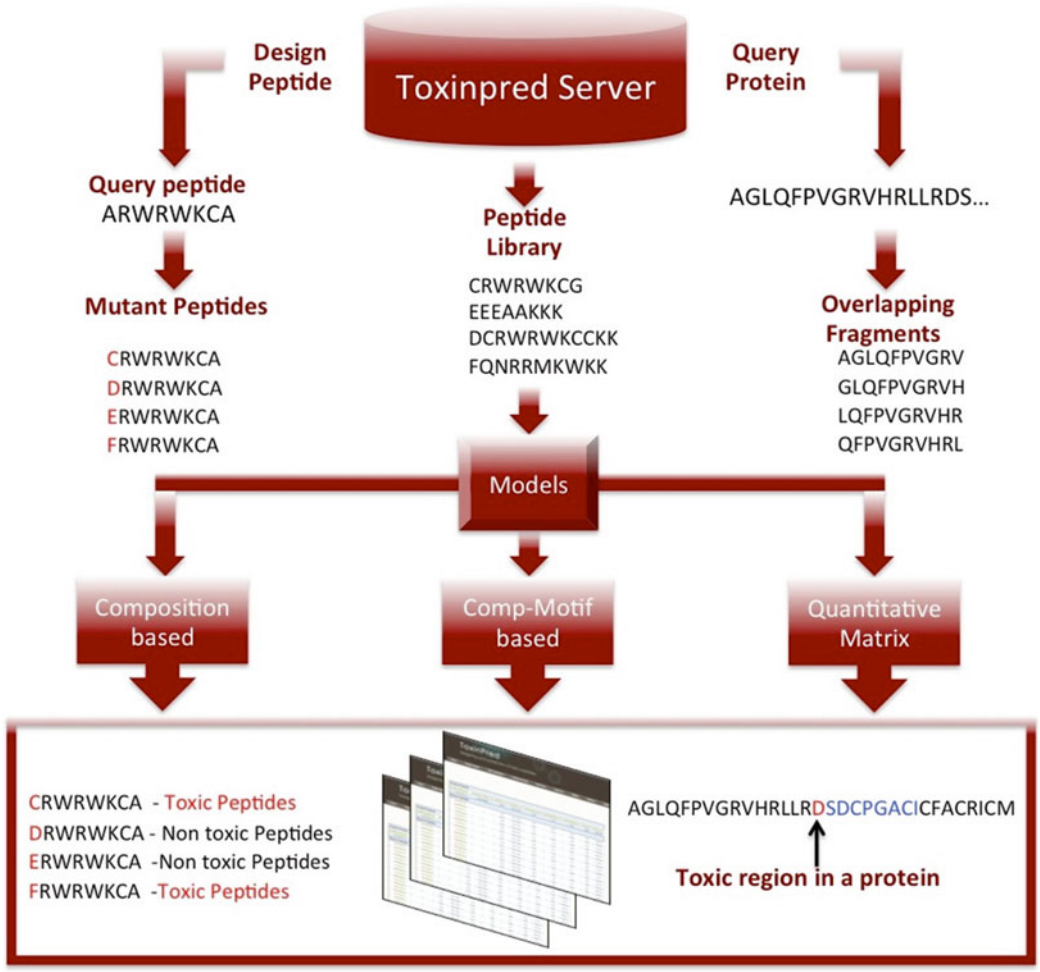


Fig. 1 Architecture of ToxinPred webserver

made, which incorporate in silico approaches for the prediction of toxins. Among several such advances (as mentioned in the computational resources section), ToxinPred is a recent method, which provides a series of tools for analysis and prediction of toxicity in peptides. This method presents a range of different amino acid sequence properties and their use in machine learning method like SVM. The highly accurate in silico models of this study have been integrated in different modules of ToxinPred webserver. The modules of webserver and their utility have been discussed in the following section:

6.1 Designing of Toxic Peptides

The SVM models based on different amino acid sequence features like amino acid composition, dipeptide composition, presence of motifs, and models based on quantitative matrix (QM) were used in this module to provide a platform for analysis and designing of

peptide with desired toxicity (Fig. 2). The virtual screening of all possible single-residue mutants, as given by this module, enables the user to look at the peptide sequences with a specific mutation made and alteration in its toxicity. The prediction results are also supplemented with different physiochemical properties of peptides such as hydrophobicity, steric hindrance, side bulk, hydropathicity, amphipathicity, hydrophilicity, and net hydrogen atom molecules and charge. This comprehensive analysis and prediction of a peptide and its possible single-residue mutants help in deciding toxicity, at a single-residue change, which in turn provides options to engineer a peptide for decreasing or increasing toxicity as required.

As an improvement to the single-peptide submission in the above module, the batch submission module was designed. This module has the ability to submit more than one peptide sequence. The result comes in two steps where the first step involves the classification of peptides in toxic and nontoxic classes. In the next step, the link of peptide leads to virtual screening of different single-residue mutants for that particular peptide.

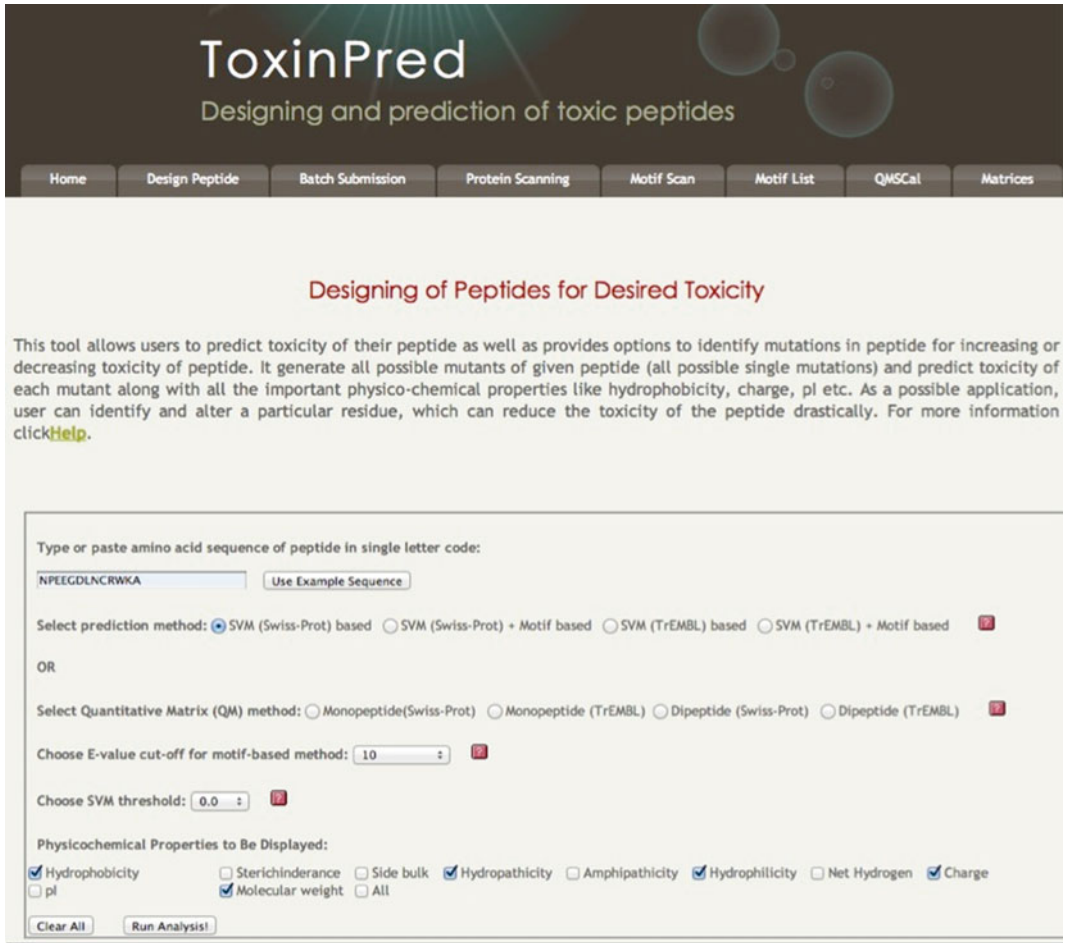
6.2 Protein Scanning

Since the in silico models are based on a dataset of 4–35 length of peptides, it was assumed that the models could be applied on proteins by scanning it in a sliding window of desired size. The protein-scanning module of ToxinPred webserver has been designed to make fragments of protein sequence by sliding window of desired size and predict its toxicity at the same time. The module results in identification of toxic regions in a protein and thus helps in editing or removing the region for lesser toxicity in the protein. Similar to the above-mentioned module of designing peptide, here also, the user can select different in silico models such composition-based, motif-based or quantitative matrix-based models. The result tables further lead to single-residue mutation table where toxicity is predicted for every possible single-residue change. The toxic region in a protein can be seen in the main result table as a colored defined region.

6.3 Motif Scanning

Toxic peptides or proteins have unique motifs, which discriminate them from the rest of the sequences. So we have used MEME/MAST suite [32] for mining exclusive motifs in toxic peptides and have given provision to scan the same in users' query protein/peptide sequence. This algorithm is widely used for motif discovery and is based upon multiple Em theory (Fig. 3).

First, motifs have been discovered using MEME in toxic peptide dataset. Then using MAST, sequence dataset with motifs has been searched. Different e-values have been selected to search motifs. A lower e-value leads to furthermore stringent motif search. So to get motif sequences with high confidence e-value must be set at a lower value.



Original Peptide

Peptide Sequence	Mutation Position	SVM score	Prediction	Hydrophobicity	Hydropathicity	Hydrophilicity	Charge	Mol wt
NPEEGDLNCRWKA	No Mutation	-0.09	Non-Toxin	-0.37	-1.59	0.67	-1.00	1531.84

Mutant Peptides

APEEGDLNCRWKA	1	-0.38	Non-Toxin	-0.30	-1.18	0.62	-1.00	1488.81
CPEEGDLNCRWKA	1	-0.36	Non-Toxin	-0.32	-1.13	0.58	-1.00	1520.87
DPEEGDLNCRWKA	1	-0.20	Non-Toxin	-0.38	-1.59	0.88	-2.00	1532.82
EPEEGDLNCRWKA	1	-0.38	Non-Toxin	-0.37	-1.59	0.88	-2.00	1546.85
FPEEGDLNCRWKA	1	-0.22	Non-Toxin	-0.27	-1.11	0.46	-1.00	1564.91
GPEEGDLNCRWKA	1	-0.30	Non-Toxin	-0.31	-1.35	0.65	-1.00	1474.79
HPEEGDLNCRWKA	1	0.06	Toxin	-0.35	-1.57	0.62	-0.50	1554.88
IPEEGDLNCRWKA	1	-0.30	Non-Toxin	-0.27	-0.98	0.52	-1.00	1530.90
KPEEGDLNCRWKA	1	-0.36	Non-Toxin	-0.41	-1.62	0.88	0.00	1545.91
LPEEGDLNCRWKA	1	-0.39	Non-Toxin	-0.28	-1.03	0.52	-1.00	1530.90
MPEEGDLNCRWKA	1	-0.68	Non-Toxin	-0.30	-1.18	0.55	-1.00	1548.93
PPEEGDLNCRWKA	1	-0.19	Non-Toxin	-0.33	-1.45	0.65	-1.00	1514.85
QPEEGDLNCRWKA	1	-0.57	Non-Toxin	-0.37	-1.59	0.67	-1.00	1545.87
RPEEGDLNCRWKA	1	-0.25	Non-Toxin	-0.46	-1.67	0.88	0.00	1573.92
SPEEGDLNCRWKA	1	-0.31	Non-Toxin	-0.34	-1.38	0.68	-1.00	1504.81
TPEEGDLNCRWKA	1	-0.33	Non-Toxin	-0.34	-1.38	0.62	-1.00	1518.84
VPEEGDLNCRWKA	1	-0.41	Non-Toxin	-0.28	-1.00	0.54	-1.00	1516.87

Fig. 2 Peptide designing module of ToxinPred

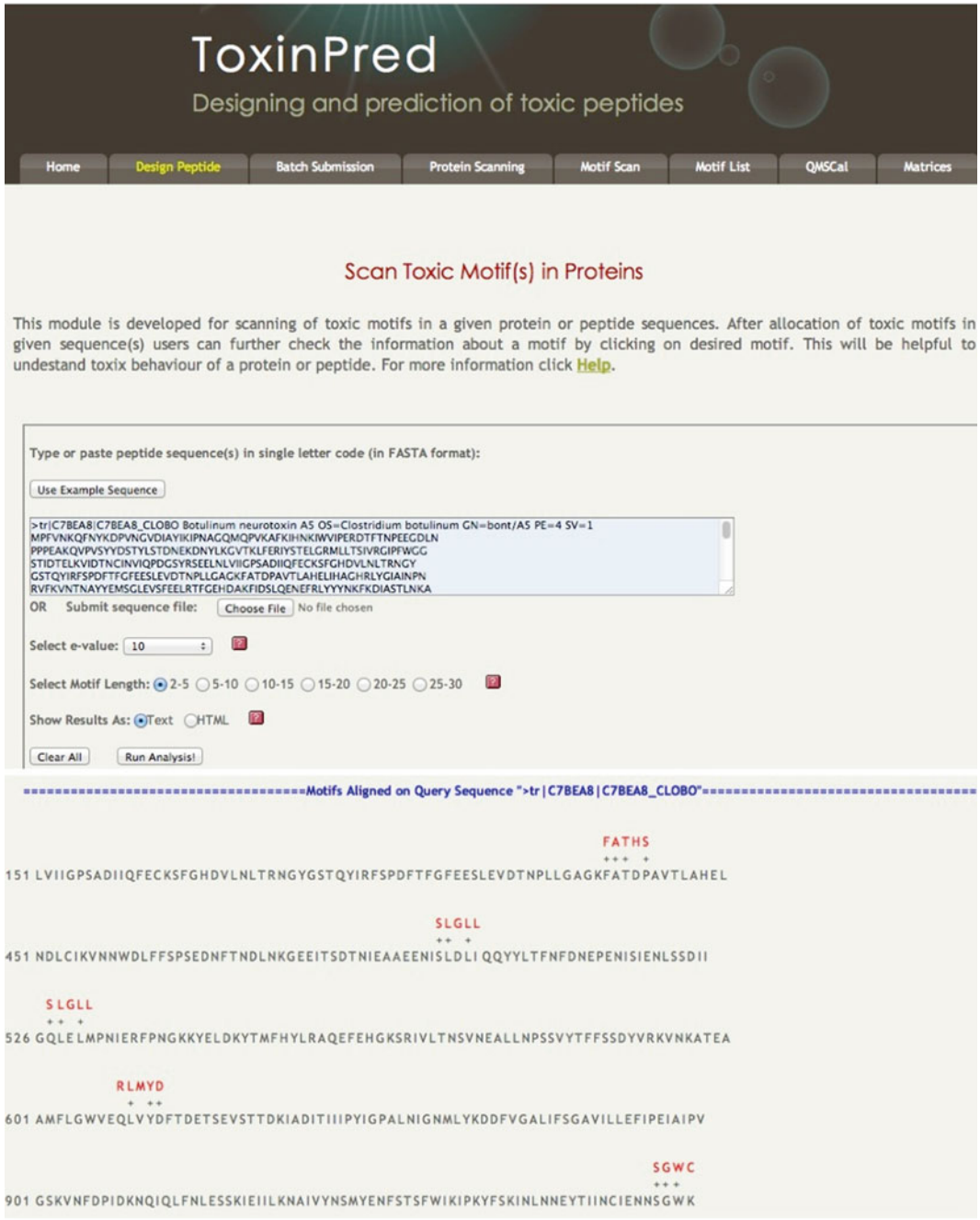


Fig. 3 Motif scan module of ToxinPred

Balance between coverage and probability of correct prediction is necessary because a biologist may be interested in those motifs only, which give 100 % confidence. But in that case, coverage of dataset decreases drastically, so by making balance, motif information can be incorporated with support vector machine (SVM)-based

Motif	No. of sites	E value	Motif	No. of sites	E value
RLMYD	40	1.10e-252	AAKVK	40	3.30e-138
NNPHV	40	8.80e-196	THPGG	40	2.90e-85
AANDK	40	5.50e-138	GAKCSRLMYD	40	3.8e-397
KGKGA	40	7.20e-184	NPACRVNNPH	40	4.9e-362
HPACG	40	4.20e-172	NPPCFANHPE	40	2.20e-241

Fig. 4 Top motifs and their presence in peptides, in ToxinPred

prediction to get higher performance. Here coverage is covered by SVM part and confidence is attributed by motif part.

Also, a motif list has been provided, which is found in our peptide dataset above a given threshold. By looking at the motifs one can have a general idea about the sequence or the peptide for its toxic nature. This list contains motif sequence, length of motif and number of times it has occurred in toxic dataset. Also, the e-value at which this stats have been generated, is provided to get an idea whether motif searching is stringent or not.

Motif-based prediction could be useful in many ways. It simplifies the prediction procedure in a way. If exclusive motif is found then straightway property can be assigned to query sequence. Alternatively, motif information incorporated with SVM score can be useful for threshold-based prediction.

Apart from this, motif scan module helps to scan a protein for the presence of toxic motifs. Here user can select e-value cutoff to make search more/less stringent according to need. Also, length of motif can be selected among the given ranges and output can be viewed in traditional MAST output or simpler text output, which shows the location of motifs found in given query sequences. This tool gives idea, which proteins are abundant in toxic motifs and further can be exploited for toxic peptide generation.

We incorporated predominantly found motifs in toxic peptide dataset for prediction. MEME software was employed to search motifs in this dataset followed by hitting query sequences on these motifs using MAST. Reliability of prediction was increased by incrementing SVM prediction score in those cases where motifs were found in the query sequences (Fig. 4).

6.4 Quantitative Matrix (QM)

QM is a representation of relative propensity of each type of residue at position-specific information in a dataset. QMs have been provided which give the probability of all natural amino acids at every position in toxic peptide dataset. This kind of matrix has direct implication in understating the probability of occurrence of a residue at a given position. Apart from individual amino acid's frequency, frequency of residues having different physicochemical properties has also been given. This shows the prevalence of major property at each position in a given sequence. Further, the QMs have been provided for single amino acid propensity and dipeptide

propensity, which is more relevant to the biologists. Sum of propensity scores of a query sequence gives an idea about the nature of that sequence being toxic or not. Matrices are sortable at every position to get propensity score of minimum and maximum scoring residue.

In addition, a QMSCal tool has been integrated which assists the user to alter a therapeutic peptide with minimum mutation to achieve maximum score based upon QMs. This tool is useful to make minimum residue-level changes to attain desired property of peptide (high/low toxicity) based upon QM-based cumulative score. Highest/lowest propensity values of each position are given along with QM-based cumulative score for a given query. The residues having more propensity value will be contributing more towards the given property (toxicity) and vice versa. So selecting residues favoring a property would make that sequence more effective and QM is a direct implication of this theory. QMSCal can assists the users to mutate minimum residues in therapeutic peptides in such a way that its toxicity reduces drastically without affecting its activity. For any given query in FASTA format, user can tweak residues to get minimum/maximum toxicity based upon QM-based cumulative probability scores for all residues in given query. QMs can also be used for prediction based upon their propensity score. When used alone, then query sequence is hit on this matrix and score is calculated. And output is assigned based upon different thresholds. QMs generally give poor results as compared to SVM. But they can also be incorporated to SVM pipeline to get better performance in a hybrid or cascade pipeline.

Quantitative matrix was generated for each residue on the basis of position-specific contribution of every residue for both datasets. The performance of QM was evaluated by using fivefold cross validation technique.

7 Limitation of Existing Methods

There are a number of webservers available in the literature, which are dedicated for the prediction of toxicity of peptides, e.g., ToxinPred and BTXpred, just to name a few. If we carefully look at the functioning of these prediction methods, they are based on machine learning techniques, e.g., SVM and HMM, and predict whether certain peptide is toxic or not. But the datasets for developing these prediction methods are primarily of bacterial origin. So, these methods will discriminate whether a certain peptide is bacterial toxin or not and not discussing about their toxicity and its impact on the human body. For any therapeutic use of peptide, first it has to be evaluated on the human body in context to toxicity and prediction methods should also be developed on such data.

Prediction methods developed on such data will be more accurate and biologically relevant and they will certainly help toxicologists to screen peptides and assign toxicity level to them, before going any further.

8 Future Prospects

Peptide-based therapies have already made a huge impact in the health and pharmaceutical industry in the area of treatment as well as diagnosis. Toxicological profile of drug candidate is an important objective in the clinical development and initiation plans of human studies. Conventional approaches to the toxicity studies of peptides include cytotoxic studies, hemolytic toxicity, and standard toxicity studies in animals. Single-dose toxicity studies, repeated-dose toxicity studies, immunotoxicity studies, and developmental toxicity studies are done to achieve safety standards of peptides. In the preliminary stage, if efforts are put to minimize the toxic effects of peptide a lot of human efforts and resources can be saved, so in silico method for toxicity prediction is a very rational approach in the field of peptide therapeutics.

As we have already developed in silico method "ToxinPred" useful in predicting toxicity of the peptides/proteins and also in designing of peptides with least toxicity. In the future, in silico models developed for toxicity prediction should be developed with a focus on the therapeutic property prioritization and toxicity profile of the peptide, i.e., positive dataset comprises peptide with therapeutic peptide in subject (e.g., tumor-homing peptide, cell-penetrating peptide, anticancer peptide, anti-hypersensitive peptides) and negative dataset (toxic peptides). This will help in designing peptide with full therapeutic potential and least toxicity. We hope that these in silico-based approaches become so accurate that they are not only used in the early screening of the peptides but also broadly acceptable by regulatory authorities and toxicologists.

References

1. Thundimadathil J (2012) Cancer treatment using peptides: current therapies and future prospects. J Amino Acids 2012:967347
2. Saladin PM, Zhang BD, Reichert JM (2009) Current trends in the clinical development of peptide therapeutics. IDrugs 12:779–784
3. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M (2010) Synthetic therapeutic peptides: science and market. Drug Discov Today 15:40–56
4. Craik DJ, Fairlie DP, Liras S, Price D (2013) The future of peptide-based drugs. Chem Biol Drug Des 81:136–147
5. Descotes J (2006) Methods of evaluating immunotoxicity. Expert Opin Drug Metab Toxicol 2:249–259
6. Dhanda SK, Gupta S, Vir P, Raghava GPS (2013) Prediction of IL4 inducing peptides. Clin Dev Immunol. doi:10.1155/2013/263952
7. Nielsen M, Lund O, Buus S, Lundegaard C (2010) MHC class II epitope predictive algorithms. Immunology 130:319–328
8. Singh H, Raghava GP (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17:1236–1237

9. Gupta S, Ansari HR, Gautam A, Raghava GP (2013) Identification of B-cell epitopes in an antigen for inducing specific class of antibodies. Biol Direct 8:27
10. Dhanda SK, Vir P, Raghava GPS (2013) Designing of interferon-gamma inducing MHC class-II binders. Biol Direct 8:30
11. Goodwin D, Simerska P, Toth I (2012) Peptides as therapeutics with enhanced bioactivity. Curr Med Chem 19:4451–4461
12. Uchide N, Ohyama K, Bessho T, Toyoda H (2009) Lactate dehydrogenase leakage as a marker for apoptotic cell degradation induced by influenza virus infection in human fetal membrane cells. Intervirology 52:164–173
13. Fotakis G, Timbrell JA (2006) In vitro cytotoxicity assays: comparison of LDH, neutral red, MTT and protein assay in hepatoma cell lines following exposure to cadmium chloride. Toxicol Lett 160:171–177
14. Cree IA, Andreotti PE (1997) Measurement of cytotoxicity by ATP-based luminescence assay in primary cell cultures and cell lines. Toxicol In Vitro 11:553–556
15. Kalcheim C, Goldstein RS (1991) Segmentation of sensory and sympathetic ganglia: interactions between neural crest and somite cells. J Physiol Paris 85:110–116
16. Kondejewski LH, Jelokhani-Niaraki M, Farmer SW, Lix B, Kay CM, Sykes BD, Hancock RE, Hodges RS (1999) Dissociation of antimicrobial and hemolytic activities in cyclic peptide diastereomers by systematic alterations in amphipathicity. J Biol Chem 274:13181–13192
17. Nell MJ, Tjabringa GS, Wafelman AR, Verrijk R, Hiemstra PS, Drijfhout JW, Grote JJ (2006) Development of novel LL-37 derived antimicrobial peptides with LPS and LTA neutralizing and antimicrobial activities for therapeutic application. Peptides 27:649–660
18. Albada HB, Prochnow P, Bobersky S, Langklotz S, Bandow JE, Metzler-Nolte N (2013) Short antibacterial peptides with significantly reduced hemolytic activity can be identified by a systematic L-to-D exchange scan of their amino acid residues. ACS Comb Sci 15:585–592
19. Veronese FM, Harris JM (2002) Introduction and overview of peptide and protein pegylation. Adv Drug Deliv Rev 54:453–456
20. Fox MA, Thwaite JE, Ulaeto DO, Atkins TP, Atkins HS (2012) Design and characterization of novel hybrid antimicrobial peptides based on cecropin A, LL-37 and magainin II. Peptides 33:197–205
21. Morris CJ, Beck K, Fox MA, Ulaeto D, Clark GC, Gumbleton M (2012) Pegylation of antimicrobial peptides maintains the active peptide conformation, model membrane interactions, and antimicrobial activity while improving lung tissue biocompatibility following airway delivery. Antimicrob Agents Chemother 56:3298–3308
22. He Q, Han W, He Q, Huo L, Zhang J, Lin Y, Chen P, Liang S (2010) ATDB 2.0: a database integrated toxin-ion channel interaction data. Toxicon. doi:10.1016/j.toxicon.2010.05.013
23. Chen L, Xiong Z, Sun L, Yang J, Jin Q (2012) VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res 40:D641–D645
24. Chakraborty A, Ghosh S, Chowdhary G, Maulik U, Chakrabarti S (2012) DBETH: a database of bacterial exotoxins for human. Nucleic Acids Res 40:D615–D620
25. Kaas Q, Yu R, Jin A-H, Dutertre S, Craik DJ (2012) ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. Nucleic Acids Res 40:D325–D330
26. Apweiler R, Bairoch A, Wu CH et al (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32:D115–D119
27. Naamati G, Askenazi M, Linial M (2009) ClanTox: a classifier of short animal toxins. Nucleic Acids Res 37:W363–W368
28. Saha S, Raghava GPS (2007) BTXpred: prediction of bacterial toxins. In Silico Biol 7:405–412
29. Saha S, Raghava GPS (2007) Prediction of neurotoxins based on their function and source. In Silico Biol 7:369–387
30. Saha S, Raghava GPS (2006) VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. Genomics Proteomics Bioinformatics 4:42–47
31. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GPS (2013) In silico approach for predicting toxicity of peptides and proteins. PLoS One 8:e73957
32. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–W208

Chapter 8

Synthetic and Structural Routes for the Rational Conversion of Peptides into Small Molecules

Pasqualina Liana Scognamiglio, Giancarlo Morelli, and Daniela Marasco

Abstract

The demand for modified peptides with improved stability profiles and pharmacokinetic properties is driving extensive research effort in this field. The conversion of peptides into organic molecules, as traditional drugs, is a long and puzzled way. Many and versatile approaches have been described for designing peptide mimetics: the substitution of natural residues with modified amino acids and the rigidification and modification of the backbone are the main structural and chemical routes walked in medicinal chemistry. All of these strategies have been successfully applied to obtain active new compounds in molecular biology, drug discovery and design. Here we propose a panoramic review of the most common methods for the preparation of modified peptides and the most interesting findings of the last decade.

Key words Peptidomimetics, Backbone modification, Conformational constraints

1 Introduction

Peptides show great pharmaceutical means as drugs and diagnostics in several clinical fields such as neurology, oncology, endocrinology, immunology, urology, and obstetrics and as functional excipients in drug delivery methods to diffuse across tissue and cellular membrane barriers [1, 2]. From a chemical point of view peptides are situated borderline between classical organic drug substances and high-molecular-weight biopharmaceuticals.

But, despite crucial in vitro advantages of peptides, there are few natural sequences commercialized as pharmaceutical products. The limitations of native peptides as drugs rely on their poor metabolic stability, low membrane permeability, limited oral bioavailability, rapid clearance, and low selectivity [3].

Nowadays bioavailability issues are often addressed by novel routes of administration (e.g., intranasal, inhalation, iontophoresis) and injectable depot formulations [1], and, furthermore, modern medicinal chemistry offers powerful synthetic and structural tools to overcome several limitations. Peptidomimetics are small

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_8,

protein-like molecules designed to mimic natural peptides or proteins in order to retain similar biological effects but with enhanced proteolytic stability, higher bioavailability, and improved selectivity and/or potency. On these basis they can be considered as good lead compounds in the discovery processes of new drugs [4].

Actually the majority of marketed peptide products are hormones or peptide derivatives that simulate the action of hormones; but, on the other hand, agonists or antagonists for receptors implicated in oncology and inflammation, or inhibitors of enzyme, are increasingly in preclinical stages, suggesting that this class of drugs might soon occupy a larger segment in the pharmaceutical market. Also antimicrobial peptides (AMP), with broader spectrum activity, promised to have a great future, especially in counteracting the loss of efficiency of conventional antibiotics [5]. The development of peptidomimetics includes the design of new molecules through the incorporation of nonnatural and/or conformationally constrained amino acids, the replacement of the peptide bond isosteres, and the introduction of further structural restraints such as cycles.

The molecular basis of the bioactivity of most peptides relies on the recognition by an interface ligand region toward the complementary surface of the receptor or of a protein complex. On the ligand side, the region involved in the interaction is generally defined pharmacophore and it can involve both a continuous stretch of amino acids that several residues separated within the primary sequence but close in space [6]. In both cases, the peptide backbone creates favorable hydrogen bonds even if it acts as the scaffold to support functional groups of the side chains directly involved in the recognition process. Two structural requirements are crucial for the design of peptidomimetics: the new molecule has to conveniently fit into the binding site and its functional moieties (polar and hydrophobic groups) need to be placed in defined positions to allow useful interactions to take place [7].

Since the properties of peptides are determined by the nature of the constituent amino acids and several of them can undergo posttranslational modifications, unnatural amino acids (i.e., those not genetically coded that naturally occur, but also other synthetically produced) are important tools for modern drug discovery research. Due to their chemical diversity and functional versatility, they are widely used as chiral building blocks and molecular scaffolds in constructing combinatorial libraries and in modifying already exiting sequences [8, 9].

The major limitation in peptide research is the conformational flexibility of most natural peptides and its high dependence from the environment: small peptides typically show high conformational flexibility due to the multiple conformations that are energetically possible for each residue [10].

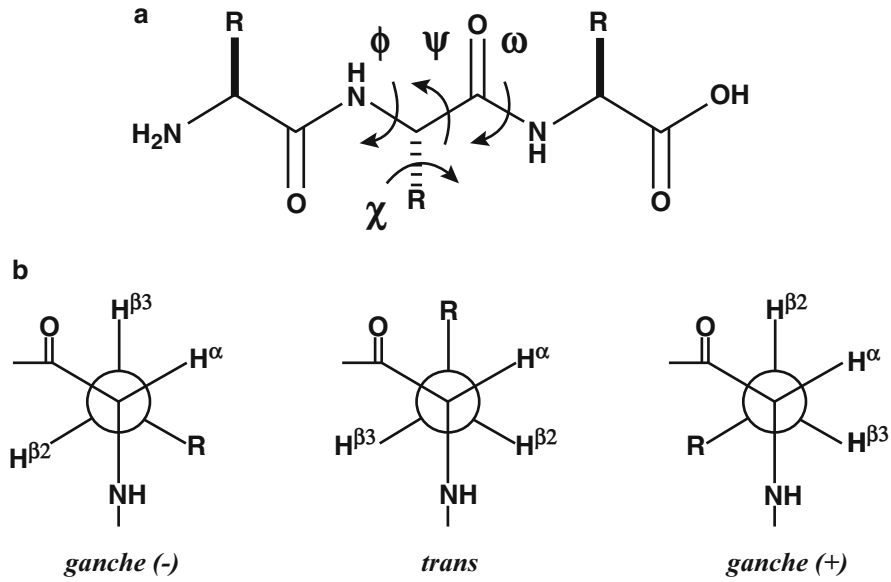


Fig. 1 (**a**) Schematic description of torsional angles in a peptide bond: φ, ψ, and ω. (**b**) Newman projections about the Cα–Cβ bond indicating the rotational isomers of the L-amino acids

The conformation of the peptide backbone can be described by three torsional angles as illustrated in the Fig. 1: φ describes rotation about the N–C(α) bond and involves the C(O)–N–Cα–C(O) bonds, ψ describes rotation about the C(α)–C(O) bond and involves the N–Cα–C(O)–N bonds, while ω describes rotation about the C(O)–N bond and involves the Cα–C(O)–N–Cα bonds. The ω angle for the peptide bond is generally *trans* ($\omega = 180°$) except for the Xaa-Pro bond, which can be *cis* ($\omega = 0°$) or *trans*, and generally it differs from these planar conformations of less than ±20°. The evaluation of the low-energy conformations of the angles φ and ψ was examined for the first time 40 years ago by Ramachandran et al. [11] and it resulted in the so-called Ramachandran plots. The conformational space accessible to the L-amino acids is about one-third of the total structural space. These regions correspond to the classical secondary structures of peptides and proteins (as α-helix and β-sheet), but also other low-energy conformations of the backbone were highlighted as β-turns and γ-turns. Nevertheless the remaining degrees of freedom still make a prediction of structure extremely difficult. There are only few examples reported in the literature, where short- to medium-sized peptides (<30–50 residues) adopt a stable structure in aqueous solution [12]; in most cases they have numerous dynamically interconverting conformations. During the design of bioactive peptides the three-dimensional structures of the side-chain moieties should be evaluated. The side-chain χ torsional angle (χ^1 is defined by the N–Cα–Cβ–Cγ) can assume three low-energy

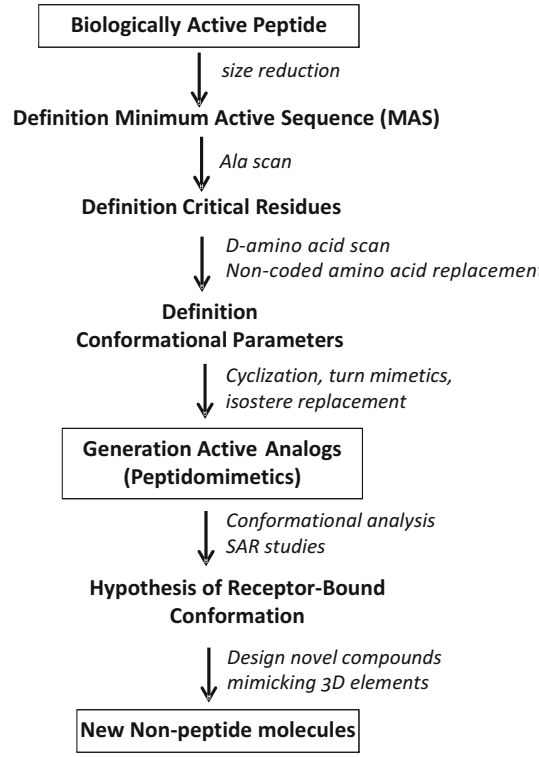
staggered conformations (rotamers): *gauche* (+), *gauche* (−), and *trans*. Though energy differences among these conformations are not so high, the orientation of the side-chain group of a L-amino acid residue relative to the peptide backbone is dramatically different: for *gauche* (−), the side chain points toward the N-terminus of the peptide chain; for *trans*, it points toward the C-terminus; and for *gauche* (+), the side chain points over the peptide backbone [13]. The consequences for both the surface of the peptide ligand and its complementarity for a receptor/acceptor are critical for successful recruitment of the target from small peptides as confirmed by structure–activity relationship studies [13]. The design of peptidomimetics and peptide models is based on the knowledge of the secondary structure elements of peptides. Based on these values it is possible to predict in the first approximation the effects of a replacement of natural amino acid by its conformationally restricted analogues [14].

2 Approaches for the Design of Peptide-Like Drugs

The principal aim in peptidomimetics design is to replace the peptide backbone as much as possible with non-peptide fragments while still maintaining the pharmacophoric groups (deriving from amino acid side chains) of the peptide. Conformational requirements represent principal guidelines in the design of new biologically active compounds [15]. Usually peptides are an ensemble of conformational states in solution; thus, if biological activity involves only one discrete conformer, this conformational ensemble represents a dilution of the biologically active species.

For the development of potent peptidomimetics, it is necessary to know the forces that govern peptide–protein interactions. These interactions are mainly based on side chain indicating that the peptide backbone itself is not an absolute requirement for high affinities [16].

The conversion of peptides into small molecule is a long and puzzled pathway and there is no guarantee of success, although several peptides have been converted successfully into non-peptide drug candidates [17]. It starts from the reduction of the peptide to the minimum active sequence, testing truncated sequences from the C- and N-termini alternatively. Then the contribution of a specific residue to the structure, stability, and biological activity is determined by systematically replacing each residue in the sequence with specific amino acids, typically alanine (a non-bulky, chemically inert residue) or D-isomer amino acids; these latter are able to be more resistant to proteolysis [5]. After the assessment of structure–activity relationship (SAR) of each residue, the bioactive conformational flexibility is reduced by introducing constraints, such as rings, into the linear peptide to force it to adopt a rigid and biologically active conformation. These features are used for the



Scheme 1 Strategies for the conversion of peptides into small molecules

design of a pharmacophore model in which functional groups crucial for activity are prepositioned. All phases of the process are summarized in Scheme 1.

3 Unnatural Amino Acid

A smart approach to introduce novel chemical groups in peptides to improve their pharmacokinetic and pharmacodynamic properties is the insertion of unnatural amino acids that represent a source of chemical and structural diversity still poorly explored and easily commercially available. Peptidomimetics, designed on the basis of conformationally restricted unnatural amino acids, are expected to be biologically active due to their better preorganization resulting in stronger interaction with target biomolecules [18]. This strategy revealed to be successful, since many studies reported on metabolically stable peptidomimetics [19–21].

Two different starting points exist for the modification of peptides at the amino acid level: in the first the amino acid side chain is usually rigidified by the use of sterically demanding groups; in the latter the backbone of the peptide is modified.

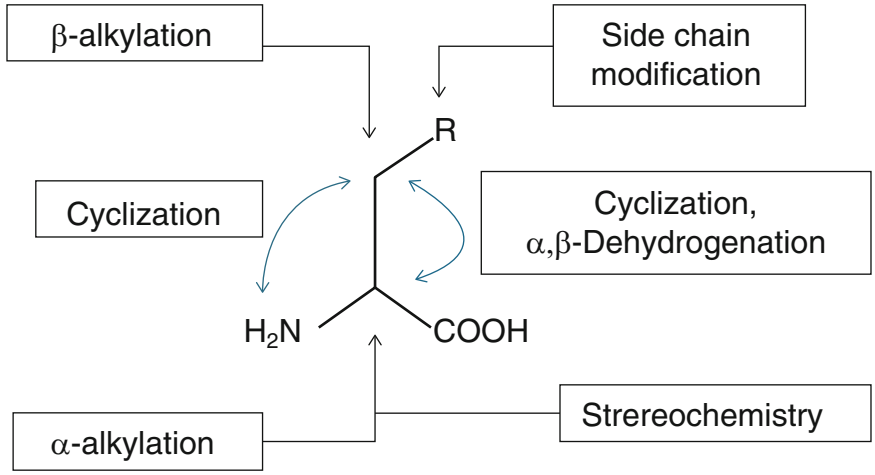


Fig. 2 Schematic description of potential amino acid sites for chemical modification

3.1 Side-Chain Modification

This approach employs amino acids with sterically constrained side chains or a stereochemistry of type D at the level of the α carbon atom. In addition, the modifications can be achieved through a α- or β-alkylation of natural amino acids and/or cyclization or through the introduction of an olefinic bond between α and β carbon atoms. Sites of potential modifications of amino acids are shown in Fig. 2.

3.1.1 α-Methyl Amino Acids

Optically active α-methylated amino acid analogues (where Hα is replaced by methyl and R is a residue corresponding to a natural or unnatural amino acid), particularly those having the L enantiomeric configuration, are known to be useful for a variety of purposes. Some α-methylated amino acids are usefully inserted in therapeutic agents, such as methyldopa (L-α-methyl-3,4-dihydroxyphenylalanine) (Fig. 3a); it is an alpha-adrenergic agonist psychoactive drug, used as a sympatholytic or antihypertensive [22]. Other α-methylated amino acids have been used as intermediates for the synthesis of hydantoin analogues that resulted active for treating or preventing inflammatory and immune cell-mediated diseases [23].

Methylation severely restricts the rotation around the N–Cα, Cα–C(O) and χ angle of amino acids (Fig. 1). Among the α-methylated amino acids, Aib (α-aminoisobutyryl acid) (Fig. 3b) is frequently introduced in peptides, reducing the allowed areas of typical values of α-helix and β-turn conformations in the Ramachandran's plot. When located in an internal position of the peptide chain, Aib induces a strong stabilization of helical secondary structures. In a study reported by Gobbo et al. [24], a series of longer analogues of the C-terminal region of RNAse A has been analyzed to assess the helix induction potential in water of α-methyl amino acids at the N-terminus of the chain. The circular dichroism data indicated that also isovaline residue (Fig. 3c) is effective in increasing the helix content of the 13-residue peptide by about 7 %.

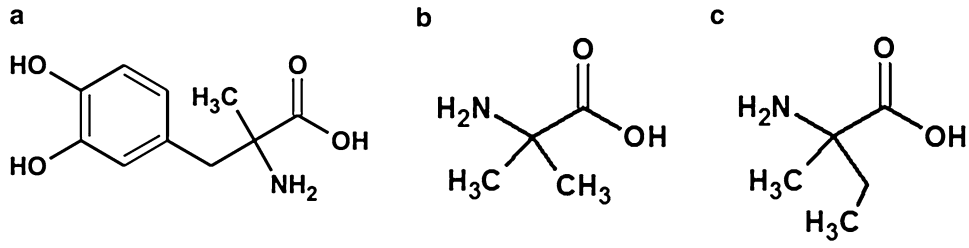


Fig. 3 Chemical structure of (**a**) methyldopa (L-α-methyl-3,4-dihydroxyphenylalanine); (**b**) α-aminoisobutyryl acid; (**c**) isovaline

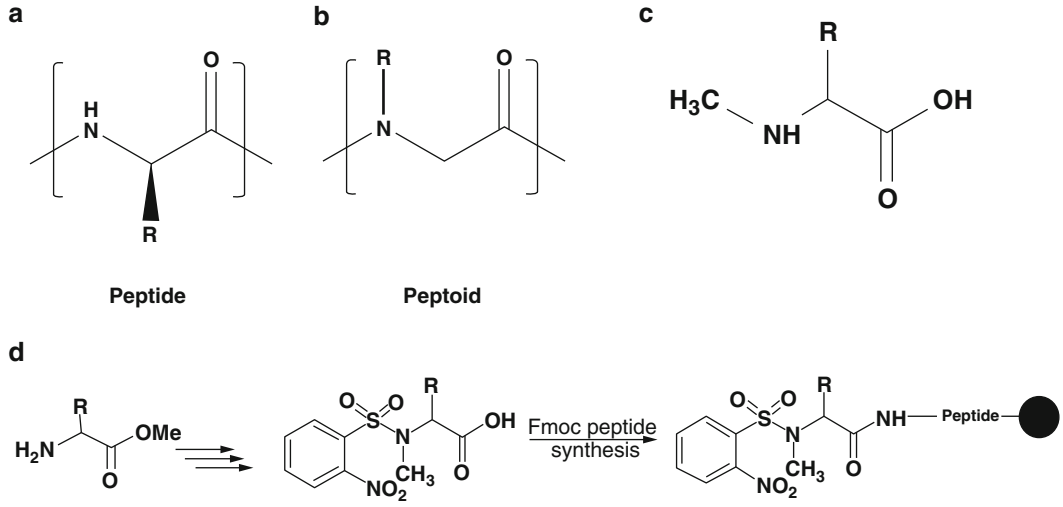


Fig. 4 Diagram of backbone of (**a**) peptide and (**b**) peptoid; (**c**) chemical structure of *N*α-methyl amino acid; (**d**) scheme of synthetic procedure for N_{α}-methylamino described by Kessler's group [29]

A recent study reported on rationally designed conformationally constrained analogues of cyclopentapeptide CXCR4 antagonists [25]. In a successful attempt to downsize the 14-mer peptide antagonist T140, Fujii's group discovered the potent cyclopentapeptide antagonist [cyclo(-Arg-Arg-2-Nal-Gly-D-Tyr-)] [26], where a further modification employed both backbone stabilization and side-chain rigidification strategies. Importantly, the introduction of an α,α-disubstituted amino acid in position 1 was beneficial for activity, supporting the predicted requirements of backbone conformation for this class of compounds to show bioactivity [27].

3.1.2 Nα-Substituted Glycines

Sequences having N-substituted glycines in which side chains are appended to the nitrogen atom of the backbone, rather than to the α-carbons (as they are in natural amino acids), are called α-peptoids [28] and are represented in Fig. 4a, b.

The conformational change in the N-substituted glycines makes the α-carbon achiral so that peptoids are less restricted in their spatial conformations. Notably, peptoids lack the amide hydrogen

which is responsible for many secondary structure elements in peptides and proteins. Like D-peptides and β peptides, peptoids are completely resistant to proteolysis [30] and are therefore advantageous for therapeutic applications, where proteolysis is a major issue. Since secondary structure in peptoids does not involve hydrogen bonding, it is not typically denatured by solvent, temperature, or chemical agents [31].

Peptoids with alpha-chiral bulky side chains are known to adopt a polyproline-type I-like conformation. Notably, since the amino portion of each residue can result from the use of any amine, thousands of commercially available amines can be used to generate unprecedented chemical diversity at costs far lower than those required for similar building blocks modified at different sites [32].

3.1.3 N^{α}-methyl Amino Acids

N-methyl amino acids (Fig. 4c) can improve the pharmacokinetic properties of bioactive peptides where they replace natural residues; indeed the introduction of *N*-methyl amino acids generally increases the enzymatic stability of peptides, thus enhancing their in vivo half-life.

Recent reports show that peptides rich in *N*-methyl phenylalanine passively diffuse across the blood-brain barrier and can be used as blood-brain barrier shuttles [33]. But *N*-methyl amino acids can enhance activity and selectivity or convert an agonist to an antagonist [34]. These changes are attributed to reduced backbone flexibility resulting from *N*-methyl groups, but also on the side chain of the neighboring amino acid [35]. The *N*-methyl groups also reduce the number of inter- and intramolecular hydrogen bonds, but it is able to affect water solubility as Aβ-derived analogues [36]. Kessler's group reported on cyclic *N*-methylated somatostatin analogues related to the Veber–Hirschmann peptide, generating a library of 30 compounds with varying degrees of methylation of the secondary amides contained in the starting macrocycle. Extensive in vitro experiments showed that specific methylation of D-Trp8, Lys9, and Phe11 gave rise to a large enhancement in membrane permeability and reasonable oral bioavailability in rat [37]. It becomes evident that multiple N-methylation is a novel technology to achieve oral bioavailability but also to improve the pharmacological properties of peptides. For example, a number of cyclic α-MSH analogues were designed on the basis of the melanotan II (MT II) from NMR structure, where the pharmacophore in arginine was mimicked via backbone N_{α}-alkylation with the introduction of a guanidinyl-butyryl group [38]. The binding affinity and adenylate cyclase activity assays of these peptidomimetics at human melanocortin receptors showed that three of the new α-MSH analogues act as antagonists and exhibited high selectivity toward the human melanocortin-4 receptor [38].

Unfortunately, their synthesis is hampered by the high price and unavailability of many N_α-methyl amino acids. An efficient and practical preparation of N_α-methyl-N_α-(*o-nitrobenzenesulfonyl*)-α-amino acids without extensive purification is described by Kessler [29] and is reported in Fig. 4d. The procedure is based on the well-known N-alkylation of N_α-arylsulfonylamino esters which was improved by using dimethyl sulfate and DBU as base. Ester cleavage is efficiently achieved by using an SN_2-type saponification with lithium iodide, avoiding racemization. Compatibility of the synthesized N_α-methylamino acids with Fmoc solid-phase peptide synthesis worked by using normal coupling conditions to efficiently prepare *N*-methyl dipeptides [29].

3.1.4 β-Methyl and β,β-Dimethyl Amino Acids

β-Substitution can provide the rigidification of the side chains: the mono-methylation on β-carbon of the amino acids influences the conformations of the side chain by steric interactions. A β-methyl or β-substituted amino acid can have four different stereochemical structures because of two chiral carbons (2*S*,3*R*; 2*S*,3*S*; 2*R*,3*S*; 2*R*,3*R*). Each isomer can adopt one of the three different conformations of the side chain (*g*−, *g*+, and *trans*) (Fig. 1). The 2*S*,3*R* isomer favors the *g*− conformation, while the 2*S*,3*S* and 2*R*,3*R* isomers prefer for *trans* conformation. The 2*R*,3*S* isomer stabilizes the *g*+ conformation. The introduction of a methyl group into the side chains of phenylalanine or tryptophan leads to β-MePhe and β-MeTrp, whose chemical structures are shown in Fig. 5a, b, respectively. The replacement of the natural amino acids Phe or Trp by their rigidified analogues provided higher activity and biological stability of the modified peptides; indeed the systematic incorporation of β-MePhe into somatostatin sequence provided a new model for the ligand-receptor interaction, based on the activity changes induced by different configurations at the β center [39].

Among the natural 20 amino acids already three residues show β-disubstitutions: valine bearing two β-methyl substituents, isoleucine which has a β-methyl and a β-ethyl substitution, and threonine which has a β-methyl and a β-hydroxy substitution. Besides β-disubstituted amino acids several other side-chain-modified amino acids are reported, for example the 2-naphthylalanine [40] (Fig. 5c) or penicillamine (β,β-dimethyl cysteine) (Fig. 5d); this latter has been employed for the synthesis of cyclic peptides through disulfide bridges of known biologically active peptides, such as angiotensin II, RGD, and opioids [41, 42]. The incorporation of penicillamine affects the angles of the disulfide bonds in terms of steric constraint. These modifications do not greatly perturb the backbone, allowing the peptide backbone and the side chain to have a certain degree of flexibility, which often is crucial for the activity of peptidomimetic.

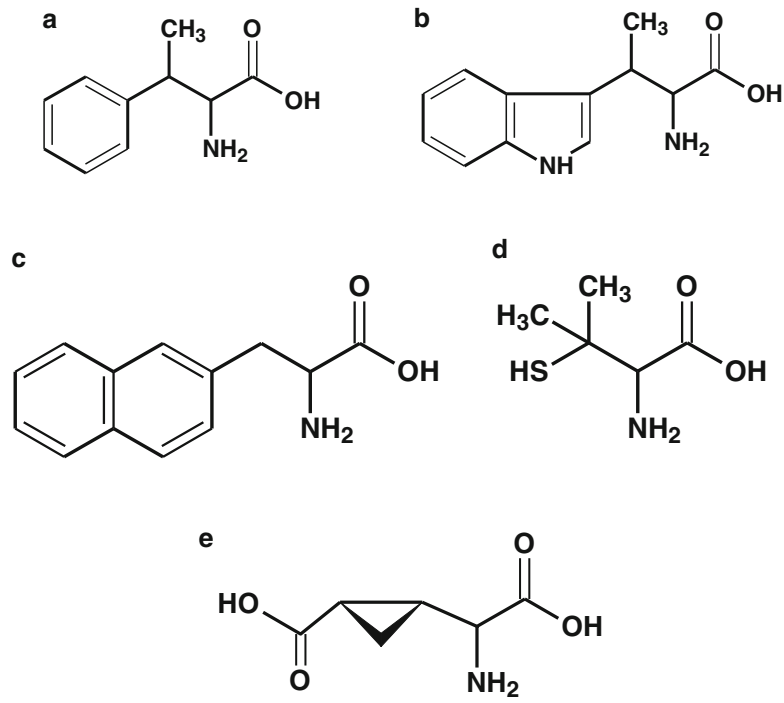


Fig. 5 Chemical structure of (**a**) β-methylphenylalanine; (**b**) β-methyltryptophan; (**c**) 2-naphthylalanine; (**d**) penicillamine (β,β-dimethyl cysteine); (**e**) 2-(carboxycyclopropyl)-glycine (CCG)

Another advantage of these modifications is that the extra alkyl groups can enhance the lipophilicity of peptide and therefore can help it to cross membrane barriers [13].

Other examples reported on the introduction of three methyl groups at the 2′-, 6′-, and β-position of natural tyrosine able to hinder the free rotation around the Cβ–Cγ bond that result in the formation of biologically active conformations [43].

In the same area of chemical modifications, a very interesting case is represented by 2-(carboxycyclopropyl)-glycine (CCG) (Fig. 5e). (2*S*,1′*S*,2′*R*)-2-(carboxycyclopropyl)-glycine is one of the four stereoisomers of a conformationally restricted glutamate analogue naturally present in the African akee apple (*Blighia sapida*) and related species [44]. The agonist potencies and selectivities of so-modified compounds for metabotropic glutamate receptors (mGluRs) were analyzed through their effects on the signal transduction of representative mGluR1, mGluR2, and mGluR4 subtypes in Chinese hamster ovarian cells [44].

3.1.5 N^{α}-C^{α}-Cyclized Amino Acids (Proline Analogues)

The distinctive cyclic structure of proline's side chain gives to proline an exceptional conformational rigidity compared to other amino acids that also affects the rate of peptide bond formation. The cyclic structure of proline's side chain locks the angle φ at approximately −60° (Fig. 1).

The conformation of proline affects the secondary structure of proteins and may account for its higher prevalence in the proteins of thermophilic organisms.

One of the most important features of proline analogues relies in the *cis/trans* isomerism (Fig. 6a, b). Proline is also commonly found in turns, and aids in the formation of β turns. Most peptide bonds adopt the *trans* isomer (typically 99.9 % under unstrained conditions), chiefly because the amide hydrogen (*trans* isomer) offers less steric repulsion to the preceding Cα atom than does the following Cα atom (*cis* isomer) [45].

Numerous proline analogues have been recovered in proteins as a result of posttranslational modifications: *cis*-4-methyl-L-proline was discovered in hydrolysates of different leucinostatine [46]. *Trans*-3-hydroxyproline and *trans*-4-hydroxyproline represent constituents of common proteins as a result of posttranslational hydroxylation, especially in collagens [47]. α-Methyl-proline is a bioactive molecule restoring normal levels of bone collagen type I synthesis [48]. Additionally proline analogues were synthesized by the introduction of alkyl chains or aromatic groups in the 3-, 4-, and 5-positions of the ring [49]. In addition to proline mimetics based on ring substitutions with alkyl and aromatic groups, it is also possible to provide the incorporation of heteroatoms into the ring, or the expansion or contraction of the proline ring.

Structures as aziridine-2-carboxylic acid (Azy), azetidine-2-carboxylic acid (Aze) (Fig. 6c), or pipecolic acid (Pip) (Fig. 6d) mime the proline, even if they present different ring sizes (3, 4, and 6 atoms, respectively).

Derivatives with additional heteroatoms and halogenated prolines such as oxazolidin-4-carboxylic acid, 3-morpholino carboxylic acid, thiazolidine-4-carboxylic acid (Fig. 6e), or 1,4-thiazine-3-carboxylic acid (Fig. 6f) were synthesized and extensively studied [50]. Instead the perhydropyridazine-3-carboxylic acid has been found as a component in a natural antagonist of oxytocin.

Substitution of 5,5-dimethylthiazolidine-4-carboxylic acid (Dtc) for Pro in angiotensin II, a key peptide in blood pressure regulation, resulted in a peptidomimetic with 39 % greater agonist activity than the natural peptide [51].

Other proline mimetics are based on ring substitutions with alkyl and aromatic groups to generate high constraint groups, as derivatives of fused bicyclic proline compounds, as (3a*S*,7a*S*)-octahydroindole-2-carboxylic acid (Oic) (Fig. 6g) or (*S*,*S*,*S*)-2-azabiciclo[3.3.0]octane-3-carboxylic acid (Aoc) (Fig. 6h). Octahydroindole-2-carboxylic acid is a proline analogue that is considered to be a very useful scaffold for the optimization of pharmacologically active peptides. Due to its bicyclic structure and lipophilicity, the incorporation of Oic into peptides may be of help to overcome several limitations to the usefulness of peptides as drugs [52]. Moreover, Oic resulted to improve the affinity toward

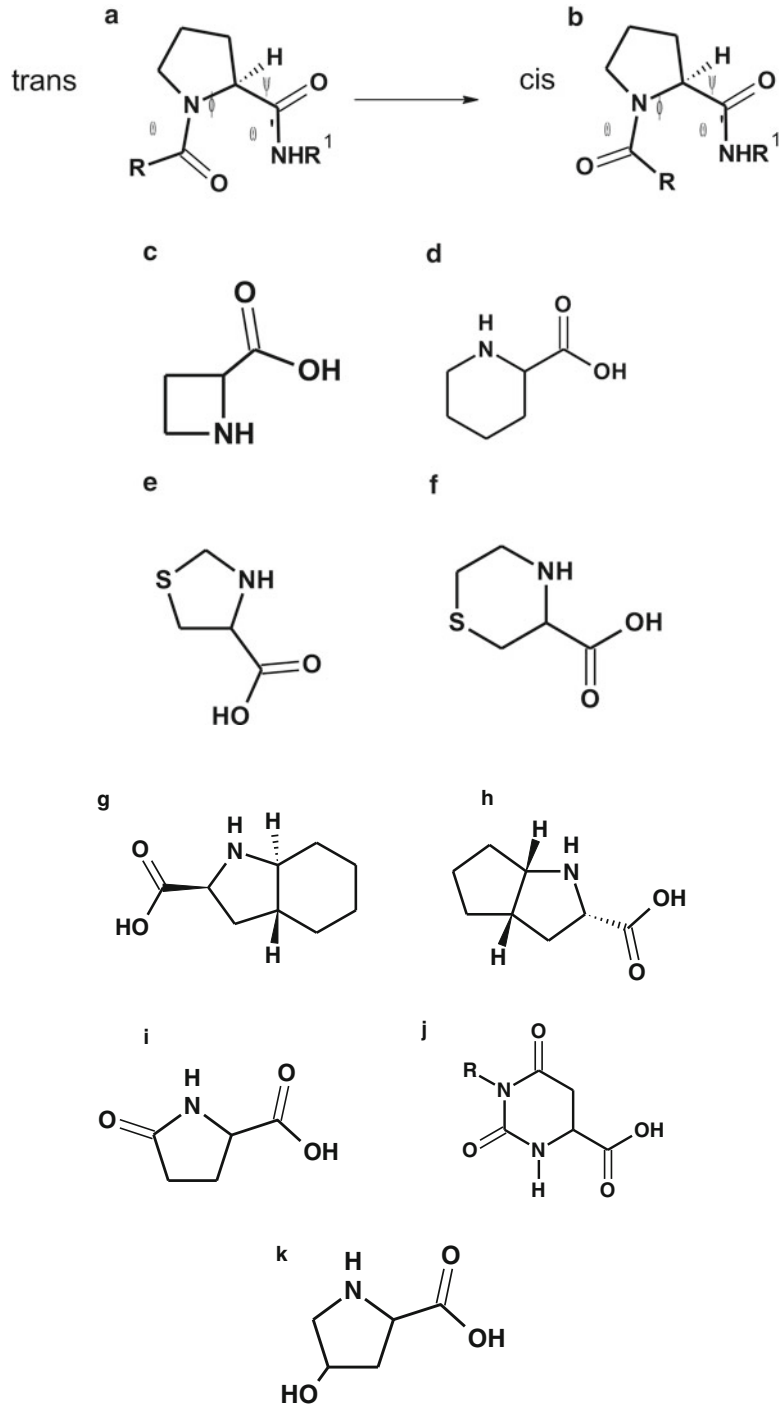


Fig. 6 Chemical structure of (**a** and **b**) *cis*/*trans* isomers of proline; (**c**) azetidine-2-carboxylic acid (Aze); (**d**) pipecolic acid (Pip); (**e**) thiazolidine-4-carboxylic acid; (**f**) 1,4-thiazine-3-carboxylic acid; (**g**) (3a*S*,7a*S*)-octahydroindole-2-carboxylic acid (Oic); (**h**) (*S*,*S*,*S*)-2-azabiciclo[3.3.0]octane-3-carboxylic acid (Aoc); (**i**) pyroglutamic acid; (**j**) (*S*)-4,5-dihydroorotic acid (Dio-OH); (**k**) hydroxyproline

certain receptors by providing better hydrophobic recognition interactions at the binding site, as in the case of tetrapeptide-based compounds, where the Oic incorporation showed to effectively better inhibit the hepatitis C virus NS3·4A protease [53].

Pyroglutamic acid or 5-oxoproline (Fig. 6i) is found as an N-terminal modification in many neuronal peptides and hormones that also include the accumulating peptides in Alzheimer's disease and familial dementia. The modification in proteins has been shown to contribute to both the structural and activity-related properties of the proteins. A series of thyrotropin-releasing hormone (TRH) analogues in which the pyroglutamic acid residue was replaced by (*S*)-4,5-dihydroorotic acid (Dio-OH) (Fig. 6j) were prepared in a Suzuki's study [54]. Of these, (1-methyl-(*S*)-4,5-dihydroorotyl)-L-histidyl-L-prolinamide showed the most potent activities, 30–90-fold greater than those of TRH.

Hydroxyproline (Fig. 6k) differs from proline by the presence of a hydroxyl group attached to the gamma carbon atom; it is produced by hydroxylation of the proline by the enzyme prolyl hydroxylase following protein synthesis and is a major component of the collagen; it plays key roles for collagen stability, along with proline [55].

3.1.6 *α,β-Unsaturated Amino Acids*

Dehydroamino acids are important precursors in the synthesis of a number of unnatural amino acids and are structural components in many biologically active peptide derivatives, indeed α,β-unsaturated amino acids are potential precursors for the formation of cross-linkages in peptides and proteins. These molecules favor the formation of β-turn (the most common form) (Fig. 7a), where the separation between the two end residues is by three bonds ($i+3$), or γ-turns when placed in the ($i+2$) position of the putative turn sequence. Dehydroamino acids (*Z* isomer more synthetically accessible than *E*) rigidify the conformation of the side chain. χ is fixed at 0° (*Z*) or 180° (*E*). Sequential placement of dehydrophenylalanine (ΔPhe) (Fig. 7b) in a peptide gives repeated β-turns, which form a 3_{10} helix [57].

Secondary structure elements were considerably influenced by the presence of dehydroamino acids. Fisher et al. synthesized three analogues of the potent vasodilator peptide bradykinin (BK), containing dehydrophenylalanine in place of the phenylalanyl residues at positions 5 and/or 8. All synthetic analogues appear to be more resistant than BK to enzymatic degradation during passage through the pulmonary vascular bed [58]. Despite the versatility of dehydroamino acids in structural determinants of peptidomimetics, they cannot have wide application since efficient synthetic procedures for their production in large amounts and without side reactions are limited.

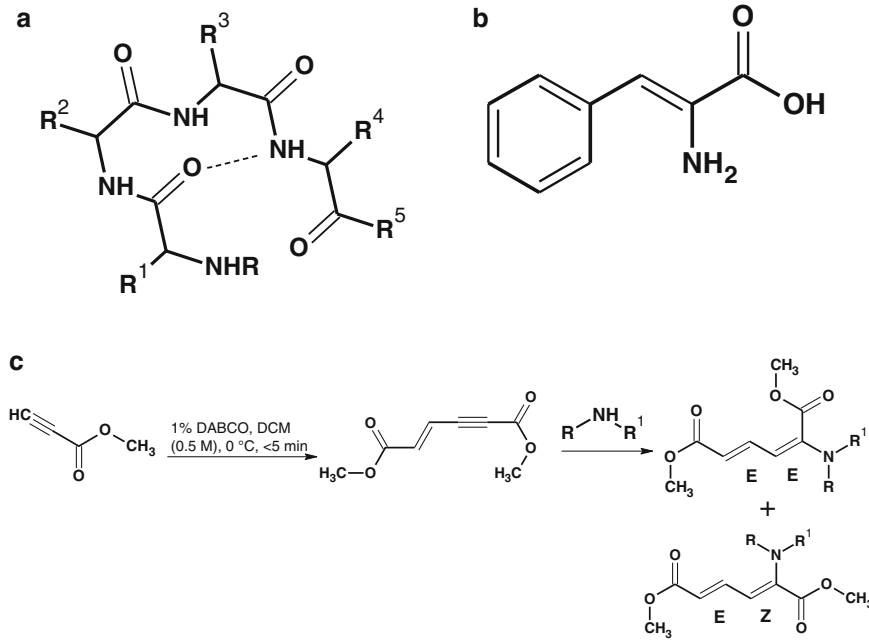


Fig. 7 (**a**) Schematic description of a β-turn; (**b**) dehydrophenylalanine (ΔPhe); (**c**) scheme of synthetic procedure of α,β-dehydroamino acid derivatives described by Chavan [56]

Recently Chavan reported the direct nucleophilic addition of alkyl amines to the α(δ′)-carbon atom of dimethyl (*E*)-hex-2-en-4-ynedioate to generate α,β-dehydroamino acid derivatives [56], as described in Fig. 7c.

3.1.7 β-Amino Acids

β-Peptides represent another class of promising peptidomimetics, consisted of β-amino acids, which have their amino group bonded to the β carbon rather than to α carbon, as in the 20 natural amino acids. The only naturally occurring β-amino acid is β-alanine; and β-peptide-based antibiotics are being explored as ways of evading antibiotic resistance [59].

A variety of β-amino acids demonstrated to be non-mutagenic by Ames tests and large elimination half-lives (3/10 h) in the serum of rodents. Conjugates of α- and β-peptides are efficient ligands for the HLA*B27 MHC class I protein, showing a fivefold increase of binding compared to a natural peptide ligand. Furthermore, β-peptides are able to mimic natural α-peptide hormones such as somatostatin. The cyclo-β-tetrapeptide derivative binds to the five human somatostatin receptors in the micromolar range [60].

β-Amino acids are subdivided into β2-, β3-, and β2,3-amino acids depending on the position of the side chain at the 3-aminopropionic acid core. β-Peptides which are readily available by standard methods either in solution or on solid support adopt a

large variety of different secondary structures in solution and in the solid state [61]. The alkyl substituents at both the α and β positions in a β-amino acid favor a gauche conformation around the bond α-β-carbon. The structures of different β-amino acids and their expected conformation are shown in Fig. 8.

The introduction of β-amino acid affects the thermodynamic stability of the structure; indeed many types of helix structures consisting of β-peptides have been reported; generally, β-peptides form more stable helices than α-peptides [62, 63].

They are smartly obtained from Fmoc-protected amino acids through a sonication of diazo ketones in dioxane in the presence of silver benzoate and water [64]; several steps of this procedure are shown in Fig. 9.

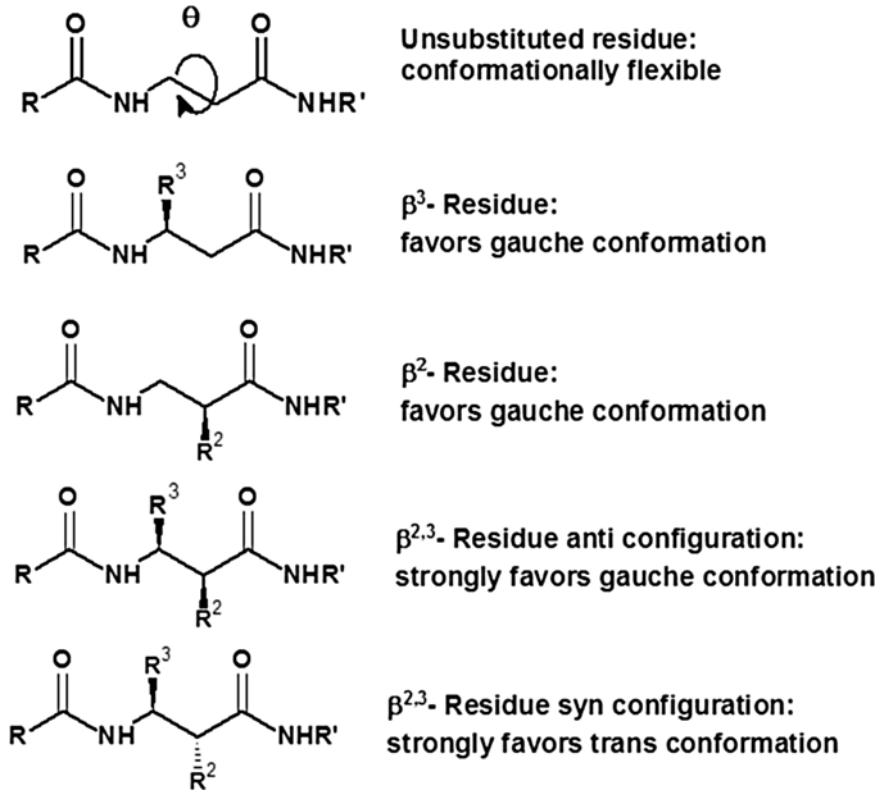


Fig. 8 Chemical structure of β-amino acids: β2-, β3-, and β2,3-amino acids and their expected conformation

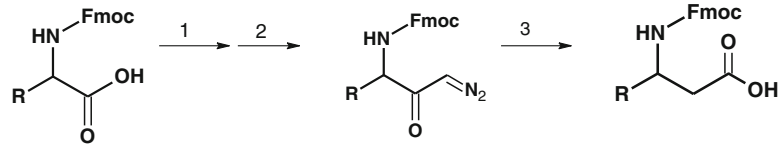


Fig. 9 Scheme of synthetic procedure of β-amino acid by Müller et al. (*1*) triethylamine, ethyl carbonochloridate, in THF at −15 °C for 15 min; (*2*) diazomethane (in Et_2O), r.t. for 3 h; (*3*) silver(I) benzoate, dioxane/H_2O, r.t., for 30 min

3.1.8 Aromatic Amino Acid Analogues

Phe, Tyr, and Trp are hydrophobic, aromatic amino acids; both Tyr and Trp can also contribute to hydrogen bond formation and they are usually over-represented at protein-binding sites, but Trp, in particular, is the most conserved of all amino acids. Conformationally constrained aromatic amino acids have found widespread application in search of novel peptide-based ligands with minor side effects [65]. Such analogues, able to enhance receptor selectivity and affinity, can be subdivided into sterically (e.g., β-methylphenylalanine, β-methyltryptophan, β-methyl-2′,6′-dimethyltyrosine) and covalently constrained derivatives [e.g., 2-aminotetralin-2-carboxylic acid (Atc) (Fig. 10a), 2-aminoindane-2-carboxylic acid (Aic) (Fig. 10b), 1,2,3,4-tetrahydroisoquinoline-3-carboxylic acid (Tic)] (Fig. 10c).

The incorporation of these residues into peptides, restricting both the conformational freedom of the aromatic ring and peptide backbone, provides valuable insights into the bioactive conformation of the peptide ligand and often leads to more potent and selective compounds. In the effort to determine the effect of side-chain conformational restriction on opioid receptor selectivity, Schiller et al. showed that Phe replacement, in the potent cyclic opioid peptide analogue H-Tyr-D-Orn-Phe-Glu-NH_2 (which lacks significant opioid receptor selectivity), with cyclic phenylalanine analogues as Aic, Atc, and Tic resulted in more potent and selective agonists [66]. These unnatural amino acids result to be particularly effective in fixing the rotation around the Cα–Cβ (χ^1) and Cβ–Cγ (χ^2) bonds. The Tic residue restricts χ^1 to −60° (*gauche* −) or 60° (*gauche* +) rotamers, while $\chi^1 = 180°$ (*trans*) for the side chain is excluded; instead χ^2 is about 160°.

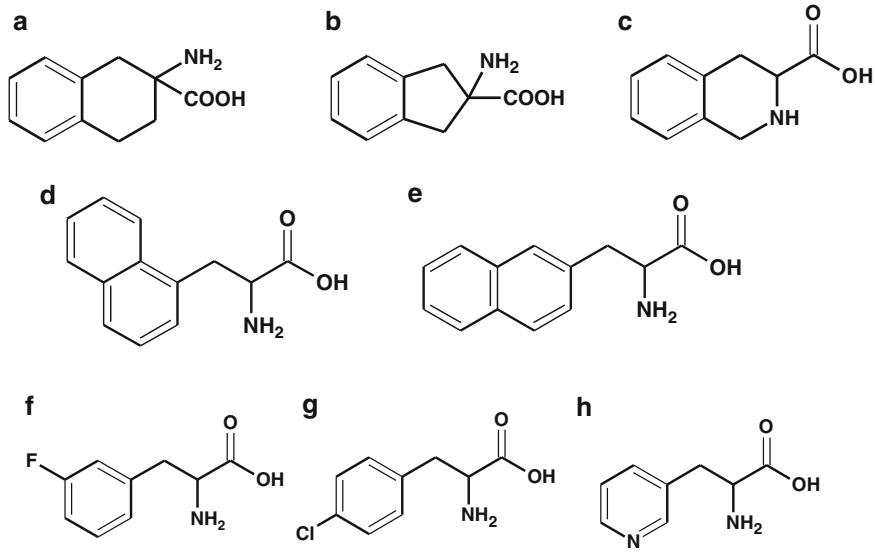


Fig. 10 Chemical structure of (**a**) 2-aminotetralin-2-carboxylic acid (Atc); (**b**) 2-aminoindane-2-carboxylic acid (Aic); (**c**) 1,2,3,4-tetrahydroisoquinoline-3-carboxylic acid (Tic); (**d**) α-1-naphthyl-alanine; (**e**) α-1-naphthyl-alanine (2NA); (**f**) α-3-fluoro-phenylalanine; (**g**) α-4-chloro-phenylalanine (1NA); (**h**) α-3-pyridyl-alanine

In other aromatic amino acid mimetics also the expansion of the aromatic ring (some examples are reported in Fig. 10d, e) and/or the incorporation of heteroatoms into the ring and/or as substituents (some examples are reported in Fig. 10f–h) have been developed.

There are many examples in literature of peptidomimetics that result from the incorporation of a heterocycle into a peptide.

Cerminara and coworkers [67] changed the core of a known class of peptidomimetic drugs, HIV protease (HIV-Pr) inhibitors, by replacing the phenyl ring with a heterocyclic ring, as thiophene or benzothiophene [68]. The choice of 4- and 5-substituted benzothiophenes was significant, not only for their isosterism with biologically active compounds containing indole rings, but also for the electronic characteristics of sulfur in interaction with the active sites of biological molecules.

Furthermore Feng and coworkers modeled the complex of a small molecular peptidomimetic inhibitor and TACE (tumor necrosis factor-α-converting enzyme) and thus they introduced 1-3-phenylalanine and 1-3-(2′-naphthyl)alanine with hydrophobic aryl side groups in the peptide segment, producing successfully novel TACE inhibitors [69].

Recently, eight analogues of prolactin-releasing peptide (PrRP20) were analyzed, in which the Phe31 was modified with a bulky side chain or a halogenated aromatic ring; they revealed in vitro and in vivo high binding potency and cell signaling in RC-4B/C cells [70]. In particular, [$PheNO_2^{31}$]PrRP20, [1-Nal^{31}]PrRP20, [2-Nal^{31}]PrRP20, and [Tyr^{31}]PrRP20 showed not only effects comparable or higher than those of PrRP20, but also a very significant and long-lasting anorexigenic effect [71].

4 Backbone Modification

Along with or alternatively to the use of modified amino acids, another widely applied approach to convert peptides into drug-like compounds entails the backbone amide replacement with amide bond surrogates, or isosteres. Various peptidomimetics containing pseudo-peptides or peptide bond surrogates, having peptide bond replaced with other chemical groups, have been designed and studied to develop new analogues with improved pharmacological properties [72]. This is mainly because such approaches create an amide bond surrogate with defined three-dimensional structures and with significant differences in polarity, hydrogen bonding capability, and acid-base character. Generally, the isosteres do not restrict global conformations, but have influence on secondary structure through different hydrogen bond patterns and lengths of backbone, even if such drastic modifications can have several

negative effects on conformation, flexibility, and hydrophobicity of new molecules.

On this basis, the choice of an amide bond surrogate represents a compromise between positive effects on pharmacokinetics and bioavailability and potential negative effects on activity and specificity [73]. The most common isosteres used are shown in Fig. 11. The psi bracket ([Ψ]) nomenclature is used for this type of modifications.

Also importantly, the structural and stereochemical integrities of the adjacent pair of α-carbon atoms in these pseudo-peptides are unchanged.

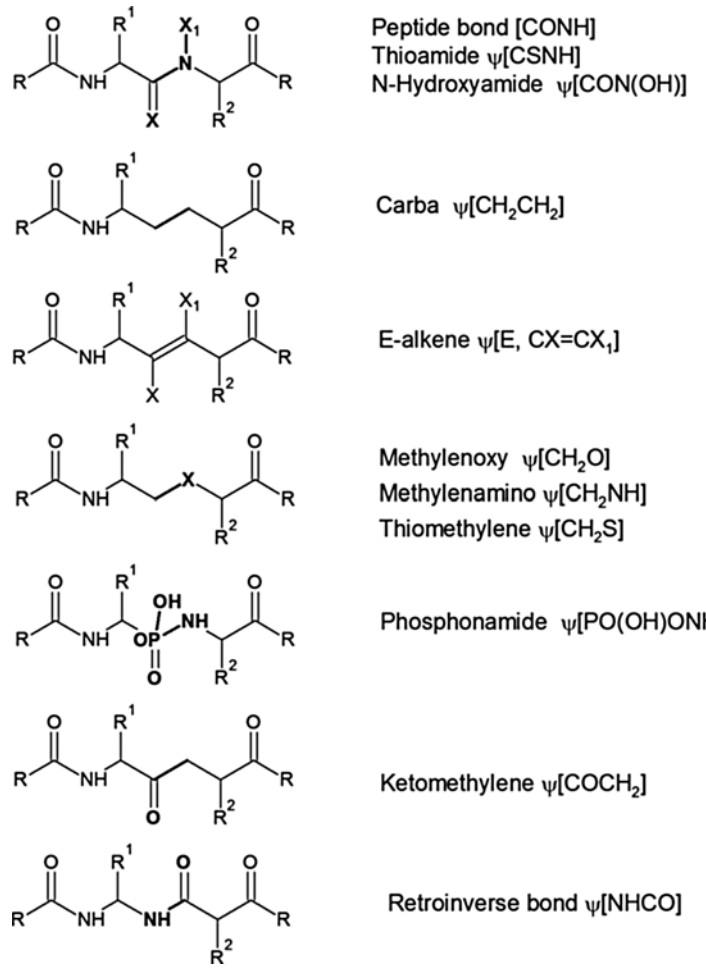


Fig. 11 Schematic description of the most common isosteres used in peptidomimetics

The incorporation of a reduced amide bond, ψ(CH_2NH), into peptide results in an increase in the net positive charge and the perturbation of a helical structure. Hah and coworkers [74] synthesized pseudodipeptide analogues with reduced amide bond and evaluated their activity as inhibitors of nitric oxide synthase (nNOS). The deletion of the carbonyl group from the amide bond either preserved or improved the potency and the selectivity.

A ψ[CS-NH] thioamide group is one of the most similar mimics of an amide linkage, even if it exhibits significantly different chemical and physical properties: the thioamide NH group is more acidic respect to its oxygenated counterpart and consequently a stronger H-bonding donor. Its *cis/trans* isomerization can be photo-triggered by irradiation at about 260 nm and it may act as a minimalist, effective quencher for any type of protein and nonprotein fluorophores.

The ketomethylene (ψ[$COCH_2$]) isostere retained hydrogen bond acceptor properties but lacks donor possibilities and is also more flexible compared to the amide bond. Retro-*inverso* peptides which contain ψ(NH–CO) bonds instead of ψ(CO–NH) bonds are made up of D-amino acids in a reversed sequence and resulted to be much more resistant to proteolysis than L-peptides. Despite their limited success in some immunological applications, retro-inverso isomers generally fail to follow the protein-binding activities of their nature peptides of an a helical nature [75].

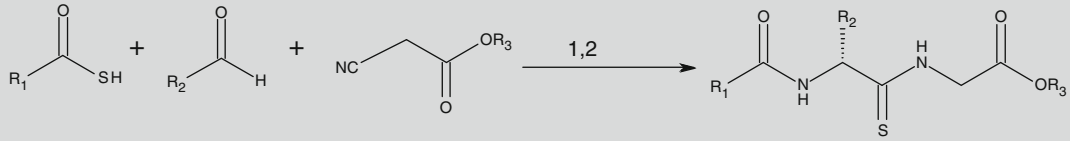
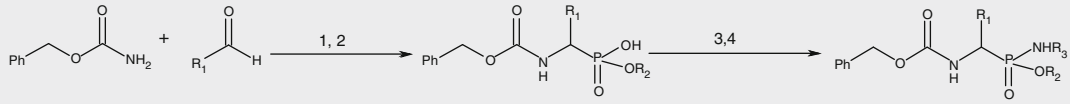
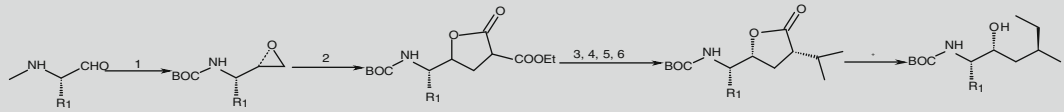
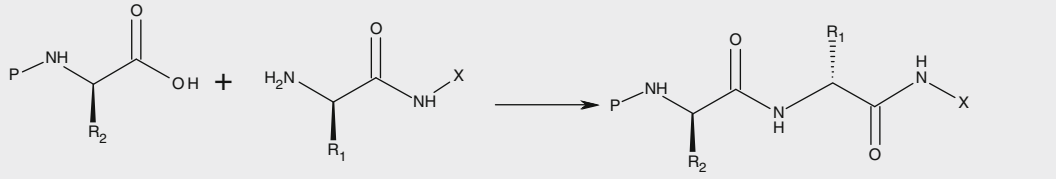
Edwards and coworkers tried to replace the peptide bond in the 4–5 position of the cyclic and linear enkephalin analogues with a thiomethylene ether linkage ψ[CH_2S]. The pseudopeptide showed high potency in both the guinea pig ileum and the mouse vas deferens assay but, therefore, it lost selectivity. It may be due to the greater flexibility of their 18-membered ring structures as a consequence of the peptide bond substitution [76].

The methylene-oxy ψ[CH_2–O] modification offers a polar, flexible, proteolytically resistant peptide bond surrogate which can be easily incorporated into biologically active peptides. The standard *trans*-amide geometries, together with methylene-oxy and methylene-thio units, were compared, showing a very close geometrical similitude of the isostere bond to the amide bond [77].

Synthetic methods for the assembly of peptidosulfonamides, phosphonopeptides, oligoureas, depsides, depsipeptides, peptoids, and azapeptides are parallel to those standard for solid-phase peptide synthesis, although different reagents and different coupling and protecting strategies need to be employed. In this field different procedures are summarized in Table 1.

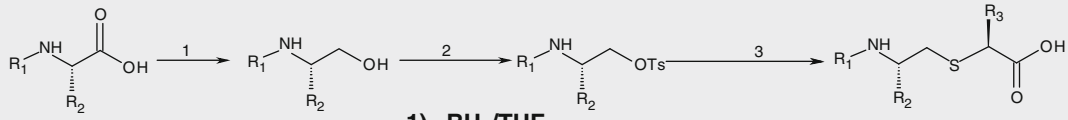
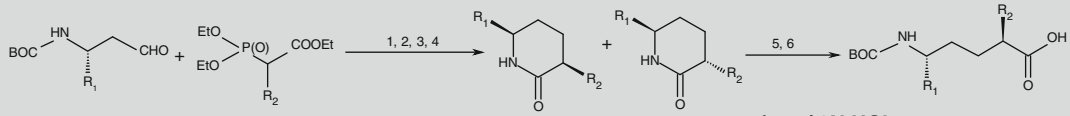
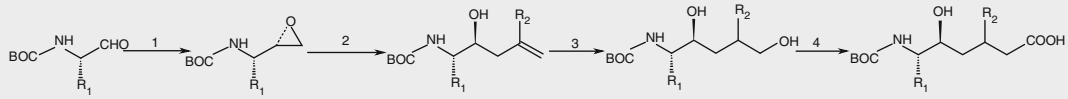
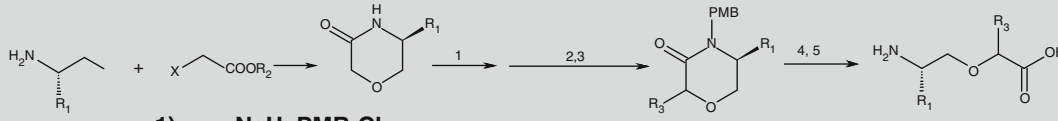
Also small heterocycles act as isosteres [87]; indeed those directly prepared from dipeptides such as oxazoles, oxazolines, oxazolidines, and their thio-derivatives resulted to be very versatile [88].

Table 1
Chemical peptide bond modification usually introduced in peptidomimetics

Peptide bond modifications
Reduced-amide bond –CH_2–NH– [78]  1,2 90 min 22°C **1) DMF, 1% AcOH** **2) NaBH3CN**
Thio-amide bond –C(S)–NH– [79]  1,2 **1) NH_4OH** **2) CF_3CH_2OH r.t., 15h**
Phosphonamide bond –P(O)OH–NH– [80]  1, 2 3,4 **1) RCHO, PCl_3, CH_2Cl_2** **2) ROH, CHCl** **3) $(COCl_2)_2$,DMF** **4) RNH, EtN, CHCl**
Ketomethylene bond –CO–CH_2– [81]  **1) CH_2SMe_2** **2) DiethylPropanedioate, $NaOC_2H_5$** **3) i-Pr-1 / $NaOC_2H_5$** **4) NaOH** **5) H+** **7) NaOH** **8) H+**
Retro-inverso bond –NH–CO– [82] 

(continued)

Table 1
(continued)

Peptide bond modifications
Thiomethylene bond $-CH_2-S-$ [83]  1) BH_3/THF 2) TOSCl/ pyridine 3) $Na^{+-}S\text{-}CHR_3COONa$, DMSO
Carba bond $-CH_2-CH_2-$ [84]  1) NaH/DME 2) H2, Pd/C, EtOH 3) TFA 4) Δ/ pyridine 5) Δ/ 6N HCl 6) $(Boc)_2O$
Hydroxyethylene bond $-CHOH-CH_2-$ [85]  1) CH_2SMe_2 2) CH_2 =C(R)CuLi 3) R3BH, NaOH 4) [O]
Methylene-oxy bond $-CH_2-O-$ [86]  1) NaH, PMB-Cl 2) LDA 3) R3X or $R3CHO/H_2$ 4) CAN 5) HCl

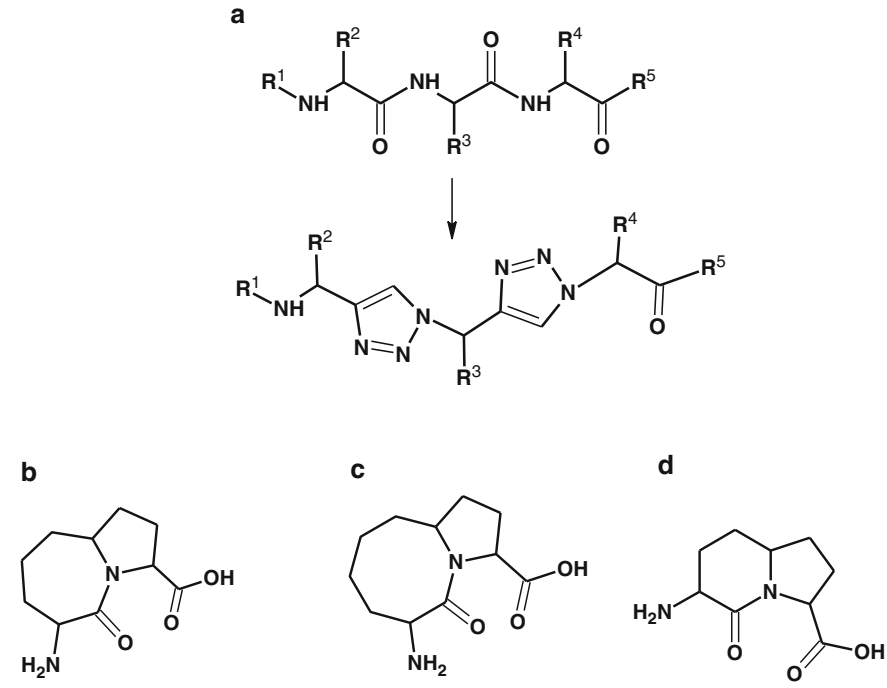


Fig. 12 Chemical structure of (**a**) 1,2,3-triazole as a bioisostere of the amide bond; (**b**) pyrroloazepin-2-one; (**c**) pyrroloazocin-2-one; (**d**) indolizidin-2-one

Recently, click chemistry routes were applied to perform synthetic modifications; for example Cu-catalyzed Huisgen 1,3-dipolar additions have been thoroughly used [89]. The 1,2,3-triazole attracts increasing attention as a bioisostere of the amide bond moiety of peptides. The similarity of the two moieties can be seen in their sizes, dipole moment, and H-bond acceptor capacities [90]. Figure 12a shows the systematic replacement of the amide bond with 1,2,3-triazoles.

The strong dipole moment of the 1,2,3-triazole ring polarizes the C(5) proton to such a degree that it can function as a hydrogen-bond donor similar to the amide NH. Furthermore, the 1,2,3-triazole ring has a large dipole that could align with that of the other amides in a given peptide secondary structure. Importantly, 1,2,3-triazoles are extremely stable to hydrolysis. Highly chemo- and regio-selective reactions are used also to insert 1,2,3-triazoles into peptide chains for macrocyclizations, and for quantitative conjugation of other subunits such as carbohydrates, polymers, or labeling agents [90].

4.1 Cyclization of Backbone

Many companies are developing new drug discovery platforms based on macrocycles: these chemical scaffolds are able to address new targets since their ring structure allows them to behave differently than most small molecules.

Macrocycles are chemically defined by a ring structure of at least 12 atoms and are typically 500–2,000 Da in size. In contrast, most small molecules weigh less than 500 Da, which has been considered the upper limit for a compound to be cell permeable and orally bioavailable [91].

Over the years, many strategies have been employed to orientate macrocyclizations using pre-organized conformations. These strategies are classified into two categories: (1) internal and (2) external conformational elements, which involve the use of scaffolds that are neither covalently attached to the sequence [92]. A similar approach relies on the development of constrained peptides by artificially linking linear peptides into specific structures possessing improved drug-like properties. The model macrocyclic drug, that inspired these efforts, could be cyclosporine, a fungal derived natural product developed as an immunosuppressant more than 30 years ago. Approved by FDA in 1983 as Sandimmune, the drug is an orally bioavailable and cell-permeable 1,200-Da cyclic peptide made up of 11 amino acids. Cyclosporine suppresses the immune system by binding to cyclophilin A, which then drives the formation of a protein-protein complex that inhibits calcineurin [93].

NMR studies have shown that cyclosporine can adopt different conformations depending on its chemical environment, which may explain how it can possess drug-like properties despite its size. Cyclosporine violates Lipinski's rules and there are many other cyclic peptide natural products with molecular weight over 1,000 that are cell permeable. Since these compounds are well outside what is normally considered drug-like, their structures might suggest a path toward the design of synthetic, non-Lipinski compounds, as we go toward more challenging targets like protein-protein interactions. Other naturally derived macrocycles include the antibiotics erythromycin and vancomycin [94] and the immunosuppressant tacrolimus [95]. However, natural macrocycles are chemically complex and difficult to synthesize, which has prevented the large-scale synthesis of compound libraries. In addition, computational challenges make SAR difficult. Early attempts to design macrocycle drugs were based on the screening of peptide libraries against targets of interest, and then attempting rational design of analogues with improved pharmacokinetic properties [8, 9, 96].

Actually several chemical methods were developed to improve the drug-like properties of peptides by constraining their structure and increasing their diversity, often in conjunction with platforms that enabled the screening of large libraries of cyclic molecules. Suga developed a method of incorporation of modified unnatural amino acids into mRNA display peptide libraries [97], while new technologies have been developed for "freezing" the 2D- and 3D-structure of short peptides in this direction, as the CLIPS (Chemical LInkage of Peptides onto Scaffolds) approach [98],

that not only rigidify the structure of the peptide, but also improve its binding activity and/or proteolytic stability to a significant extent. It involves the (multiple) cyclization of linear peptides via reaction with a small rigid entity (chemical scaffold) that carries 2, 3, or 4 anchoring points; these points react exclusively with one type of functionality of the peptide (i.e., thiols) and attach to the peptide via multiple covalent bonds. The peptide folds around the scaffold and loses flexibility while slowly adopting a well-defined 3D structure. This technology is used to affix linear peptides into (poly)cyclic structures and to bring together different parts of a protein-binding site. A similar approach relies in MATCH (macrocyclic template chemistry) platform [99], in which macrocycles incorporate three recognition moieties locked in a defined, cyclic, three-dimensional structure by a chemical fragment called tether, having a molecular weight of 100–200, while the recognition motifs originate from either natural or nonnatural amino acids and allow the interaction with targeted receptor. The tethers define and control their unique conformation, ensuring tighter binding and improved potency.

5 Conformational Constraints

Unlike proteins, peptides of under 15 or so residues in length tend not to exhibit a stable or even a preferred solution conformation; thus they have generally poor hydrophobic properties that can be sequestered from the polar environment in the folding event. Usually there will be an ensemble of conformational states in solution and if, biological activity involves only one discrete conformer, a dilution of the biologically active species occurs. The problem is most acute for peptides mimicking protein regions. Indeed while in their native environment these regions can rely on the protein's structural rigidity to hold them in a particular conformation, as free peptides do not have such influence and are endowed of intrinsic flexibility. Conformationally restrained structures can minimize binding to no target receptors and enhance the activity at the desired receptor.

Conformational constraints can be distinguished in:

- Local constraints, involving limited conformational mobility of a single residue.
- Regional constraints that affect a group of residues forming some secondary structural unit.
- Global constraints covering the whole peptide structure.

As local restrictions the side-chain modifications (Subheading 3) represent the most successful approaches [92].

Many if not most protein-protein interactions are mediated by three main recognition motifs: α-helix, β-turn, and β-strand; consequently, an attractive approach for the discovery of modulators of protein-protein interactions is to mimic the key interaction residues using small-molecule mimetics of these three major recognition motifs. An important approach involves the design of conformationally restricted analogues that mimic and/or stabilize characteristics of the receptor-bound conformation of the endogenous peptide.

Turns are prominent features in peptide secondary structure and have often been implicated in biological activities. Turns or chain reversals must arise in cyclic peptides or in peptides that occur as short loops in proteins; but linear peptides can also fold into conformations containing turns as well. There are steric constraints placed on the side chains and backbone torsions of the corner residues in a turn. Gly and Pro residues most readily accommodate these constraints, and their appearance in a sequence is suggestive of a potential turn structure. The β-turn is a common feature in biologically active peptides and is defined as any tetrapeptide sequence, with a ten-membered intramolecularly H-bonded ring, in which the Cα(i) to Cα($i+3$) distance varies from 4 to 7 Å.

Several studies showed turn mimetics, as cyclic moieties designed to replace the internal residues of a turn maintaining the overall geometry associated with it [100].

Scaffold peptidomimetics represent important pharmacophoric residues that are held in the appropriate orientation by a rigid template. Much efforts have been devoted to the design and synthesis of conformationally constrained compounds that mimic, or induce, specific secondary structural features of peptides and proteins. There are at least 14 types of β-turn structures, described in literature [101].

The scaffold most frequently applied to design a turn is the γ-lactam that is able to force the C-terminal amide, favoring a type II′ beta-turn geometry.

Ideal mimic will have a rigid scaffold that orients the side-chain residues in the same direction as the natural peptide while conferring better solubility and/or resistance to enzymatic degradation. Prominent groups, mimicking type II′ β-turns, are azabicycloalkanone amino acid scaffold as dipeptide surrogates [102]. Many variations in size ring systems have been studied as reported in Fig. 12b–d, and a variety of sites for substitution can provide the desirable structural flexibility. X-ray crystallographic analysis confirmed that the dihedral angles within the pyrrolizidine ring carboxylate were consistent with those of the central residues of a type II′ β-turn.

A recent study reported on the screening of a library against the human opioid receptors (KOR, MOR, and DOR): they identified not only the activity of library members expected to

mimic the opioid receptor peptide ligands but also additional side-chain combinations that provided enhanced receptor binding selectivities (>100-fold) and affinities (as low as Ki = 80 nM for KOR). A key insight to emerge from this study is that the phenol of Tyr in endogenous ligands bearing the H-Tyr-Pro-Trp/Phe-Phe-NH_2 β-turn that is important for MOR binding but may not be important for KOR (accommodated, but not preferred) and that the resulting selectivity for KOR observed with its removal can be increased by replacing the phenol OH with a chlorine substituent, further enhancing KOR affinity [103].

As global restraints, the simplest way is represented by cyclization that reduces the degrees of freedom of each amino acid within the loop. The most common cyclization ways used are shown in Fig. 13. This change substantially reduces the flexibility of the native linear molecule and stabilizes its specific secondary structure. The flexibility of φ, ψ, ω, and χ angles will vary depending on the size of the ring that is formed. Cyclizations characterized by rings of small size significantly reduce the accessible conformational space.

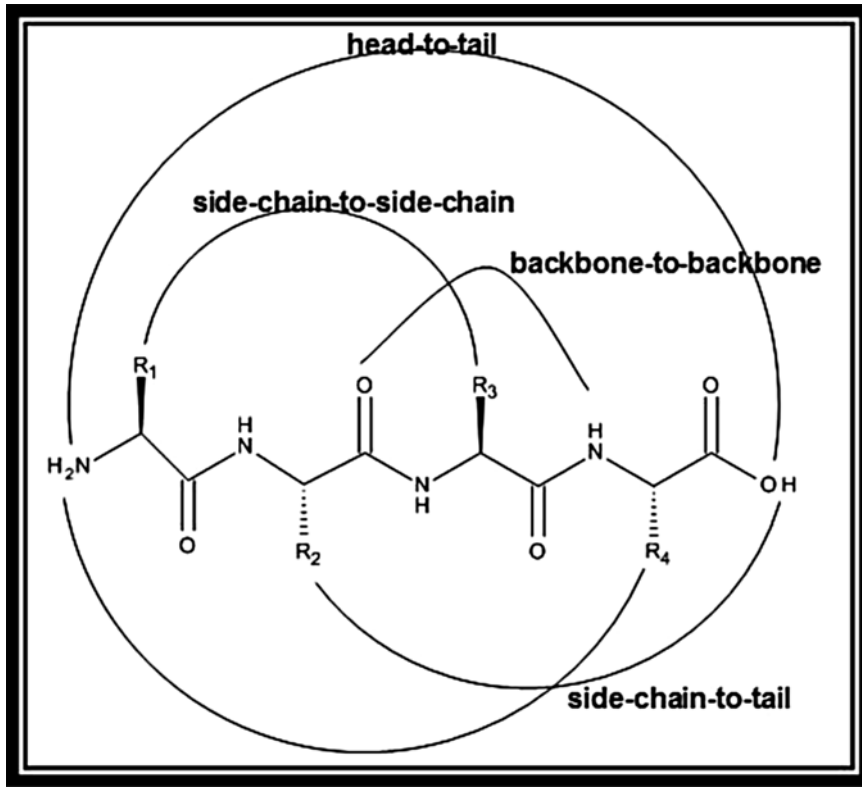


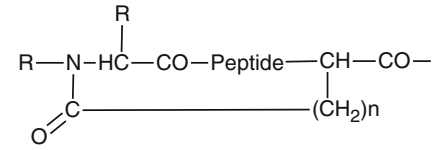
Fig. 13 Schematic description of several cyclization methods

Typically cycles increase the in vivo stability of peptides and have been observed in many natural peptides such as somatostatin, oxytocin, cyclosporine A, calcitonin, and peptide antibiotics [104]. Cyclization can be obtained by connecting:

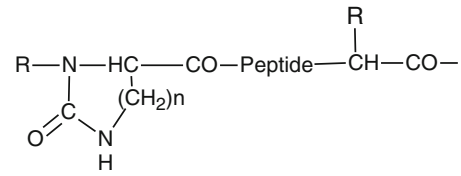
- The N- with the C-terminus (head-to-tail) portion of the peptide sequence.
- The N- or the C-terminus with one of the side chains (backbone to side chain): the most types of involved groups are reported in Fig. 14.
- Groups of side chains not involved in specific interactions with other (side chain to side chain): the most involved groups are reported in Fig. 15.

The most common side chain-to-side chain cyclization is the oxidation of two Cys residues with the formation of a disulfide bond. Alternatively, the formation of amide bonds between the side chains of Lys and Asp/Glu can occur. One limiting factor of

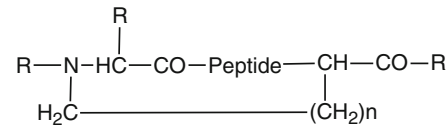
➢ **Acylation**



➢ **Trans guanidination**



➢ **Alkylation**



➢ **Thioether**

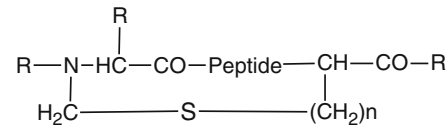
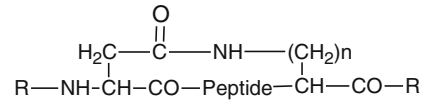
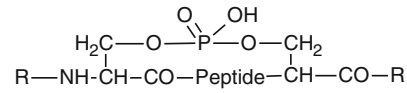


Fig. 14 Schematic description of the link between the N- or the C-terminus with one of the side chains (backbone to side chain)

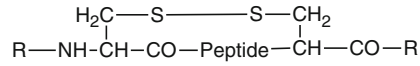
- **Amide bond**



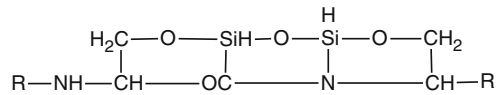
- **Phosphodiester bond**



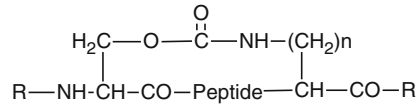
- **Disulfide bond**



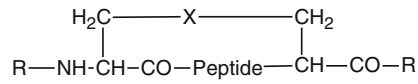
- **SiOSi bond**



- **Urethane bond**



- **Ether/Thioether bond**



- **Ester bond**

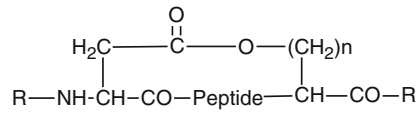


Fig. 15 Schematic description of the link among groups of side chains not involved in specific interactions with other (side chain to side chain)

side chain-to-side chain cyclization is that a limited section of the polypeptide is constrained. For this reason several covalent bridges may be incorporated into one sequence to overcome this problem [105], for example the formation of a cyclic structure using two –OH groups of side chains by a phosphodiester bond or through the formation of a disiloxane bridge. The phosphodiester bond serves similarly to the disulfide bond to maintain or stabilize

the structure of a protein. A urethane bond can instead be useful to connect a $-NH_2$ residue with an –OH group in the side chain. In addition, this bond has a high tendency to assume a *cis* configuration, thereby facilitating the formation of cyclic peptides [106].

Another cyclization is the link that connects a carboxylic group and a –OH residual. This change is common in many natural antibiotics and among renin inhibitors [107].

The highly chemo- and regio-selective reaction of 1,2,3-triazole synthesis has been used also to insert it into peptide chains for macrocyclizations, and for quantitative conjugation of other subunits such as carbohydrates, polymers, or labeling agents [108].

Similarly, important advances in the understanding of membrane permeability have come from the Kessler group [109], based on their earlier postulation that a combination of macrocyclization and N-methylation of a peptide may be a general strategy to confer the combination of membrane permeability and resistance to proteolytic degradation that is required to achieve oral bioavailability [110].

6 Conclusions

Drug design based on the natural peptide pharmacophore has become an established paradigm in novel drug discovery. The systematic SAR and screening in combination with conformational and topographical side-chain constraints can provide novel structures with novel biological activity profiles and can successfully deliver useful drug leads for various receptors. Historically, peptide vaccines to viral infections and antibacterial peptides led the way in clinical development, but recently many other diseases have been targeted, including the big sellers AIDS, cancer, and Alzheimer's disease.

In this direction the discovery of small synthetic molecules that mimic natural peptide is the main goal. These small synthetic mimics do not undergo proteolytic degradation, an advantage they hold over their natural counterparts. Small synthetic molecules make up a number of life-saving marketed drugs that inhibit certain physiologically relevant proteases. Highly stereocontrolled methods of synthesis have led to a variety of functionally diverse molecules that function as peptidomimetics because they have isosteric subunits not affected by proteolytic enzymes. Further studies to optimize biological activity and achieve desirable pharmacokinetic profiles can eventually lead to drug substances. The practice of constraining natural amino acids like their conformationally rigid counterparts has been highly successful in the design and synthesis of peptidomimetic molecules. With some notable exceptions, structural information gathered from protein X-ray crystallography of therapeutically relevant target enzymes, alone or in complex forms with inhibitor molecules, has been instrumental

in the design of peptidomimetics. Obtaining critical 3D structural information for the initial pharmacophore leads, when they are complexed with the receptors, is an additional tool made possible by the increasing potentialities of computational methods.

A synthetic peptidomimetic is needed to be resistant to proteolysis but to be able to maintain its biological activity. Conformationally constrained monocyclic and bicyclic unnatural amino acids can be directly incorporated in a potential inhibitor molecule as part of the design element [111].

Cyclic peptidomimetics are endowed with several new pharmacological properties that make them good potential lead compounds. Among these properties cell permeability is the central issue for systematically unlocking intracellular targets, even if much of what dictates cell permeability is completely unknown.

The mechanisms by which cycles can be taken up into cells can be broadly grouped into two categories—passive diffusion and active transport. In the passive diffusion the route of cell entry also used by small molecules is conceptually straightforward. Molecules diffuse from the blood through the cell membrane into the cell. But the properties that control passive diffusion for macrocycles differ from those that have previously been deciphered for small molecules: masking amide bonds, often by N-methylation, facilitates macrocycles passing through the cell membrane. Indeed, N-methylation is almost certainly critical for the cell penetration of cyclosporin, in which 7 of the 11 amide bonds are N-methylated. A Pfizer-academic team reported that selectively N-methylating accessible amide sites in a cyclic peptide resulted in about a fivefold increase in passive membrane permeability compared with that of the un-methylated parent molecule. Mislocalization or mistrafficking during endocytosis is going to be a huge task to get intracellularly active cyclic peptides. A higher hurdle is oral bioavailability that requires that a molecule is stable in the digestive track and passes through the endothelial cell barrier of the intestine before being exposed in the portal vein to the liver, where it must avoid metabolism or excretion to enter the bloodstream.

Actually protein-protein interactions represent therapeutic targets of growing interest in many diseases, but they are hardly addressed by small molecule approaches. Therefore, one can suggest that peptides and proteins have superior abilities in modulating disease states and, in some cases, returning the dysfunctional state to a homeostatic state with little or no toxicity. Furthermore, peptides—even rather small peptides with two to ten residues—have inherent three-dimensional (3D) properties. Consequently, ancillary sites enable chemical modification (e.g., with fluorophores, drugs, imaging agents) with no or minimal loss in biological activity. Peptides, therefore, have revolutionary potential in serving as drugs for early detection and treatment of disease [112].

In this context peptide macrocycles can have a prominent role. One important structural advance would be to better define pockets where macrocycles might bind, or to identify cryptic pockets—pockets that are not formed until after a ligand binds—that are potentially druggable with macrocycles. These allosteric pockets outside of a direct protein-protein interface could be important to identify because they may modulate an otherwise intractable protein target or may provide a unique mechanistic effect that cannot be achieved by directly blocking a protein-protein interaction.

Therefore, the peptidomimetic process, which aims at using peptides derived from either polypeptide-binding partner directly or after modification to improve affinity and physicochemical properties, continues to be very attractive.

References

1. Nestor JJ Jr (2009) The medicinal chemistry of peptides. Curr Med Chem 33:4399–4418
2. Aloj L, Morelli G (2004) Design synthesis and preclinical evaluation of radiolabeled peptides for diagnosis and therapy. Curr Pharm Des 10:3009–3031
3. Craik DJ, Fairlie DP, Liras S et al (2013) The future of peptide-based drugs. Chem Biol Drug Des 81:136–147
4. Hefti FF (2008) Requirements for a lead compound to become a clinical candidate. BMC Neurosci. doi:10.1186/1471-2202-9-S3-S7
5. Thaker HD, Sgolastra F, Clements D et al (2011) Synthetic mimics of antimicrobial peptides from triaryl scaffolds. J Med Chem 54: 2241–2254
6. Yang SY (2010) Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug Discov Today 15: 444–450
7. Gante J (1994) Peptidomimetics—tailored enzyme inhibitors. Angew Chem Int Ed Engl 33:1699–1720
8. Marasco D, Perretta G, Sabatella M et al (2008) Past and future perspectives of synthetic peptide libraries. Curr Protein Pept Sci 9:447–467
9. Scognamiglio PL, Di Natale C, Perretta G et al (2013) From peptides to small molecules: an intriguing but intricated way to new drugs. Curr Med Chem 20:3803–3817
10. Karle IL (1996) Flexibility in peptide molecules and restraints imposed by hydrogen bonds, the AiB residue, and core inserts. Biopolymers 1:157–180
11. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 4:95–99
12. Grauer A, König B (2009) Peptidomimetics—a versatile route to biologically active compounds. Eur J Org Chem 30:5099–5111
13. Hruby VJ, Li GG, Haskell Luevano C et al (1997) Design of peptides, proteins, and peptidomimetics in chi space. Biopolymers 3: 219–266
14. Komarov IV, Grigorenko AO, Turov AV et al (2004) Conformationally rigid cyclic α-amino acids in the design of peptidomimetics, peptide models and biologically active compounds. Usp Khim 73:849–876
15. Silverman RB (2004) Drug discovery design and development E7 peptidomimetics. In: Silverman RB (ed) The organic chemistry of drug design and drug action. Elsevier Academic, Evanston, IL, pp 47–50
16. London N, Movshovitz-Attias D, Schueler-Furman O (2010) The structural basis of peptide-protein binding strategies. Structure 18:188–199
17. Hruby VJ, Balse PM (2000) Conformational and topographical considerations in designing agonist peptidomimetics from peptide leads. Curr Med Chem 7:945–970
18. Toniolo C, Goodman M (2003) Introduction to the synthesis of peptidomimetics. In: Goodman M (ed) Methods of organic chemistry: synthesis of peptides and peptidomimetics. Thieme, Stuttgart, NY, pp 1–2
19. Rochon K, Proteau-Gagné A, Bourassa P et al (2013) Preparation and evaluation at the delta opioid receptor of a series of linear leu-enkephalin analogues obtained by systematic replacement of the amides. ACS Chem Neurosci 4:1204–1216

20. Dhanik A, McMurray JS, Kavraki LE (2012) Binding modes of peptidomimetics designed to inhibit STAT3. PLoS One 7:51603
21. Abdou WM, Barghash RF, Bekheit MS (2012) Carbodiimides in the synthesis of enamino- and α-aminophosphonates as peptidomimetics of analgesic/antiinflammatory and anticancer agents. Arch Pharm 345:884–895
22. Ibrahim IAA, Shahzad N, Al-Joudi FS et al (2013) In vitro and in vivo study of effect of α-adrenergic agonist-methyldopa on the serum biochemical laboratory findings. Clin Exp Pharmacol. doi:10.4172/2161-1459.1000136
23. International Application No. PCT/US98/04254
24. Gobbo M, Biondi L, Filira F et al (1998) Helix induction potential of N-terminal α-methyl, α-amino acids. Lett Pept Sci 5:105–107
25. Choi WT, Duggineni S, Xu Y et al (2012) Targeting the CXC chemokine receptor 4 (CXCR4). J Med Chem 55:977–994
26. Fujii N, Oishi S, Hiramatsu K et al (2003) Molecular-size reduction of a potent CXCR4-chemokine antagonist using orthogonal combination of conformation- and sequence-based libraries. Angew Chem Int Ed 42:3251–3253
27. Mungalpara J, Thiele S, Eriksen Ø et al (2012) Rational design of conformationally constrained cyclopentapeptide antagonists for C-X-C Chemokine receptor 4 (CXCR4). J Med Chem 55:10287–10291
28. Zuckermann RN, Kodadek T (2009) Peptoids as potential therapeutics. Curr Opin Mol Ther 11:299–307
29. Biron E, Kessler H (2005) Convenient synthesis of N-methylamino acids compatible with Fmoc solid-phase peptide synthesis. J Org Chem 70:5183–5189
30. Miller SM, Simon RJ, Zuckermann RN et al (1995) Comparison of the proteolytic susceptibilities of homologous L-amino acid, D-amino acid, and N-substituted glycine peptide and peptoid oligomers. Drug Dev Res 35:20–32
31. Yoo B, Kirshenbaum K (2008) Peptoid architectures: elaboration, actuation, and application. Curr Opin Chem Biol 12:714–721
32. Armand P, Kirshenbaum K, Goldsmith RA et al (1998) NMR determination of the major solution conformation of a peptoid pentamer with chiral side chains. Proc Natl Acad Sci U S A 95:4309–4314
33. Malakoutikhah M, Prades R, Teixidó M et al (2010) N-methyl phenylalanine-rich peptides as highly versatile blood-brain barrier shuttles. J Med Chem 25:2354–2363
34. Doedens L, Opperer F, Cai M et al (2010) Multiple N-methylation of MT-II backbone amide bonds leads to melanocortin receptor subtype hMC1R selectivity; pharmacological and conformational studies. J Am Chem Soc 132:8115–8128
35. Bach AC, Eyermann CJ, Groos JD et al (1994) Structural studies of a family of high affinity ligands for IIb/IIIa. J Am Chem Soc 116:3207–3219
36. Li H, Zemel R, Lopes DHJ et al (2012) A two-step strategy for SAR studies of N-methylated Aβ42 C-terminal fragments as Aβ42 toxicity inhibitors. Chem Med Chem 5:515–522
37. Biron E, Chatterjee J, Ovadia O et al (2008) Improving oral bioavailability of peptides via multiple N-methylation: somatostatin analogs. Angew Chem Int Ed 47:2595–2599
38. Ying J, Gu X, Cai M et al (2006) Design, synthesis, and biological evaluation of new cyclic melanotropin peptide analogues selective for the human melanocortin-4 receptor. J Med Chem 49:6888–6896
39. Huang Z, He YB, Raynor K et al (1992) Main-chain and side-chain chiral methylated somatostatin analogs: synthesis and conformational analyses. J Am Chem Soc 114:9390–9401
40. Tamamura H, Hiramatsu K, Ueda S et al (2005) Stereoselective synthesis of [L-Arg-L/D-3-(2-naphthyl)alanine]-type (E)-alkene dipeptide isosteres and its application to the synthesis and biological evaluation of pseudopeptide analogues of the CXCR4 antagonist FC131. J Med Chem 48:380–391
41. Mosberg HI, Hurst R, Hruby VJ et al (1983) Bis-penicillamine enkephalins possess highly improved specificity toward delta opioid receptors. Proc Natl Acad Sci U S A 80:5871–5874
42. Spear KL, Brown MS, Reinhard EJ et al (1990) Conformational restriction of angiotensin II: cyclic analogues having high potency. J Med Chem 33:1935–1940
43. Lu Y, Nguyen TM, Weltrowska G (2001) [2′,6′-Dimethyltyrosine]dynorphin A(1-11)-NH2 analogues lacking an N-terminal amino group: potent and selective kappa opioid antagonists. J Med Chem 44:3048–3053
44. Moussa CEH, Mitrovic AD, Vandenberg RJ (2002) Effects of L-glutamate transport inhibition by a conformationally restricted glutamate analogue (2S,1′S,2′R)-2-(carboxycyclopropyl) glycine (L-CCG III) on metabolism in brain tissue in vitro analysed by NMR spectroscopy. Neurochem Res 27:27–35
45. Stewart DE, Sarkar A, Wampler JE (1990) Occurrence and role of cis peptide bonds in protein structures. J Mol Biol 214:253–260
46. Degenkolb T, Berg A, Gams AW et al (2003) The occurrence of peptaibols and structurally

related peptaibiotics in fungi and their mass spectrometric identification via diagnostic fragment ions. J Pept Sci 9:666–678

47. Olsen BR, Ninomiya Y (1998) Collagens. In: Kreis T, Vale R (eds) Guidebook to the extracellular matrix and adhesion proteins. Oxford University Press, Oxford, p 40
48. Lubec G, Labudova O, Seebach D et al (1995) Alpha-methyl-proline restores normal levels of bone collagen Type I synthesis in ovariectomized rats. Life Sci 57:2245–2252
49. Thamm P, Musiol H-J, Moroder L (2003) Synthesis of peptides containing proline analogues. In: Goodman M (ed) Methods of organic chemistry: synthesis of peptides and peptidomimetics. Thieme, Stuttgart, NY, pp 52–86
50. Bhagwanth S, Mishra RK, Johnson RL (2013) Development of peptidomimetic ligands of Pro-Leu-Gly-NH2 as allosteric modulators of the dopamine D2 receptor. J Org Chem 9: 204–214
51. Samanen J, Cash T, Narindray D et al (1991) An investigation of angiotensin II agonist and antagonist analogues with 5,5-dimethylthiazolidine-4-carboxylic acid and other constrained amino acids. J Med Chem 34:3036–3043
52. Adessi C, Soto C (2002) Converting a peptide into a drug: strategies to improve stability and bioavailability. Curr Med Chem 9:963–978
53. Perni RB, Chandorkar G, Cottrell KM et al (2007) Inhibitors of hepatitis C virus NS3.4A protease. Effect of P4 capping groups on inhibitory potency and pharmacokinetics. Bioorg Med Chem Lett 17:3406–3411
54. Suzuki M, Sugano H, Matsumoto K et al (1990) Synthesis and central nervous system actions of thyrotropin-releasing hormone analogues containing a dihydroorotic acid moiety. J Med Chem 33:2130–2137
55. Szpak P (2011) Fish bone chemistry and ultrastructure: implications for taphonomy and stable isotope analysis. J Archaeol Sci 38: 3358–3372
56. Chavan AS, Deng JC, Chuang SC (2013) α(δ′)-Michael addition of alkyl amines to dimethyl (E)-hex-2-en-4-ynedioate: synthesis of α, β-dehydroamino acid derivatives. Molecules 18:2611–2622
57. Pathak S, Chauhan VS (2011) Rationale-based, de novo design of dehydrophenylalanine-containing antibiotic peptides and systematic modification in sequence for enhanced potency [down-pointing small open triangle]. Antimicrob Agents Chemother 55:2178–2188
58. Fisher GH, Berryer P, Ryan JW et al (1981) Dehydrophenylalanyl analogs of bradykinin: synthesis and biological activities. Arch Biochem Biophys 211:269–275
59. Appella DH, Christianson LA, Karle IL et al (1996) β-peptide foldamers: robust helix formation in a new family of amino acid oligomers. J Am Chem Soc 118:13071–13072
60. Gademann K, Hintermann T, Schreiber JV (1999) Beta-peptides: twisting and turning. Curr Med Chem 6:905–925
61. Weiner B, Szymański W, Janssen DB et al (2010) Recent advances in the catalytic asymmetric synthesis of β-amino acids. Chem Soc Rev 39:1656–1691
62. Murray JK, Farooqi B, Sadowsky JD et al (2005) Efficient synthesis of a beta-peptide combinatorial library with microwave irradiation. J Am Chem Soc 127:13271–13280
63. Cheng RP, Gellman SH, DeGrado WF (2001) beta-Peptides: from structure to function. Chem Rev 101:3219–3232
64. Müller A, Vogt C, Sewald N (2006) Synthesis of Fmoc-β-homoamino acids by ultrasounds-promoted Wolff rearrangement. Synthesis 837–841
65. Ballet S, Feytens D, De Wachter R et al (2009) Conformationally constrained opioid ligands: the Dmt-Aba and Dmt-Aia vs. Dmt-Tic scaffold. Bioorg Med Chem Lett 19:433–437
66. Schiller PW, Nguyen TM, Weltrowska G et al (1993) Differential stereochemical requirements of mu vs. delta opioid receptors for ligand binding and signal transduction: development of a class of potent and highly delta-selective peptide antagonists. Proc Natl Acad Sci U S A 89:11871–11875
67. Cerminara I, Chiummiento L, Funicello M et al (2012) Heterocycles in peptidomimetics and pseudopeptides: design and synthesis. Pharmaceuticals 5:297–316
68. Chiummiento L, Funicello M, Lupattelli P et al (2012) Synthesis and biological evaluation of novel small non peptidic HIV-1PIs: the benzothiophene ring as an effective moiety. Bioorg Med Chem. doi:10.1016/j.bmcl.2012.02.046
69. Feng W, Zhao Y, Huang W et al (2010) Molecular modeling and biological effects of peptidomimetic inhibitors of TACE activity. J Enzyme Inhib Med Chem 25:459–466
70. Maletinska L, Spolcova A, Maixnerova J et al (2011) Biological properties of prolactin-releasing peptide analogs with modified aromatic ring of C-terminal phenylalanine amide. Biopolymers 96:481
71. Findeisen M, Rathmann D, Annette G (2011) RFamide peptides: structure, function, mechanisms

and pharmaceutical potential. Pharmaceuticals 4:1248–1280
72. Choudhary A, Raines RT (2011) An evaluation of peptide-bond isosteres. Chembiochem 12:1801–1807
73. Stawikowski M, Cudic P (2007) Depsipeptide synthesis. Methods Mol Biol 386:321–339
74. Hah JM, Martásek P, Roman LJ et al (2003) Aromatic reduced amide bond peptidomimetics as selective inhibitors of neuronal nitric oxide synthase. J Med Chem 46:1661–1669
75. Li C, Pazgier M, Li J et al (2010) Limitations of peptide retro-inverso isomerization in molecular mimicry. J Biol Chem 18:19572–19581
76. Edwards JV, Spatola AF, Lemieux C et al (1986) In vitro activity profiles of cyclic and linear enkephalin pseudopeptide analogs. Biochem Biophys Res Commun 136: 730–736
77. Rubini E, Gilon C, Selinger Z et al (1986) Synthesis of isosteric methylene-oxy pseudodipeptide analogues as novel amide bond surrogate units. Tetrahedron 42:6039–6045
78. Fields CG, Fields GB (1994) Solvents for solid-phase peptide synthesis. In: Pennington MW, Dunn BM (eds) Peptide synthesis protocols, vol 35, Methods in molecular biology. Humana Press, Inc., Totowa, NJ, pp 29–40
79. Kazmaier U, Persch A (2010) A straightforward approach towards 5-substituted thiazolylpeptides via the thio-Ugi-reaction. Org Biomol Chem 8:5442–5447
80. Cressin E, Lloyd AJ, De Pascale G et al (2009) Inhibition of tRNAdependent ligase MurM from Streptococcus pneumonia by phosphonate and sulfonamide inhibitors. Bioorg Med Chem 17:3443–3455
81. Hoffman RV, Tao J (1997) A simple, stereoselective synthesis of ketomethylene dipeptide isosteres. Tetrahedron 53: 7119–7126
82. Fletcher MM, Campbell MM (1998) Partially modified retro-inverso peptides: development, synthesis, and conformational behavior. Chem Rev 98:763–796
83. Crozet Y, Wen JJ, Loo RO et al (1997–1998) Synthesis and characterization of cyclic pseudopeptide libraries containing thiomethylene and thiomethylene-sulfoxide amide bond surrogates. Mol Divers 3:261–276
84. Rodriguez M, Heitz A, Martinez J (1990) "Carba" peptide bond surrogates: synthesis of Boc-L-Leu-(CH2-CH2)-L-Phe-OH and Boc-L-Leu-ψ-(CH2-CH2)-D-Phe-OH through a horner-emmons reaction. Tetrahedron Lett 30:7319–7322
85. Pégorier L, Larchevêque M (1995) A general stereocontrolled synthesis of hydroxyethylene dipeptide isosteres. Tetrahedron Lett 36: 2753–2756
86. Norman BH, Kroin JS (1996) Alkylation studies of N-protected-5-substituted morpholin-3-ones. A stereoselective approach to novel methylene ether dipeptide isosteres. J Org Chem 61:4990–4998
87. Goodman M (2003) Synthesis of peptides and peptidomimetics. In: Goodman M (ed) Houben-Weyl methods in organic chemistry. Georg Thieme, Stuttgart, NY, pp 101–141
88. Wipf P, Wang X (2002) Parallel synthesis of oxazolines and thiazolines by tandem condensation-cyclodehydration of carboxylic acids with amino alcohols and aminothiols. J Comb Chem 4:656–660
89. Angell YL, Burgess K (2007) Peptidomimetics via copper-catalyzed azide-alkyne cycloadditions. Chem Soc Rev 36:1674–1689
90. Tron GC, Pirali T, Billington RA et al (2008) Click chemistry reactions in medicinal chemistry: applications of the 1,3-dipolar cycloaddition between azides and alkynes. Med Res Rev 28:278–308
91. Lipinski CA (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23: 3–25
92. White CJ, Yudin AK (2011) Contemporary strategies for peptide macrocyclization. Nat Chem 3:509–524
93. Hess AD, Colombani PM, Esa AH (1986) Cyclosporine and the immune response: basic aspects. Crit Rev Immunol 6:123–149
94. Walsh CT (2002) Combinatorial biosynthesis of antibiotics: challenges and opportunities. Chembiochem 3:125–134
95. Thomson AW, Starzl TE (1992) FK 506 and autoimmune disease: perspective and prospects. Autoimmunity 12:303–313
96. Gunasekera S, Aboye TL, Madian WA et al (2013) Making ends meet: microwave-accelerated synthesis of cyclic and disulfide rich proteins via in situ thioesterification and native chemical ligation. Int J Pept Res Ther 19:43–54
97. Passiouraand T, Suga H (2013) Flexizyme-mediated genetic reprogramming as a tool for noncanonical peptide synthesis and drug discovery. Chemistry 19:6530–6536
98. http://www.pepscan.com/
99. http://www.ocerainc.com/technology/match

100. Blomberg D, Hedenström M, Kreye P et al (2004) Synthesis and conformational studies of a beta-turn mimetic incorporated in Leu-enkephalin. J Org Chem 14:3500–3508
101. Ressurreiçao ASM, Delatouche R, Gennari C et al (2011) Bifunctional 2,5-diketopiperazines as rigid three-dimensional scaffolds in receptors and peptidomimetics. Eur J Org Chem 2011:217–228
102. Cluzeau J, Lubell WD (2005) Design, synthesis, and application of azabicyclo[X.Y.O] alkanone amino acids as constrained dipeptide surrogates and peptide mimics. Biopolymers 80:98–150
103. Whitby LR, Ando Y, Setola V et al (2011) Design, synthesis, and validation of a β-turn mimetic library targeting protein–protein and peptide–receptor interactions. J Am Chem Soc 133:10184–10194
104. Goodwin D, Simerska P, Toth I (2012) Peptides as therapeutics with enhanced bioactivity. Curr Med Chem 19:4451–4461
105. Grauer A, König B (2009) Peptidomimetics—a versatile route to biologically active compounds. Eur J Org Chem 30:5099–5111
106. Parkinson GN, Wu Y, Fan P et al (1994) Crystal structure and NMR conformation of a cyclic pseudotetrapeptide containing urethane backbone linkages. Biopolymers 34: 403–414
107. Dutta AS, Gormley JJ, McLachlan PF et al (1990) Novel inhibitors of human renin. Cyclic peptides based on the tetrapeptide sequence Glu-D-Phe-Lys-D-Trp. J Med Chem 33:2552–2560
108. White CJ, Yudin AK (2011) Contemporary strategies for peptide macrocyclization. Nat Chem. doi:10.1038/nchem.1062
109. Beck JG, Chatterjee J, Laufer B et al (2012) Intestinal permeability of cyclic peptides: common key backbone motifs identified. J Am Chem Soc 134:12125–12133
110. Chatterjee J, Gilon C, Hoffman A et al (2008) N-methylation of peptides: a new perspective in medicinal chemistry. Acc Chem Res 41: 1331–1342
111. Hanessian S, Auzzas L (2008) The practice of ring constraint in peptidomimetics using bicyclic and polycyclic amino acids. Acc Chem Res 41:1241–1251
112. Hruby VJ, Cai M (2013) Design of peptide and peptidomimetic ligands with novel pharmacological activity profiles. Annu Rev Pharmacol Toxicol 53:557–580

Chapter 9

In Silico Design of Antimicrobial Peptides

Giuseppe Maccari, Mariagrazia Di Luca, and Riccardo Nifosì

Abstract

The rapid spread of drug-resistant pathogenic microbial strains has created an urgent need for the development of new anti-infective molecules, having different mechanism of action in comparison to existing drugs. Natural antimicrobial peptides (AMPs) represent a novel class of molecules with a broad spectrum of activity and a low rate in inducing bacterial resistance. In particular, linear alpha-helical cationic antimicrobial peptides are among the most widespread membrane-disruptive AMPs in nature, representing a particularly successful structural arrangement of the innate defense against microbes. However, until now, many AMPs have failed in clinical trials because of several drawbacks that strongly limit their applicability such as degradation, cytotoxicity, and high production cost. Thus, to overcome the limitations of native peptides, a rational in silico approach to AMPs design becomes a promising strategy that drastically reduce production costs and the time required for evaluation of activity and toxicity.

This chapter focuses on the strategies and methods for de novo design of potentially active AMPs. In particular, statistical-based design strategies and MD methods for modelling AMPs are elucidated.

Key words AMPs, Drug resistance, QSAR, Molecular dynamics, De novo peptide design

1 Introduction

The appearance and rapid spread of antibiotic-resistant bacteria represents a major global health problem. Infections caused by resistant microorganisms often fail to respond to conventional treatment, resulting in prolonged illness, greater risk of death, and higher costs. The decline in effectiveness of current therapies spurs research for the identification of novel molecules endowed with antimicrobial activities and new mechanisms of action.

Antimicrobial peptides (AMPs) are small evolutionally conserved molecules, representing an exciting class of drug candidates, particularly because their mechanism of action is unlikely to induce drug resistance and some of them are also active against microbial biofilms [1]. Furthermore, AMPs have been applied not only as direct antimicrobial agents but also as potential endosomolytic moieties promoting the release of biomolecules into cells for

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_9,

delivery purposes [2]. Although some AMPs are already in clinical and commercial use, the future design of novel molecules will need to minimize the toxicity against eukaryotic cells and enhance the resistance to proteolytic degradation, with a key opportunity being offered by the introduction of non-natural amino acids (AA) to contrast host resistance and increase compound's life.

AMPs belong to a vast and various class of molecules, featuring different structure, amino acid composition, and chemophysical characteristics. Therefore, an understanding of AMPs physico-chemical characteristics and modes of action is mandatory in order to develop proper design and optimization strategies. Despite their great variability, most AMPs act by perturbing the cytoplasmic membrane, thus determining cell death by osmotic shock. Membrane perturbation activity is usually determined by at least three mechanisms [3]. The best-characterized models, the "barrel-stave" and the "toroidal-pore" models, rely on the peptide ability to form transmembrane channels/pores, while in the so-called "carpet model," the peptides disrupt the bilayer in a detergent-like manner, eventually leading to the formation of micelles [4] (Fig. 1). The mechanism of membrane disruption involves several molecular properties of the peptides, each one related to individual stages of the process:

- The process of cell attachment is facilitated by a positive net charge because of the bacterial membrane constituent.

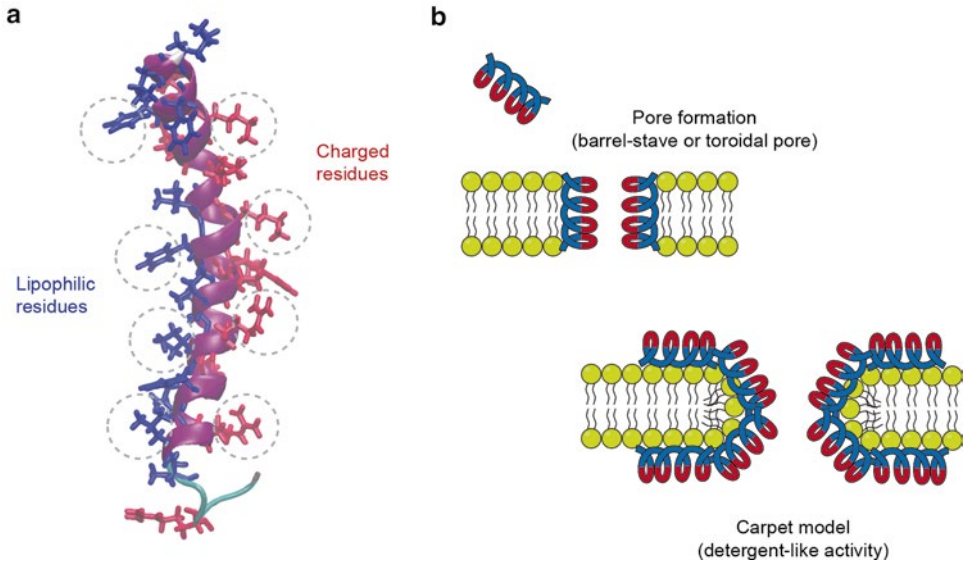


Fig. 1 AMPs chemophysical features and mode of action. (**a**) NMR structure of LL-37 (PMID: 18818205). In *red* are highlighted charged residues, while in *blue* lipophilic ones. (**b**) AMPs membrane perturbation activity. *Top left*, the "barrel-stave" and "toroidal-pore" models; *bottom-right*, the carpet model

- Aggregation facilitates the formation of a carpet on the outer side of the bacterial membrane, eventually leading to the destabilization of the lipidic bilayer. Amphipathic alpha-helical peptides better interact electrostatically with the target cell membrane.
- The overall lipophilicity rules the mechanism of permeation into the membrane, leading to a destabilization or a pore formation.

A balanced combination of these properties determines the mode of action and the overall peptide activity and cytotoxicity (Fig. 1).

Recent research on AMPs has focused on methods to search through the constellation of known or predicted peptide sequences—either empirically or computationally—for molecules with desired properties, and these approaches are continually evolving. Multi-scale approaches are increasingly applied to in silico rational design of bioactive molecules, because of their ability to study biophysical problems from multiple points of view. Multiscale approach for molecular design consists of at least two phases. The first (coarse grain) provides a fast exploration of the objective space, in order to sample its relevant regions in an approximate way. Afterwards, in a second phase, the coarse grain representation is transformed to a more detailed one, able to represent each aspect of the biological process.

Statistical-based peptide design and prediction methods are usually valid choices for unbiased screening, where speed and accuracy is a fundamental requirement. In these methods, the primary sequence information is associated with a measure of peptide activity—either quantitative or qualitative—through a series of sample sequences derived from experimentally validated peptides. A statistical model is then constructed by regression models and/or lexical methods in order to derive a rule explaining the biological activity. The derived model is then applied to stochastic or deterministic methods in order to explore the major possible number of candidates.

In contrast, computationally intensive biophysical studies are applied in order to valuate peptide folding, interaction, and mode of action of a screened list of candidates. In particular, molecular dynamics (MD) has been extensively applied for the study of AMPs in order to unravel the molecular mechanisms supporting their activity. MD simulations target the motion of the molecular system by numerically solving Newton's dynamic equation. Different resolutions can be used in the simulations, varying from all-atom one to different degrees of coarse grain, in which groups of atoms are packed into single interaction centers. From the motion of the studied systems biomolecular interactions can be inferred, and the molecular mechanisms underlying certain biological processes can be elucidated.

In this chapter, both statistical and MD design methods are discussed. In the first part, common steps in statistical-based design strategies are surveyed, from the dataset preparation procedure to the mathematical model training and validation. Furthermore, application of the designed model to deterministic and stochastic peptide design is illustrated. The second part describes MD methods for modelling AMP and their interaction with the membrane. Finally, experimental procedures for in vitro validation and measure of AMP activity are listed.

2 Statistic-Based AMP Design

In common statistic-based peptide design methods, a dataset of molecules is collected to extrapolate an adequate number of features in order to represent the desired activity. The dataset can contain quantitative information about the peptide activity such as the Minimal Inhibitory Concentration (MIC), or qualitative information such as active or inactive. In the latter case, the screening process will return a confidence score about peptide activity. Depending on the information available, each peptide in the dataset is encoded in some computer-friendly variables best representing the activity, and a regression or a classification algorithm is employed in order to distinguish peptide activity in a qualitative or quantitative fashion. In this paragraph, the process of dataset construction, model preparation, and validation are exhaustively outlined.

2.1 Dataset Preparation

In statistical analysis, the process of dataset preparation is one of the most delicate in model construction. During this phase, a list of peptides is collected in an ordered database and a specific activity is associated with primary and/or secondary structure information. Because of the remarkable variety of AMPs in terms of sequence and secondary structure, a rich and complete dataset of active and inactive peptides is difficult to obtain without introducing biases. For these reasons, during the years different bioinformatics methods were applied in order to collect as much information as possible, on natural and synthetic AMPs from the literature, facilitating the process of dataset preparation. Although information gathering can be automated (for example by iterative scanning of public sources [5]), because of the difficulty and sensitivity of the information crawling process, manually attended datasets are more appreciated (Table 1). AMPs datasets can be prepared ad hoc by experimentally screening random peptides libraries. This method has the advantage of giving precise and uniform quantitative or qualitative information of the peptide activity [6], required by complex prediction models in order to fit the correspondingly large set of parameters. Solid-phase synthesis and high-throughput screening of large peptide arrays has become a

Table 1
A chronological list of AMPs databases

Year	Database	Web site	Content
2002	AMSDb	http://www.bbcm.univ.trieste.it/~tossi/pag1.htm	Plant and animal AMPs
2007	AMPer	http://marray.cmdr.ubc.ca/cgi-bin/amp.pl	Plant and animal AMPs
2007	BACTIBASE	http://bactibase.pfba-lab-tun.org/main.php	Bacteriocins
2008	RAPD	http://faculty.ist.unomaha.edu/chen/rapd/	Recombinant AMPs
2009	PhytAMP	http://phytamp.pfba-lab-tun.org/main.php	Plant AMPs
2009	APD2	http://aps.unmc.edu/AP/main.php	Natural AMPs
2010	CAMP	http://www.bicnirrh.res.in/antimicrobial/	All AMPs
2012	DAMPD	http://apps.sanbi.ac.za/dampd/	All AMPs
2012	YADAMP	http://yadamp.unisa.it/	All AMPs
2014	BaAMPs	http://www.baamps.it	Biofilm-active AMPs

common practice in drug discovery. However, systematic studies tent to limit the number of peptides by analyzing a fixed number of amino acids positions with a precise combination of substitution [7]. Indeed, the huge number of amino acidic combinations makes an exhaustive screening of random libraries unfeasible. For example, a full combinatorial assay of peptides with length up to ten residues would result in 20^{10} different sequences, an unfeasible number of combinations. On the basis of the analysis of natural AMPs, the amino acidic space is limited to charged residues and moderately hydrophobic sequences; to avoid technical problems during the synthesis phase, cysteine and methionine residues are excluded, owing to potential cross linking or oxidation. In this way the number of combinations is extremely reduced, at the cost of some bias introduction, since a large number of substitutions are excluded a priori.

When the aim of the dataset preparation is to classify bioactive peptides, two or more different classes of sample peptides must be prepared. For the simplest case, a dataset of experimentally validated AMPs must be compared with a dataset of non-active peptides. Therefore, a list of inactive peptides must be compiled. Unfortunately, few peptides are annotated as non-antimicrobial in literature [8], and therefore negative datasets must be inferred in different ways. One is the fuzzy and unbiased random selection of peptide fragments from datasets of known proteins. Obviously, this approach can cause the unwanted inclusion of bioactive peptides in the negative dataset. In order to reduce the possibility to introduce false negative, protein datasets can be screened with knowledge-based approaches. Gene Ontology (GO) annotations

are used to mark experimentally or computationally known protein's functions and pathways, interactions, and organelles involved in their function and activity [9]. These keywords can be combined to narrow the search process into particular districts or within specific functions. AMPs are usually released from the cell in the extracellular matrix, and thus a possible strategy can be to exclude proteins and peptides marked as "secreted" or exclusively present in specific cell compartments.

Care should be taken not to introduce bias in this process, as the bigger and wider the dataset, the more precise and complete the classification. Each particular class of protein should be represented equally, as the over-representation of a particular motif or amino-acidic combination can compromise the entire dataset. For this reason, peptide datasets are usually pruned for repetitive and over-representative sequences. CD-HIT [10] implements an algorithm that, given a threshold, clusters and trims out sequences based on their similarity. Usually, a threshold of 75 % of identity is enough to assure a proper variability in the dataset. An additional method to avoid overtraining is to split the dataset into a training set—for the model training—and a test set for the validation of the model performance.

Regression models have different requirements from classification models, as the resulting function must express a measure of AMP activity using a continuous function. A dataset containing quantitative values of experimentally tested activity is therefore mandatory. Even if precise and exhaustive datasets of AMPs with quantitative activity exists [11], their use in regression model is difficult. Data collected from different works and workgroups usually is scattered, resulting in imprecise and biased models. A solution can be to distinguish categories of AMPs, grouping together highly active peptides and low active peptides on the basis of a predetermined threshold. This choice allows for a certain tolerance, thus giving some quantitative information about peptide's activity.

2.2 Peptide Representation

In order to present the dataset to a classification or regression model, each sequence must be encoded in a computer-interpretable way, able to represent peptide's salient characteristic. Amino acids can be considered the basic unit of AMPs; therefore each peptide must be represented on the basis of its sequence composition and order. The simplest and most intuitive way to represent AMPs sequences is through a linguistic model, where sequences are considered as "words" and amino acids are represented with one-letter code. As a consequence, text motives can be identified through the analysis of recurrences and grammar rules, giving useful hints about the importance of specific amino acids and residue positions to peptide activity. However, such local approaches fail to account for amino acidic position-specific interactions. Furthermore, there

is no understanding of the physicochemical variables influencing peptides activity. As an evolution of this grammar model, in order to introduce secondary structure information, different strategies have been adopted, like sequence alignment or position-specific scoring matrix (PSSM) [5, 12]. However, these approaches are limited to natural amino acids, since there is not enough sequence information of nonnatural amino acidic substitutions to build an exhaustive statistical model.

In the effort to overcome these limitations, quantitative structure–activity relationship (QSAR) models have been employed to describe the relationship between chemophysical characteristics and biological activity. These chemophysical characteristics, named *descriptors*, can be derived from experimental measures such as molecular weight, partition coefficient, or HPLC retention time, but also theoretically calculated. Calculated descriptors can be related to peptide's primary structure or chemical composition, as well as secondary or tertiary structure. Moreover, single descriptors can be combined to describe different—but related—chemophysical characteristics, like polarity and hydrophobicity, in order to reduce variable hyperspace.

In AMP design and classification, the choice of representative QSAR descriptors is directly influenced by their mechanism of action. The positive net charge and hydrophobicity are important features for the attachment and the permeation of the bacterial membrane, respectively. It is likely that only those peptides which possess a balanced combination of these properties can achieve sufficient activity in each step of the concerted mechanism and attain higher levels of antimicrobial effects. Furthermore, the overall distribution of these chemicophysical properties influences the activity. As a consequence, global descriptors can be applied to account for whole-molecular properties—such as polarity, lipophilicity, or molecular weight—while topological descriptors account for sequence order information and secondary structure (Fig. 2). A measure of sequence information can be considered by analyzing the correlation between QSAR descriptors along the primary sequence. Auto and Cross-covariance (ACC) analysis is a measure originally introduced by Wold [13]. Although various methods have been employed [14–16], the concept remains that different chemophysical descriptors are correlated between each other in a given order along the primary sequence. Basically, for a given protein sequence, ACC variables describe the average interactions between residues distributed a certain *lag* apart throughout the whole sequence. Higher lag values result in describing distant interactions along the peptide sequence. Besides encoding the sequence order, ACC has the ability to transform each amino acid sequence of variable size into uniform equal-length vectors. This feature is very important in data mining methods, where a

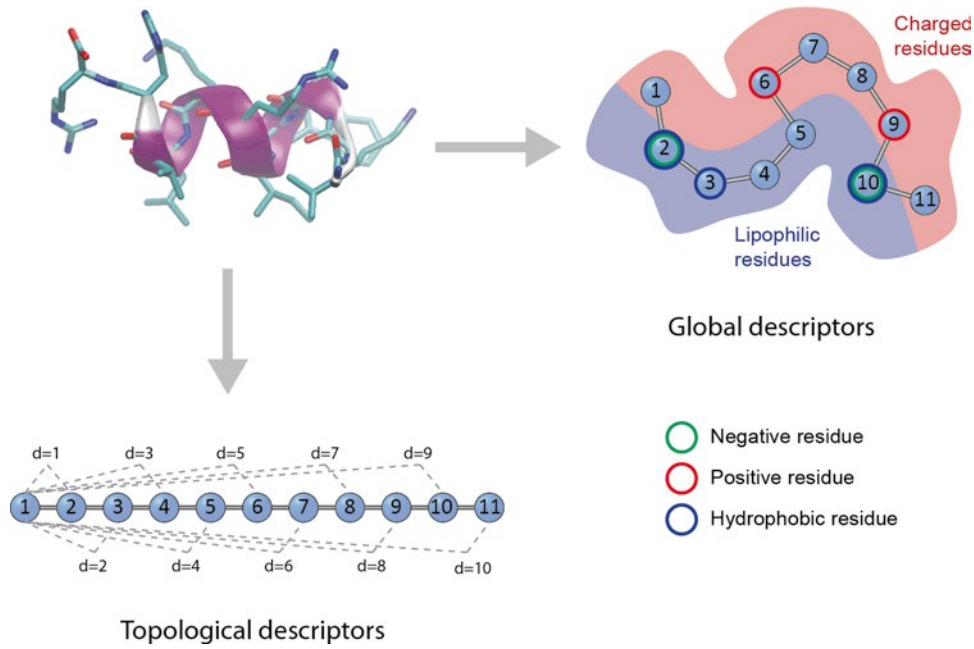


Fig. 2 Schematic representation of the feature selection process. Global and topological features are selected in order to represent the overall chemophysical characteristics and their distribution, respectively

fixed length vector describing each instance is required. Even if each ACC variable is able to represent in a certain measure the amino acidic order along the primary sequence, their effectiveness should be evaluated for every single case. A list of ACC descriptors is summarized in Table 2. Another method to include structure information can be integrated in the model by taking advantage of 3D structure information. Inductive QSAR descriptors are based on the intramolecular steric effects, electronegativities, and intramolecular and intermolecular interaction energies [17]. However, it should be noted that this type of descriptors profoundly depends on AMP structure, and therefore they are not suitable for the analysis of mixed datasets, where different structures are present.

2.3 Prediction Model

The development of novel AMPs and the optimization of known ones, require an understanding of how the activity is correlated to the molecular chemicophysical features. In order to develop such correlations, different statistical models and multivariate approaches can be employed. Advanced methods for data mining can be employed in connection to QSAR variables to quantitatively or qualitatively discriminate between AMP and non-AMP sequences. This paragraph is not meant to be exhaustive about this topic; however, the most important and used techniques are highlighted and discussed.

Table 2
Autocorrelation and cross-correlation descriptors. d is defined as the lag of the autocorrelation; P_i and $P_i + d$ are the normalized properties of the amino acid at position i and $i + d$, respectively

Name	Formula	Description
Normalize Moreau–Broto autocorrelation [16]	$F(d) = \sum_{i=1}^{N-d} P_i \cdot P_{i+d}$	Properties values are used as a measure of spatial autocorrelation
Moran autocorrelation [15]	$F(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}\left(P_i - \bar{P}\right)\left(P_{i+d} - \bar{P}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - \bar{P}\right)^2}$; $\bar{P} = \frac{\sum_{i=1}^{N}(P_i)}{N}$	Property deviations from the average values as a measure of spatial autocorrelation
Geary autocorrelation [14]	$F(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}\left(P_i - P_{i+d}\right)^2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(P_i - \bar{P}\right)^2}$	Square difference of property values as a measure of spatial autocorrelation

Depending on the type of dataset created and the information available, two main categories of models can be distinguished: regression models for a quantitative measure of the biological activity and classification models for the qualitative one (in this case, AMP or non-AMP). The choice of a prediction technique also involves a trade-off between model accuracy and meaningfulness. Linear methods have been widely used in AMPs design because of their simple calculation and interpretation. Principal Component Analysis (PCA) is a mathematical procedure able to transform a number of possibly correlated variables into a smaller number of uncorrelated ones called principal components. Support Vector Machine (SVM) is a linear method where two or more classes are represented in a variable hyperspace and each class is separated by critical boundary instances called *support vectors*. A linear discriminant function is then built to separate each class as widely as possible. On the other hand, nonlinear techniques, like artificial neural networks (ANN), are considered to give better results when the correlation between QSAR descriptors and biological activity is not completely clear. ANN is a mathematical model based on the simulation of some properties of biological neural networks. A network of descriptors is defined as *input nodes* or *neurons*. These nodes are connected together, forming a network that interacts in a hidden layer and sums up into an *output node*. For the purpose of classification, the nonlinear techniques are

considered to give superior results, but at the cost of introducing rather opaque models that cannot easily be used to shed light on the underlying mechanisms involved.

Finally, decision trees are another method to classify an unknown instance in different classes. Each node in the tree represents a particular attribute to test. Unknown instances are routed down the tree according to the values of the attributes tested in successive nodes. The instance is then classified according to the class assigned to the leaf reached. Random Forest (RF) is one of the most popular decision trees in biological data mining, mainly because of two important qualities: high prediction accuracy and information on variable importance for classification [18]. RF is an ensemble recursive partitioning method where many decision trees are trained using subsets of samples and descriptors with replacement. RF has been widely used in AMP prediction and optimization, with performances that compare well to other classification algorithms such as SVM and ANN [12].

RF has several properties that allow extracting relevant trends from data with complex variable relations, which are ubiquitous in datasets generated in the Life Sciences. The classification model can be analyzed a posteriori to infer the similarity between samples, calculated as the number of times the two samples end up in the same terminal node of the tree [19, 20]. In this way, cluster analysis can be applied to identify peptides that have similar features to other AMPs and direct the design to a particular branch of the tree.

After the choice of the statistical model, a required step can be the normalization of the descriptor set. In fact, depending on the chosen descriptors, the scale of values can be varied even of three or more logarithms. Thus, their normalization can help in improving the accuracy of the training. Some classification systems, otherwise, do not require a normalization phase. Generally speaking, Decision Trees are robust enough to handle highly varying variables, while ANN and SVM require for descriptor hyperspace to be normalized. Another consideration is that redundant descriptors can condition the classification performance. Furthermore, in the selection of the descriptors a tradeoff should be found between the performance of the encoding and the requirement of minimizing the number of descriptors. Indeed, on equal terms of performance, a lower number of features is preferable, since the resulting model is less computationally expensive and the interpretation of resulting models is simpler. Therefore, a description selection procedure can be performed using automatic methods, such as genetic algorithms (GA) [21] or iterative methods like Incremental Feature Selection (IFS) [22].

2.4 In Silico Sequence Screening

Once that a sophisticated activity estimator model is constructed, an automatic method for the fast and efficient design and optimization of peptides must be adopted. AMPs design needs to

explore a huge number of amino acidic combinations in order to perform an unbiased analysis of the probability space, and therefore a deterministic approach would be unfeasible. Stochastic optimization methods, like GA or Ant Colony Optimization have been extensively used in virtual peptide design [23, 24]. In particular, GA represents a versatile and powerful tool for AMP design. GA are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. The algorithms follow the principle of Nature adaptive approach to the environment, in which the evolution process is performed by successive generation or mutation and only the fittest individuals resist. Each potential AMP candidate is treated like an entity belonging to a population, and the statistical model is used as a fitness function in order to reflect its biological activity. At the beginning of the selection process, a certain number of random sequences is generated. As the simulation goes on, the population tents to present an increasing average fitness value, until convergence. In AMP design, sometime the simultaneous optimization of one or more conflicting objective is required, like the sequence length or a particular amino acidic composition. Multi-objective evolutionary algorithms (MOEA) are a class of GA, able to optimize different objectives separately. As a result, a list of candidate solutions are screened without favoring one particular objective [25].

2.5 Notes in Statistic-Based AMP Design

- For the model training and validation it is a good habit to have two distinct dataset, one for the training and the other one for the validation. However, when few data are available, the N-fold cross-validation is a good alternative. Basically, the dataset is divided into N parts, where N is usually set to 10. $N-1$ parts are used for the model training, while the remaining is used for validation. This operation is repeated N times and the average of the performance estimator (see below) is computed.
- A good choice of descriptors is imperative for a valid and non-redundant representation of the antimicrobial activity. A mix of global descriptors (describing the overall characteristics of the molecule) and topological descriptors (describing the distribution of them along the sequence) is suggested. Various methods are available for the systematic analysis of descriptor sets. However, one of the most used methods in literature because of its simplicity of use is the Maximum Relevance, Minimum Redundancy (mRMR) method [22], where descriptors are sorted in descending order of importance on the basis of their relevance and redundancy.

- The quality of a classification model can be measured by four parameters: true positive rate for sensitivity, false positive rate for selectivity, predictive accuracy, and MCC, as defined below.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{MCC} = \frac{(TP \cdot TN) - (FN \cdot FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

Where *TP*, *TN*, *FP*, and *FN* are the number of true positive, true negative, false positive, and false negative, respectively, resulting from the model. MCC is an important index used to evaluate the performance of the predictor when the dataset is not balanced. The MCC value ranges from −1 to +1, where a value above 0.5 is considered to be predictive.

For regression models, the Pearson correlation coefficient (PCC) is used as a predictive ability estimator:

$$\text{PCC} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(\Upsilon_i - \bar{\Upsilon})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(\Upsilon_i - \bar{\Upsilon})}}$$

Where X_i and Υ_i are the expected and predicted activity, respectively; N is the number of data points; $\bar{X}$ and $\bar{\Upsilon}$ are the average value of X and Υ, respectively.

3 Molecular Dynamics Simulations OF AMPs

In Force-Field based Molecular Dynamics simulations, atoms in the system are propagated by numerically solving Newton's dynamic equation, with forces described by computationally amenable functions of the coordinates. The set of terms, including covalent interactions (describing bond stretching, angle bending, and dihedral torsion) and non-bonded interactions (electrostatics, hard-core repulsive and dispersive forces), is called the force field. The detail with which each molecule is described can vary from the highest resolution possible in all-atom methods, in which each atom is taken into account, to different degree of coarse grain, in which the atoms are suitably grouped into interaction centers, sometimes also grouping different small

molecules together (for example 3–4 water molecules together). The result of these simulations is a trajectory (a sort of molecular "movie") recording the detailed dynamics of each molecule and how it interacts with the other components.

Current all-atom simulations of molecular systems relevant to this chapter, containing several tens of thousands of atoms, span timescale of hundreds of nanoseconds to some microseconds, the limiting factor being the small time step (1–2 fs) required to integrate Newton's equations of motion, resulting in 10^8–10^9 integration steps to reach these timescales. With coarse-grain force fields the simulation is sped up by 2–3 orders of magnitude thanks to (1) the possibility to use longer time steps (tens of fs) due to elimination of fast degrees of freedom, (2) fictitious speed up of the dynamics due to a smoother potential-energy surface, (3) the reduction in the number of interaction centers (though this is usually compensated by simulating systems of larger sizes).

MD simulations are playing a growing role in elucidating the mechanisms of peptide–bilayer interactions (for recent reviews *see* ref. 26–30). By computing the evolution of suitably prepared initial configurations one can in principle obtain atomic-resolution data on a vast variety of processes. However, due to the empirical nature of molecular mechanics force fields and to the necessarily limited sampling of the configuration space, MD simulations lack "absolute" prediction accuracy, and should be generally validated against experimental findings. Their role should be that of complementing experimental measurements providing the information needed to bridge the gap between the various experimental techniques.

This section provides a brief outline of issues and techniques specific to the MD simulations of AMPs. The reader is assumed to be familiar with the concepts behind MD simulations, such as the molecular mechanics force fields and the algorithms needed to solve Newton's equation of motion. For introductory material *see* ref. 31.

3.1 Force Fields

The force fields commonly employed for biomolecular simulations, and for simulation of AMPs in particular, are AMBER, CHARMM, GROMOS, and OPLS (for reviews and original references *see* refs. 32, 33). Each of these is actually a family of force fields, containing several versions of an original force field, based on a common parameterization strategy. A different version may therefore include extension to different molecules (Charmm36 contains the lipid force field, while the protein and nucleic acid part is that of Charmm27), different parameterization procedures (for example the partial charges in the ff03 Amber force field are obtained starting from a DFT quantum mechanics calculations, rather than the HF in the original version), or modification of certain torsion

terms (for example with respect to Charmm22, Charmm27 contains an additional cross term for backbone torsions).

Validation studies, comparing several different force fields applied to peptide simulations [34–37], have highlighted their strengths and drawbacks. Generally, the latest versions are better at reproducing a series of experimental findings such as peptide helix content, beta-hairpin formation, and NMR chemical shifts and coupling, though caution should be placed in using force fields out of the conditions in which they were parameterized (for example around standard temperature and pressure conditions, 300 K and 1 atm, respectively).

Lipid force fields have been developed in connection to AMBER, CHARMM, and GROMOS. The validation of these force fields is done by trying to reproduce physiochemical properties of the bilayer for different lipid compositions (either homogeneous or mixtures), such as thickness, area per lipid, NMR order parameters (related to the order of the lipid alkyl chains), surface tension, and isothermal area compressibility [38, 39].

Force fields commonly used in peptide/lipid simulations treat electrostatic interactions using fixed partial charges sitting on the atom positions. As such, they do not account for polarization, i.e., the variation in electronic density in response to local electrostatic perturbations. The inclusion of these effects has been pursued for some time, though the use of polarizable force fields is still somewhat limited, due to higher computational costs and absence of extensive benchmarking/validation studies. Existing biomolecular force field accounting for polarization are, among others, Amoeba [40], SIBFA [41], and the polarizable versions of Amber, Amber ff02 [42]. Research is still active on these "next generation" force fields, and inclusion of polarization will be eventually needed to remedy for the deficiencies of additive (i.e., non-polarizable) force fields.

Currently available computational resources limit the size and timescales addressable with all-atoms force fields. An attractive way to speed up the calculations is to reduce the number of degrees of freedom by "coarse graining" (CG) the system, i.e., describing suitably chosen chemical group by single effective interaction centers [43, 44]. Martini is a widely used coarse-grained force field for proteins and lipids [45], which has been specifically applied to peptide–bilayer simulations. The coarse graining in Martini is moderate, in that 3–4 atoms are grouped in "beads", so that single beads are assigned to the smallest amino acids such as Gly and Ala, while four beads are used to describe the biggest such as Tyr or Trp. With Martini the reachable temporal and spatial scales are expanded by 2–3 orders of magnitude, so that simulations of peptide insertion and assembly in the bilayer can be achieved. The disadvantages are that peptide secondary structure needs be assigned a priori, so that no secondary structure change can be simulated.

In addition the grouping of three water molecules in the same bead may conceal the observation of transient water filled pores, and implicit screening of charges may lead to overestimation of the energy required for pore formation [46]. To overcome such drawbacks, multiscale approaches can be adopted, in which the resolution of the system is suitably changed from coarse grain to all-atom and vice versa [47].

3.2 Enhanced Sampling Schemes

Besides coarse graining, other schemes have been devised to overcome the problem of limited conformational sampling in MD simulations. These schemes may exploit collective variables tracing the relevant conformation states (umbrella sampling and metadynamics), or they may facilitate crossing of free-energy barrier through coupling with higher-temperature simulations (parallel tempering).

In Umbrella Sampling [48] a generalized coordinate ξ(R) (also termed collective variable) is defined as function(s) of atom coordinates R. In the context of peptide–bilayer simulations relevant coordinates may be the distance of the peptide center of mass to the bilayer center, or the peptide orientation with respect to the bilayer normal. The sought quantity is the free energy along the generalized coordinate, also called the potential of mean force $W(\xi)$, defined by

$$W\left(\xi\right) = -k_{\mathrm{B}}T\ln\left(\rho\left(\xi\right)\right)$$

where k_B is the Boltzmann constant and $\rho(\xi)$ is the equilibrium distribution of the coordinate. In principle, a sufficiently long simulation would span the relevant configuration space, and from the distribution of ξ one could extract the potential of mean force. However, the presence of free-energy barriers will generally restrain the simulation to limited free-energy basins. The umbrella sampling method forces the sampling of all relevant values of ξ by performing a sort of scan along ξ. This is accomplished by performing several simulations in which an extra term is added to the normal potential energy of the molecular system. This term may have the form

$$U = k\left(\xi(R) - \bar{\xi}_i\right)^2$$

where $\bar{\xi}_i$ are successive values of ξ, and k is a spring constant. In the case of the peptide–bilayer distance, the $\bar{\xi}_i$ may be suitably spaced value from 0 nm (peptide completely immersed in the bilayer) to 6 nm or more (peptide in the bulk solvent). For each window, $W(\xi)$ is obtained from the biased distribution of ξ during the MD simulation, $\rho_{Ui}(\xi)$, by

$$W_i\left(\xi\right) = -k_{\mathrm{B}}T\ln\left(\rho_{Ui}\left(\xi\right)\right) - k\left(\xi - \bar{\xi}_i\right)^2 + \Delta_i$$

where Δ_i are unknown constants that may be found by matching together the various segment of $W(\xi)$. Clearly, for each simulation i the values of ξ will be restrained around $\bar{\xi}_i$. However, provided that there is enough overlapping between the explored values of ξ, the continuous profile of $W(\xi)$ can be reconstructed automatically through, for example, the weighted histogram analysis method (WHAM) [49].

Though it is possible to perform multidimensional umbrella sampling, the number of needed simulation windows grows rapidly for two- and three-dimensional scans. In addition, a lot of computational time may be spent in "uninteresting" windows of ξ. The metadynamics approach [50], albeit less accurate than WHAM, at least in the original formulation, is both more amenable for treating multi collective variables and "self" regulating in the time spent exploring the various regions in the conformational space. The idea behind metadynamics is to perform an MD simulation where the system is "discouraged" to explore the same free-energy regions (described by the set of collective variables) by adding a history dependent potential that gradually fills the free-energy basins. In the original formulation, the potential energy is modified by periodically adding Gaussian functions with suitably chosen heights and widths, and centered on the current values of the ξ_i. The process is repeated until free diffusion in the collective variable space is achieved. The (one-dimensional or multidimensional) free-energy profile is then obtained as the negative of the sum of all added Gaussians. Several variants were based on the same idea of a history dependent potential: local elevation [51], conformational flooding [52], adaptively biased molecular dynamics [53], among others.

A common issue with both umbrella sampling and metadynamics methods is that they assume that the degrees of freedom orthogonal to the chosen collective variables be sufficiently sampled, i.e., that the relaxation times of these degrees of freedom are shorter than the time spent in each "bin" of the free energy surface. Through careful choice of the collective variables in multidimensional scans these problems can be alleviated, but still "hidden" variables coupled to the relevant reaction coordinate may play important roles. In lipid-membrane studies, a typical indicator of poor sampling in umbrella sampling simulations is the hysteresis between, for example, insertion of the peptide in the bilayer and extraction [54].

Parallel tempering, also known as replica exchange [55], enables free-energy barrier crossing by coupling the simulation at the desired temperature with higher-temperature simulations. This coupling is accomplished by exchanging the coordinates among the replica following a Metropolis scheme. More in detail, n replicas are evolved through MD, each maintained at a temperature T_i. After a number of MD steps an exchange between the coordinates

of replica at T_i and T_{i+1} (the higher successive temperature in the ladder) is performed with a probability given by

$$p = \min\left(1, e^{(E_i - E_{i+1})\left(\frac{1}{k_B T_i} - \frac{1}{k_B T i+1}\right)}\right)$$

i.e., the exchange is performed with probability 1 if E_{i+1} (the potential energy of replica $i+1$) is lower than E_i; otherwise it is performed with a probability given by the exponential term in the previous equation. This probabilistic exchange ensures the detailed balance condition, and that the MD at each temperature samples a canonical ensemble.

In a typical replica exchange molecular dynamics (REMD) simulation a set of temperatures from $T_0 = 300$ K to $T_n = 600$–900 K is chosen, and exchanges are attempted each 50–500 time steps. The spacing between successive replicas should be such as to allow for a 10–40 % successful exchange, implying a suitable overlapping between potential energies distributions at different temperatures. Unfortunately, these distributions become narrower at increasing number of degrees of freedom, and for systems of tens or hundreds thousands atoms an unfeasible number of replica is needed (>500). A solution to this problem is provided by the so-called Hamiltonian replica exchange (HREX) [56, 57] in which only a subsystem is "heated," by actually scaling its potential energy function. For example, one may choose to "heat" only the peptides, or peptides and bilayer: without the solvent degrees of freedom the number of replica is greatly reduced.

The schemes described above can be coupled together, and their simultaneous application may ensure both a sufficient sampling on the chosen coordinate, via Umbrella Sampling or Metadynamics, and rapid relaxation for the orthogonal degrees of freedom through REMD [58].

3.3 Issues with Peptide–Bilayer Simulations

This subsection schematically lists some of the points needing particular care in peptide–bilayer simulations. Explicit solvent simulations with periodic boundary conditions are assumed.

- Choice of the membrane model. With respect to the cellular membrane, the simulated systems contain only few lipid components and no protein or carbohydrate. They are closer to experimental studies involving artificial bilayers, with controlled lipid composition. However, for the peptide shown to destabilize the bilayer in artificial vesicles, simple bilayer models may for the most part be appropriate.
- The size of the simulated bilayer patch should be carefully chosen. Some peptide may act by selectively modifying the surface tension of the outer leaflet of the bilayer, thereby inducing curvature [33]. Such effects may be hidden by the use of periodic boundary conditions if the bilayer patch is too small.

- Different ensembles may be used in MD simulations. The microcanonical ensemble, NVE (constant energy, temperature, and particle number), is rarely used because it does not allow for temperature control and volume fluctuation. NVT, or constant temperature, is the ensemble of choice when simulating biomolecules in aqueous solvent. Several algorithms have been proposed to approximate the canonical ensemble, which may rely on stochastic terms or on the introduction of fictitious degree of freedom representing heat bath. In the isothermal–isobaric ensemble (NPT), pressure is controlled by suitably scaling the atomic coordinates, thereby changing the total volume. This can be accomplished with the Berendsen algorithm, the Nosé–Hoover Langevin piston [59], or the Parrinello–Rahman method [60]. Bilayer simulations frequently use semi-isotropic pressure schemes, in which the control of pressure on the orientation normal to the bilayer is separated from the other two dimensions, i.e., the scaling in the lateral directions is independent from the one in the normal. This decoupling is required because of the different compressibility of water and of the bilayer. Clearly, choosing a constant volume ensemble fixes the area-per-lipid value, and this may correct for force-field artifacts. However, insertion of the peptides into the bilayer may require significant rearrangements for which flexibility in the lateral directions may be more realistic.
- Non-bonded interactions in principle require infinite summation over the pairs of particles in the periodic cells. For short-ranged potential, such as the r^{-6} attractive tail of the Lennard–Jones potential, cutoff schemes are used, i.e., only the particles within a certain distance (cutoff) are accounted for. Coulomb interactions are long ranged, so a cutoff approach (truncation schemes) may be too crude an approximation, leading to potential inconsistent behavior, such as artificial ordering [61]. The method of choice is the so-called Particle Mesh Ewald, or PME, in which the interaction is separated into a short-range and long-range part, calculated separately, the first using a cutoff scheme, the second by spreading the charges on a 3D grid and accounting for all the periodic images. PME needs an overall neutral system, and failing to add neutralizing counterions may lead to serious artifact such as thinning of the bilayer [54]. The cutoff to be used for both Lennard–Jones and Coulomb interactions can vary in the range 8–12 Å, and is force-field dependent. In addition MD codes such as GROMACS or NAMD employ smoothing schemes to avoid the discontinuity at the cutoff. This schemes and the cutoff distance need be carefully tuned, and discussions on the topics can be found in the literature [38].

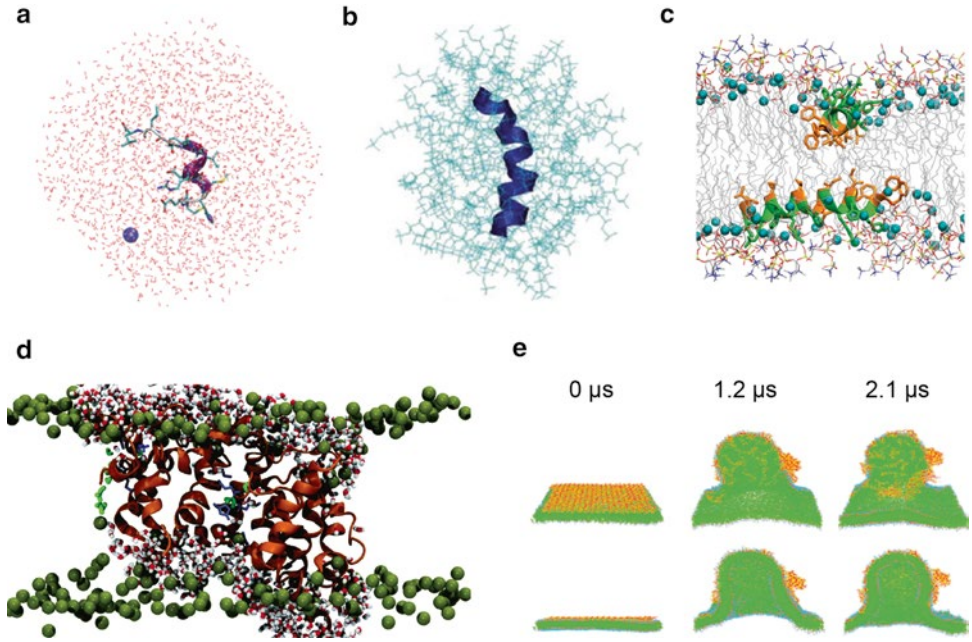


Fig. 3 MD simulation of AMPs. (**a**) Snapshots from all-atom simulation of a peptide in water and (**b**) in a micelle (**c**) Snapshot from all-atom simulation of AMP Piscidin 1 and Piscidin 3 in lipid bilayers. Reprinted (adapted) with permission from ref. 68. Copyright 2011 American Chemical Society. (**d**) All-atom MD simulation of pore formation by a cluster of 16 Maculatin 1.1 peptides (*orange*) in a lipid bilayer. Reprinted (adapted) with permission from ref. 69, copyright (2012) American Chemical Society (**e**) Snapshots from coarse grained simulations of several Magainine 2 peptides (*orange*) placed on one side of a pure DPPC bilayer (in *green*). Reprinted with permission from ref. 67. Copyright 2011 American Chemical Society

3.4 Systems and Processes

MD simulations can be used to predict structural properties of the peptide in different environments, such as, in order of complexity, water, organic solvents and water–organic solvent mixtures, micelles, and lipid bilayer (*see* Fig. 3). Simulation of the monomeric AMPs, though not directly targeting their mechanism of action on the cellular membrane, are useful for extracting information such as presence and stability of secondary structure motives (alpha-helix, beta-sheets, and turns), and other physicochemical characteristics such as solvent exposed surface. These quantities may be then related to antimicrobial activity and toxicity of various examined peptide sequences through multiple linear regression algorithms, such as in quantitative structure–activity relationship (QSAR) studies.

Structure predictions of peptides in solvents, water in particular, are achievable with an adequate degree of confidence, given that the force fields have been rather extensively tried and optimized for such tasks, as long as standard pressure and temperature (around 1 atm and 300 K, respectively) conditions are considered. In addition, the limited number of degrees of freedom of peptides allows for rather exhaustive sampling of their configuration space,

at least with the enhanced sampling techniques mentioned below. Furthermore for this kind of systems the simulation protocols are rather robust and well established. Water–organic solvent mixtures can be used to assess structural properties in various environments. For example, MD simulations in pure water and water–TFE mixtures (TFE, or 2,2,2-trifluoroethanol, provides a low dielectric environment partially mimicking the conditions inside the bilayer) have been used to assist bioinformatics algorithms in designing novel AMP sequences [20]. A more realistic model of the environment inside lipid bilayer is provided by micelles, self-assembled structures of amphipathic molecules, with a highly hydrophobic interior and anionic or zwitterionic heads exposed to the solvent. Micelles mimicking the bilayer are used in NMR experiments, because of their faster relaxation times. Simulations of AMPs in micelles are at an intermediate level of complexity between those in solvent and lipid bilayer (*see* ref. 62).

The insertion of a peptide in the bilayer is a prohibitively slow process for all-atom MD simulations, and coarse-grained force fields or enhanced sampling techniques need to be used (see below). Studies of peptide structure and position inside the membrane may therefore start from an initial configuration where the peptide is already embedded in the bilayer. A different approach consists in starting from a non-formed bilayer, i.e., from a random mixture of water molecules and lipids, which are known to form a bilayer in tens to hundreds of nanoseconds [63]. In this way the self-assembly of the bilayer is simulated and the position of the peptide is not biased toward the starting configuration thanks to the high fluidity of the system during the self-assembly process.

Simulating the aggregation of several peptides in the bilayer is yet a more ambitious goal, because also the relative configurations of the various peptides need to be sampled. In MD studies of peptide aggregation in the bilayer the aggregates may be preassembled to study their stability and function, or the self-assembly process itself may be pursued. For example, different putative structure of a pore may be tried, and the stable ones be selected as the most probable structures [64], or peptides may be inserted at unbiased position in the bilayer and aggregation and pore formation be observed [65]. With CG force fields the whole process of peptide adsorption and pore formation [66, 67], in systems of thousands of peptides and lipid patches of lateral dimension up to 0.1 μm [67].

4 Experimental Validation of Amps: Minimal Inhibitory Concentration (MIC)

The in vitro activity of AMPs is tested using the Microtiter Broth Dilution Method in order to determine the MIC value, as recommended for the antibiotic testing by the NCLSS (National Committee of Laboratory Safety and Standards) [70].

Here, we suggest a modified version of this method as recommended by R. E. W. Hancock (University of British Columbia, Vancouver, British Columbia, Canada) for testing antimicrobial peptides (http: //www.interchg.ubc.ca/bobh/MIC.htm).

4.1 Materials

1. Sterile tubes (15 mL).
2. Mueller Hinton Broth (MHB).
3. Mueller Hinton agar plates (MHA).
4. Sterile 96-well polypropylene microtiter plates.
5. Polypropylene microcentrifuge tubes.
6. Sterile petri dishes.
7. sterile deionized water (dH_2O).

4.2 Methods

1. Inoculate 5 mL MHB in tubes with test strains from MHA plates and grow overnight at 37 °C on a shaker (160 rpm).
2. Make serial dilutions of test peptides in sterile deionized water in polypropylene tubes:
   - Dissolve test peptide in dH_2O at ten times the required maximal concentration;
   - Do twofold dilutions in dH_2O to get serial dilutions of peptides at ten times the required test concentrations, e.g., 640, 320, 160, …, 2.5 μg/mL.
3. Dilute overnight bacterial cultures in MHB to give 5×10^5 colony forming units/mL.
4. Dispense 90 μL of bacterial suspension in each well from column 1 to column 11. Do not add bacteria to column 12, and instead dispense 100 μL of MHB (sterility control and blank for the plate scanner).
5. Add 10 μL of 10× test peptide each well from column 1 to column 10 (column 11 is a control for bacteria alone, with no peptide, where 10 μL of dH_2O is added).
6. Incubate the plates at 37 °C for 18–24 h.
7. MIC can be taken as the lowest concentration of drug that reduces growth by more than 50 %.
8. Plate l0 μL 10^{-6} dilution of overnight cultures on MHA plates to determine a viable count. The MBC (Minimal bactericidal concentration) can be determined by plating out the contents of the first three wells showing no visible growth of bacteria onto MHA plates and incubating at 37 °C for 18 h. MBC is defined as the lowest concentration of the peptide causing a reduction in the numbers of viable bacteria of $\geq 3\log_{10}$ with respect to the CFU/mL inoculated.

4.3 Notes

- It is important that you use the material mentioned above. For example, do not substitute polystyrene for polypropylene tubes or microtiter plates. Cationic peptides bind polystyrene (especially "tissue culture treated" polystyrene).

References

1. Di Luca M, Maccari G, Nifosì R (2014) Treatment of microbial biofilms in the post antibiotic era: prophylactic and therapeutic use of antimicrobial peptides and their design by bioinformatics tools. Pathog Dis. http://www.ncbi.nlm.nih.gov/pubmed/24515391. Accessed 14 Feb 2014
2. Salomone F, Cardarelli F, Signore G, Boccardi C, Beltram F (2013) In vitro efficient transfection by CM18-Tat11 hybrid peptide: a new tool for gene-delivery applications. PLoS One. doi:10.1371/journal.pone.0070108
3. Bahar A, Ren D (2013) Antimicrobial peptides. Pharmaceuticals 6:1543–1575. http://www.mdpi.com/1424-8247/6/12/1543/. Accessed 29 Nov 2013
4. Shai Y, Oren Z (2001) From "carpet" mechanism to de-novo designed diastereomeric cell-selective antimicrobial peptides. Peptides 22: 1629–1641. http://www.ncbi.nlm.nih.gov/pubmed/11587791. Accessed 29 Dec 2012
5. Fjell CD, Hancock REW, Cherkasov A (2007) AMPer: a database and an automated discovery tool for antimicrobial peptides. Bioinformatics 23:1148–1155. http://www.ncbi.nlm.nih.gov/pubmed/17341497. Accessed 10 May 2013
6. Rathinakumar R, Wimley WC (2008) Biomolecular engineering by combinatorial design and high-throughput screening: small, soluble peptides that permeabilize membranes. J Am Chem Soc 130:9849–9858. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2582735&tool=pmcentrez&rendertype=abstract. Accessed 22 May 2013
7. Marks JR, Placone J, Hristova K, Wimley WC (2011) Spontaneous membrane-translocating peptides by orthogonal high-throughput screening. J Am Chem Soc 133:8995–9004. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3118567&tool=pmcentrez&rendertype=abstract. Accessed 3 Jan 2013
8. Wang P, Hu L, Liu G, Jiang N, Chen X et al (2011) Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PLoS One 6: e18476. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3076375&tool=pmcentrez&rendertype=abstract. Accessed 15 Mar 2012
9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419&tool=pmcentrez&rendertype=abstract. Accessed 21 Jan 2014
10. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. http://www.ncbi.nlm.nih.gov/pubmed/16731699. Accessed 30 July 2012
11. Piotto SP, Sessa L, Concilio S, Iannelli P (2012) YADAMP: yet another database of antimicrobial peptides. Int J Antimicrob Agents 39:346–351. http://www.ncbi.nlm.nih.gov/pubmed/22325123. Accessed 23 Aug 2012
12. Thomas S, Karnik S, Barai RS, Jayaraman VK, Idicula-Thomas S (2010) CAMP: a useful resource for research on antimicrobial peptides. Nucleic Acids Res 38: D774–D780. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808926&tool=pmcentrez&rendertype=abstract. Accessed 23 Aug 2012
13. Wold S, Jonsson J, Sjörström M, Sandberg M, Rännar S (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 277:239–253. http://linkinghub.elsevier.com/retrieve/pii/000326709380437P. Accessed 23 July 2012
14. Sokal RR, Thomson BA (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. Am J Phys Anthropol 129:121–131. http://www.ncbi.nlm.nih.gov/pubmed/16261547. Accessed 13 Feb 2014
15. Horne DS (1988) Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. Biopolymers 27:451–477. http://www.ncbi.nlm.nih.gov/pubmed/3359010. Accessed 13 Feb 2014
16. Feng ZP, Zhang CT (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 19:269–275. http://www.ncbi.nlm.nih.gov/pubmed/11043931. Accessed 13 Feb 2014
17. Jaiswal K, Naik PK (2008) Distinguishing compounds with anticancer activity by ANN using

inductive QSAR descriptors. Bioinformation 2:441–451. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2561164&tool=pmcentrez&rendertype=abstract. Accessed 13 Feb 2014

18. Michaelson JJ, Sebat J (2012) forestSV: structural variant discovery through statistical learning. Nat Methods 9:819–821. http://www.ncbi.nlm.nih.gov/pubmed/22751202. Accessed 24 Aug 2012
19. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J et al (2012) Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Brief Bioinform. http://www.ncbi.nlm.nih.gov/pubmed/22786785. Accessed 17 July 2012
20. Maccari G, Di Luca M, Nifosí R, Cardarelli F, Signore G, et al (2013) Antimicrobial peptides design by evolutionary multiobjective optimization. PLoS Comput Biol 9: e1003212. http://www.ploscompbiol.org/article/metrics/info:doi/10.1371/journal.pcbi.1003212. Accessed 23 Sept 2013
21. Hansen L, Lee EA, Hestir K, Williams LT, Farrelly D (2009) Controlling feature selection in random forests of decision trees using a genetic algorithm: classification of class I MHC peptides. Comb Chem High Throughput Screen 12: 514–519. http://www.ncbi.nlm.nih.gov/pubmed/19519331. Accessed 14 Feb 2014
22. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. http://www.ncbi.nlm.nih.gov/pubmed/16119262. Accessed 23 July 2012
23. Hiss JA, Bredenbeck A, Losch FO, Wrede P, Walden P et al (2007) Design of MHC I stabilizing peptides by agent-based exploration of sequence space. Protein Eng Des Sel 20:99–108. http://www.ncbi.nlm.nih.gov/pubmed/17314106. Accessed 14 Feb 2014
24. Fjell CD, Jenssen H, Cheung WA, Hancock REW, Cherkasov A (2011) Optimization of antibacterial peptides by genetic algorithms and cheminformatics. Chem Biol Drug Des 77:48–56. http://www.ncbi.nlm.nih.gov/pubmed/20942839. Accessed 25 May 2012
25. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6:182–197. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=996017. Accessed 14 July 2012
26. Bocchinfuso G, Bobone S, Mazzuca C, Palleschi A, Stella L (2011) Fluorescence spectroscopy and molecular dynamics simulations in studies on the mechanism of membrane destabilization by antimicrobial peptides. Cell Mol Life Sci 68:2281–2301. http://www.ncbi.nlm.nih.gov/pubmed/21584808. Accessed 6 Aug 2013
27. Gurtovenko AA, Anwar J, Vattulainen I (2010) Defect-mediated trafficking across cell membranes: insights from in silico modeling. Chem Rev 110: 6077–6103. http://www.ncbi.nlm.nih.gov/pubmed/20690701. Accessed 7 Aug 2013
28. Marrink SJ, de Vries AH, Tieleman DP (2009) Lipids on the move: simulations of membrane pores, domains, stalks and curves. Biochim Biophys Acta 1788:149–168. http://www.ncbi.nlm.nih.gov/pubmed/19013128. Accessed 7 Aug 2013
29. Bolintineanu DS, Kaznessis YN (2011) Computational studies of protegrin antimicrobial peptides: a review. Peptides 32:188–201. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013618&tool=pmcentrez&rendertype=abstract. Accessed 7 Aug 2013
30. Chen L, Gao L (2012) How the antimicrobial peptides kill bacteria: computational physics insights. Commun Comput Phys. http://www.global-sci.com/issue/abstract/readabs.php?vol=11&page=709&issue=3&ppage=725&year=2012. Accessed 7 Aug 2013
31. Leach A (2001) Molecular modelling: principles and applications, 2nd edn. Prentice Hall, NJ
32. Ponder JW, Case DA (2003) Force fields for protein simulations. In: Daggett V (ed) Protein simulations, vol 66. Academic, New York, pp 27–85. doi:10.1016/S0065-3233(03)66002-X
33. Mackerell AD (2004) Empirical force fields for biological macromolecules: overview and issues. J Comput Chem 25:1584–1604. doi:10.1002/jcc.20082
34. Lange OF, van der Spoel D, de Groot BL (2010) Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. Biophys J 99:647–655. doi:10.1016/j.bpj.2010.04.062
35. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO et al (2012) Systematic validation of protein force fields against experimental data. PLoS One 7: e32131. http://dx.plos.org/10.1371/journal.pone.0032131. Accessed 21 May 2013
36. Beauchamp KA, Lin Y-S, Das R, Pande VS (2012) Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. J Chem Theory Comp 8: 1409–1414. doi:10.1021/ct2007814
37. Cino EA, Choy W-Y, Karttunen M (2012) Comparison of secondary structure formation

using 10 different force fields in microsecond molecular dynamics simulations. J Chem Theory Comp 8:2725–2740. doi:10.1021/ct300323g

38. Piggot TJ, Piñeiro Á, Khalid S (2012) Molecular dynamics simulations of phosphatidylcholine membranes: a comparative force field study. J Chem Theory Comp 8:4593–4609. doi:10.1021/ct3003157
39. Jämbeck JPM, Lyubartsev AP (2012) An extension and further validation of an all-atomistic force field for biological membranes. J Chem Theory Comp 8:2938–2948. doi:10.1021/ct300342n
40. Shi Y, Xia Z, Zhang J, Best R, Wu C et al (2013) The polarizable atomic multipole-based AMOEBA force field for proteins. J Chem Theory Comp 9:4046–4063. doi:10.1021/ct4003702
41. Guo H, Gresh N, Roques BP, Salahub DR (2000) J Phys Chem B 104:9746–9754
42. Cieplak P, Caldwell J, Kollman P (2001) Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/. J Comput Chem 22:1048–1057. doi:10.1002/jcc.1065
43. Tozzini V (2005) Coarse-grained models for proteins. Curr Opin Struct Biol 15:144–150. http://www.ncbi.nlm.nih.gov/pubmed/15837171. Accessed 3 June 2013
44. Baaden M, Marrink SJ (2013) Coarse-grain modelling of protein-protein interactions. Curr Opin Struct Biol 23:878–886. doi:10.1016/j.sbi.2013.09.004
45. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP et al (2008) The MARTINI coarse-grained force field: extension to proteins. J Chem Theory Comp 4:819–834. doi:10.1021/ct700324x
46. Bennett WFD, Tieleman DP (2011) Water defect and pore formation in atomistic and coarse-grained lipid membranes : pushing the limits of coarse graining. J Chem Theory Comp 12:2981–2988
47. Ayton GS, Noid WG, Voth GA (2007) Multiscale modeling of biomolecular systems: in serial and in parallel. Curr Opin Struct Biol 17:192–198. doi:10.1016/j.sbi.2007.03.004
48. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. J Comput Phys 23:187–199. doi:10.1016/0021-9991(77)90121-8
49. Roux B (1995) The calculation of the potential of mean force using computer simulations. Comput Phys Commun 91:275–282. doi:10.1016/0010-4655(95)00053-I
50. Laio A, Parrinello M (2002) Escaping free-energy minima. Proc Natl Acad Sci U S A 99:12562–12566. doi:10.1073/pnas.202427399
51. Huber T, Torda AE, Gunsteren WF (1994) Local elevation: a method for improving the searching properties of molecular dynamics simulation. J Comput Aided Mol Des 8:695–708. doi:10.1007/BF00124016
52. Grubmüller H (1995) Predicting slow structural transitions in macromolecular systems: conformational flooding. Phys Rev E Stat Plasmas Fluids Retat Interdiscip Topics 52:2893–2906. doi:10.1103/PhysRevE.52.2893
53. Adamson S, Kharlampidi D, Dementiev A (2008) Stabilization of resonance states by an asymptotic Coulomb potential. J Chem Phys 128:024101. doi:10.1063/1.2821102
54. Yesylevskyy S, Marrink S-J, Mark AE (2009) Alternative mechanisms for the interaction of the cell-penetrating peptides penetratin and the TAT peptide with lipid bilayers. Biophys J 97: 40–49. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2711361&tool=pmcentrez&rendertype=abstract. Accessed 6 Aug 2013
55. Sugita Y, Yuko Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151
56. Sugita Y, Okamoto Y (2000) Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. Chem Phys Lett 329:261–270
57. Wang L, Friesner RA, Berne BJ (2011) Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). J Phys Chem B 115:9431–9438. doi:10.1021/jp204407d
58. Bussi G, Gervasio FL, Laio A, Parrinello M (2006) Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. J Am Chem Soc 128:13435–13441. doi:10.1021/ja062463w
59. Feller SE, Zhang Y, Pastor RW, Brooks BR (1995) Constant pressure molecular dynamics simulation: the Langevin piston method. J Chem Phys 103:4613. doi:10.1063/1.470648
60. Parrinello M (1981) Polymorphic transitions in single crystals: a new molecular dynamics method. J Appl Phys 52:7182. doi:10.1063/1.328693

61. Patra M, Karttunen M, Hyvönen MT, Falck E, Vattulainen I (2004) Lipid bilayers driven to a wrong lane in molecular dynamics simulations by subtle changes in long-range electrostatic interactions. J Phys Chem B 108:4485–4494. doi:10.1021/jp031281a
62. Langham A, Kaznessis YN (2010) Molecular simulations of antimicrobial peptides. Methods Mol Biol 618:267–285. doi:10.1007/978-1-60761-594-1_17
63. Venturoli M, Smit B (1999) Simulating the self-assembly of model membranes. Phys Chem Comm 2:45. doi:10.1039/a906472i
64. Peter Tieleman D, Hess B, Sansom MSP (2002) Analysis and evaluation of channel models: simulations of alamethicin. Biophys J 83:2393–2407. http://linkinghub.elsevier.com/retrieve/pii/S0006349502752533. Accessed 7 Aug 2013
65. Thøgersen L, Schiøtt B, Vosegaard T, Nielsen NC, Tajkhorshid E (2008) Peptide aggregation and pore formation in a lipid bilayer: a combined coarse-grained and all atom molecular dynamics study. Biophys J 95:4337–4347. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567951&tool=pmcentrez&rendertype=abstract. Accessed 7 Aug 2013
66. Gkeka P, Sarkisov L (2009) Spontaneous formation of a barrel-stave pore in a coarse-grained model of the synthetic LS3 peptide and a DPPC lipid bilayer. J Phys Chem B 113:6–8. doi:10.1021/jp808417a
67. Woo H-J, Wallqvist A (2011) Spontaneous buckling of lipid bilayer and vesicle budding induced by antimicrobial peptide magainin 2: a coarse-grained simulation study. J Phys Chem B 115:8122–8129. http://www.ncbi.nlm.nih.gov/pubmed/21651300. Accessed 7 Aug 2013
68. Perrin BS, Tian Y, Fu R, Grant C V, Chekmenev EY et al (2014) High-resolution structures and orientations of antimicrobial peptides piscidin 1 and piscidin 3 in fluid bilayers reveal tilting, kinking, and bilayer immersion. J Am Chem Soc. http://www.ncbi.nlm.nih.gov/pubmed/24410116. Accessed 14 Feb 2014
69. Parton DL, Akhmatskaya EV, Sansom MSP (2012) Multiscale simulations of the antimicrobial peptide maculatin 1.1: water permeation through disordered aggregates. J Phys Chem B 116:8485–8493. doi:10.1021/jp212358y
70. National Committee for Clinical Laboratory Standards (2000) Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically; M7-A5, 5th edn. National Committee for Clinical Laboratory Standards, Wayne, PA

Chapter 10

Information-Driven Modeling of Protein-Peptide Complexes

Mikael Trellet, Adrien S.J. Melquiond, and Alexandre M.J.J. Bonvin

Abstract

Despite their biological importance in many regulatory processes, protein-peptide recognition mechanisms are difficult to study experimentally at the structural level because of the inherent flexibility of peptides and the often transient interactions on which they rely. Complementary methods like biomolecular docking are therefore required. The prediction of the three-dimensional structure of protein-peptide complexes raises unique challenges for computational algorithms, as exemplified by the recent introduction of protein-peptide targets in the blind international experiment CAPRI (Critical Assessment of PRedicted Interactions). Conventional protein-protein docking approaches are often struggling with the high flexibility of peptides whose short sizes impede protocols and scoring functions developed for larger interfaces. On the other side, protein-small ligand docking methods are unable to cope with the larger number of degrees of freedom in peptides compared to small molecules and the typically reduced available information to define the binding site. In this chapter, we describe a protocol to model protein-peptide complexes using the HADDOCK web server, working through a test case to illustrate every steps. The flexibility challenge that peptides represent is dealt with by combining elements of conformational selection and induced fit molecular recognition theories.

Key words Biomolecular interactions, Information-driven docking, Conformational changes, Flexibility, HADDOCK, Molecular modeling

1 Introduction

A large variety of methods are available to scientists to investigate the 3D structure of biomolecular complexes. Experimental determination of protein-peptide complexes is, however, often nontrivial due to the dynamic nature of the transient interactions they mediate. While X-ray crystallography is struggling with the high flexibility of peptides, hybrid approaches that rely on an experimental characterization of the binding site (NMR, cross-linking mass spectrometry …) and/or NMR-derived restraints to limit the conformational space of the peptide (e.g. dihedral angle restraints), in combination with computational modeling, have demonstrated

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_10,

their accuracy for various protein-peptide systems [1–6]. Structural characterisation of low affinity interactions remain unfortunately out of reach for most experimental methods. There is therefore a need for improving existing computational methods.

Modeling of protein-protein complexes has a long-standing history that started back in the late 1970s with the first automated computer analysis of protein-protein interactions [7]. Macromolecular docking made its first proof of concept with the successful prediction of the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase in 1996 [8]. Protein-peptide interactions, in contrast, have only been studied computationally recently. The recognition mechanisms underlying their assembly are still debated [9–12]. Flexibility is a key characteristic of peptides, which are short polypeptidic chains ranging from 5 to 30 amino acids and, in most cases, do not adopt a well-defined conformation when unbound, i.e., in their free state. This represents a major challenge for classical docking algorithms where both constituents are usually treated as rigid in first instance, to be refined at later stages, allowing some degrees of flexibility at the interface.

Over the last years, a number of new algorithms or adaptations of existing docking methods have been released to address the unique challenges raises by protein-peptide interactions [13–21]. Based on the HADDOCK framework [22], we have developed an original approach that combines ensemble docking and enhanced flexibility to improve the sampling of peptides [23]. HADDOCK is an information-driven docking software [24] using CNS (Crystallography and NMR system) [25, 26] as computational engine and the OPLS united atom force field [27] to calculate the non bonded interactions (with a cutoff of 8.5 Å). It allows the integration of a variety of experimental data to drive the docking process, such as NMR chemical shift perturbation and mutagenesis data. HADDOCK also introduces flexibility into the subunits during the docking process, ending with a final refinement of the models in explicit solvent. Currently, HADDOCK is one of the most cited docking software [28], counts a large community of 3,700+ users worldwide, and ranks among the best performing docking methods based on CAPRI (Critical Assessment of PRediction of Interactions) [29], a community wide experiment where participants have a limited time to predict the structure of a complex given only the structures, sometimes even only the sequences, of its free constituents.

We have recently optimized HADDOCK's protocol for protein-peptide docking against a benchmark of 101 protein-peptide complex structures, achieving a remarkable overall performance when starting from unbound structures [23]. In this chapter, we describe step by step this protocol using the HADDOCK web server.

2 Theory

This section describes the different steps and their background in order to perform a protein-peptide docking run and achieve the overall best performances with HADDOCK.

2.1 Peptide Conformation Sampling

Unlike protein-protein docking, we usually do not have access either to the free form of the peptide or to any structural template that could be used to generate a reliable starting 3D model of the structure of the peptide. To solve this problem, we have proposed a specific protocol for flexible docking of short peptides (5–15 amino acids) that starts from an ensemble of three different conformations of the peptide (α-helix, polyproline-II, and extended—*see* Subheading 3 for more details about how to generate this ensemble). This canonical ensemble does not aim at discretizing the conformational space sampled by the free peptide, but rather represents conformations often observed in protein-peptide complexes. Indeed, taken together, these three conformations cover about 80 % of the observed peptide-bound structures in the Protein Data Base [30]. Building onto the ensemble docking capability of HADDOCK, protein-peptide docking can start from these three distinct conformations and, hopefully, select be best suited peptide conformation for the complex under study, following a conformational selection mechanism.

2.2 Interface Restraints

HADDOCK uses ambiguous and unambiguous restraints throughout the entire docking process to drive the complex formation (*see* Subheading 2.3 for more details). These restraints can be derived from various experimental information sources such as NMR chemical shifts perturbations, hydrogen/deuterium exchange, chemical cross-linking detected by mass spectrometry, mutagenesis … [31, 32]. All this information is usually translated into distance or angle restraints used both for sampling and scoring. In this protocol we describe a classical scenario in which no information is available about which residues of the peptide are involved in binding, treating it as fully "passive," which means peptide residues can make contacts but no penalty will be paid if they do not. One the protein side we define a large surface centred on the native interface. For each docking trial, we randomly select half of the so-called active residues that belong to this surface (making the assumption that they are directly involved in the binding) and define ambiguous restraints toward the peptide.

2.3 Protein-Peptide HADDOCKing

2.3.1 Docking Protocol

The docking protocol in HADDOCK consists of three successive steps:

- *it0*: Rigid-body energy minimization (RBEM)
- *it1*: Semiflexible simulated annealing (SA) in torsion angle space (TAD/SA)
- *Water*: Final restrained molecular dynamics in explicit solvent

Pre- and post-processing steps are performed: (1) to build missing atoms in the preliminary step and (2) to launch energetic, intermolecular, and restraint analyses in the final step. For further details please refer to [22, 24].

One critical aspect in protein-peptide recognition is the importance of long-range electrostatic interactions [17]. Therefore, the user should specify charged Cter and Nter (default in HADDOCK) when working with naturally occurring peptides or uncharged termini when the peptide is a fragment of protein or capped in the experiment, this to avoid undesired interaction with the termini in the latter case (*see* **Note 1**).

Rigid-Body Energy Minimization (RBEM, it0)

In this initial docking stage, the interacting partners are first separated in space and randomly rotated around their respective center of mass. As a result, the starting positions of peptides adopt a spherical distribution around the protein receptor. The number of models generated in this step should typically be increased from the default 1,000–6,000, to ensure that each of the three distinct peptide conformations from the canonical ensemble is sampled 2,000 times. The resulting models are ranked according to the HADDOCK score (*see* below), and the top ranking models (here the top 400) are selected for further flexible refinement.

Semiflexible Simulated Annealing in Torsion Angle Space (TAD/SA, it1)

Four stages of SA are performed in *it1* influencing, respectively, the orientation of the components, the side chains at the interface, and finally both side chains and backbone of the interface residues. This semiflexible refinement stage is quite crucial in protein-peptide binding since it allows the peptide to fold and adapt its conformation to the protein binding site. To maximize the chance of finding a correct conformation at this stage, the peptide is treated as fully flexible over all four stages of the simulated annealing refinement. The protein is treated as default, with its interface residues becoming flexible in the last two stages. Further, we increase the number of simulation steps by a factor 4 for the successive stages of the simulated annealing refinement (from the default 500/500/1,000/1,000 to 2,000/2,000/4,000/4,000) to increase sampling. In order to avoid deformation of helical models that may have been selected after it0, dihedral angle restraints are applied to these (*see* **Note 2**).

Restrained Molecular Dynamics in Explicit Solvent (Water)

The structures obtained after simulated annealing are finally refined in an explicit solvent layer to further improve their scoring. This is done by molecular dynamics simulation in water, solvating the complex in an 8 Å shell of TIP3P water molecules [33].

2.3.2 Clustering of Final Solutions

The final models generated by HADDOCK are clustered based on their interface-RMSD using a 5 Å cutoff instead of the 7.5 Å cutoff used for clustering protein-protein poses (*see* **Note 3**).

A smaller value is required in order to ensure conformational homogeneity of the clusters due to the smaller size of the peptides compared to full proteins.

2.3.3 Quality Criteria

To assess the quality of the generated models, we follow the CAPRI standards [34, 35]. We will use mainly the interface-RMSD (i-RMSD), which is calculated on backbone atoms of both protein and peptide residues which are within 10 Å from each other in the reference crystal structures of the complex. The calculation of i-RMSD between a model and a reference is done in two steps that are illustrated in Fig. 1:

1. We fit the protein of the model onto the protein of the reference.
2. We calculate the positional root-mean-square deviation between the model and the reference structures for the backbone atoms of the interface residues (protein + peptide).

To account for the small size of peptides, the standard CAPRI acceptability thresholds need to be decreased:

- Not acceptable: i-RMSD > 2 Å.
- Near-native prediction: 1 Å ≤ i-RMSD ≤ 2 Å.
- High-quality (subangstrom) prediction: i-RMSD < 1 Å.

3 Methods

In order to successfully run this protocol using the HADDOCK web server, two software programs need to be installed locally. First, the input ensemble of three conformations for the peptide can be generated using the PyMOL script provided in the supplementary material associated with this chapter from the *Springer extra* web site (http://extras.springer.com). PyMOL [36] is a molecular visualization system, free for educational use (http://www.pymol.org). Secondly, the models are compared based on RMSD values calculated using ProFit, a free program for protein structure least squares fitting (http://www.bioinf.org.uk/software/profit/index.html). Finally, a web browser, an internet connexion and registration to the HADDOCK web server are the only pre-requisites to access the HADDOCK web server.

In the following sections, we illustrate our protocol on a test case taken from the benchmark dataset [11]. The protocol should be run on a GNU/Linux system or under Mac OSX.

3.1 Modeling of Complexes with HADDOCK

In this section, we model the peptide `DAIDALSSDFT`, corresponding to the disordered region of the calpastatin inhibitory domain C, in complex with the calpain domain VI, a proteolytic enzyme

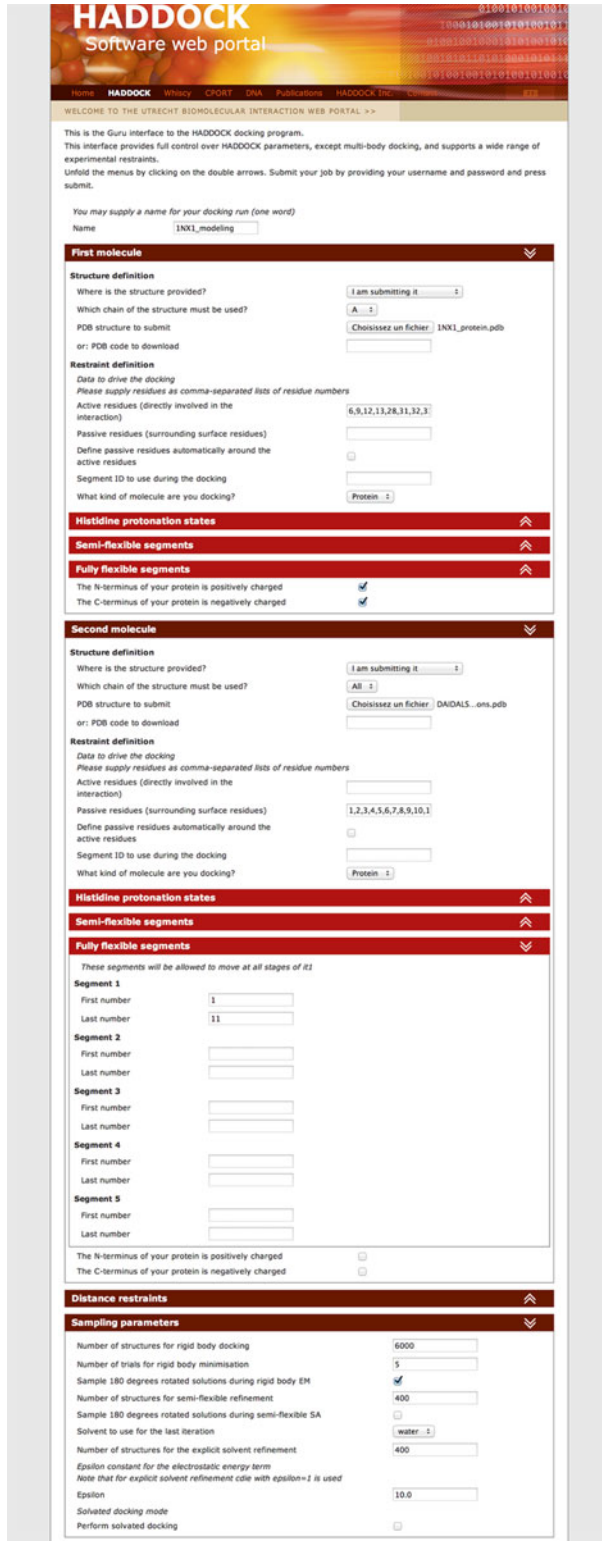


Fig. 1 Overview of the HADDOCK web server Guru interface (accessible from http://haddock.science.uu.nl/services/HADDOCK). A click on the *right arrows* will expand the associated sections to display HADDOCK parameters and/or input fields. In the current view, the *First molecule* and *Second molecule* and *Sampling parameters* sections are expanded. Fields are filled with necessary input for the docking example provided in Subheading 3.1.2

involved in a number of cell functions such as cell mobility and cell cycle progression. The coordinates of both the complex (PDBid: 1NX1) and the unbound structure of the calpain domain VI (PDBid: 1ALV) are available.

3.1.1 Preparation of PDB Files

Each PDB provided to HADDOCK has to respect the PDB format with proper syntax and clear chain identifiers (*see* **Note 4**). The input ensemble for the peptide will be composed of three artificially generated models using PyMOL [36]. Each model corresponds to a specific conformation of the peptide we want to dock onto its associated protein receptor. A PyMol script adapted from an original script of Robert L. Campbell (http://pldserver1.biochem.queensu.ca/~rlc/work/pymol/) is provided to facilitate the creation of the ensemble.

1. Open PyMol and execute the script to access its functions, in the PyMol console, type:

   ```
   > run 3c_build_seq.py
   ```

2. Use the building function provided by the script. For instance, to create the three conformations of the calpastatin peptide required to start the docking run, we type in PyMol:

   ```
   > build_seq extended_pept, DAIDALSSDFT, ss=extended
   > build_seq helical_pept, DAIDALSSDFT, ss=helix
   > build_seq polypro_pept, DAIDALSSDFT, ss=polypro
   ```

3. You can now save the structure coordinates in the PDB format via the Menu File->Save Molecule…

4. Once the three conformations (extended/helix/polyproline II) have been built and saved, the corresponding PDB files have to be merged into a unique PDB file before we can use them as input in HADDOCK. Each conformation must be defined as a unique MODEL, just alike NMR ensemble, meaning that the coordinates of each model must start with a MODEL statement and end with an ENDMDL statements in the PDB coordinate file. This can easily be done with a simple text editor.

The PDB file of the protein must be checked to avoid any double occupancies or residue insertions. This can be done manually or using for example the PDB cleaner website (http://www.igs.cnrs-mrs.fr/Caspr2/magicPDB.cgi) [37]. The input files for both the protein and the ensemble of models for the peptides are provided in supplementary material, respectively, named `1NX1_protein.pdb` and `DAIDALSSDFT_3conformations.pdb`.

3.1.2 Docking the Capstatin Peptide onto Capsain with the HADDOCK Web Server

For this docking, we will make use of the Guru interface of the HADDOCK web server (http://haddock.science.uu.nl/services/HADDOCK/haddockserver-guru.html). Note that the Guru interface is available for registered users with appropriate access rights.

1. Open an Internet browser and go to haddock.science.uu.nl/services/HADDOCK. Choose the Guru interface. You will find the page illustrated in Figs. 1 and 2.
2. We advise to give a name to your docking run. Be aware that no space or special characters other than "-" or "_" are allowed. We propose here to name the run 1NX1_modeling.
3. The PDB file of the largest molecule, in this case the calpain domain IV, has to be entered first (*see* **Note 5**). Expand the section `First molecule`. At the entry `Where is the structure provided?` click on the drop-down menu next to it and select `I am submitting it`. Set `Which chain of the structure must be used?` to *A*ll (*see* **Note 4**). Next to `PDB structure to submit` press the `Browse`... button and move to the location where the tutorial data were unpacked. Go to the `pdbs/` directory and select the `1NX1_protein.pdb` file.
4. Specify the interface by defining active and passive residues. We listed the residues that are considered active in Table 1. Fill in the numbers of the active residues in the textbox next to `Active residues`.
5. Specify the `Segment ID to use during the docking` for the first molecule as A (*see* **Note 4**).
6. We leave the proteins flexibility settings to their defaults values: no residues will be considered as fully flexible and semi-flexible segments will be determined automatically by HADDOCK.
7. Both N-terminus and C-terminus of the protein will be considered as charged, the default value (*see* **Note 6**).
8. Expand the `Second molecule` section. The peptide will require some specific settings, which we will explain in the following steps. If a parameter is not mentioned in the following steps, its default value should be kept.
9. At the entry `Where is the structure provided?` click on the dropdown menu next to it and select `I am submitting it`. Set `Which chain of the structure must be used?` to `All` (*see* **Note 1**). Next to `PDB structure to submit` press the `Browse`... button and move to the location where the tutorial data were unpacked. Go to the `pdbs/` directory and select the `DAIDALSSDFT_3conformations.pdb` file.
10. As explained before, the entire peptide will be considered as passive during the docking process. For this, enter each residue number present in the peptide PDB (for one model) separated by a comma in the *Passive residues* textbox as indicated in the Table 1 (*see* **Note 10**).

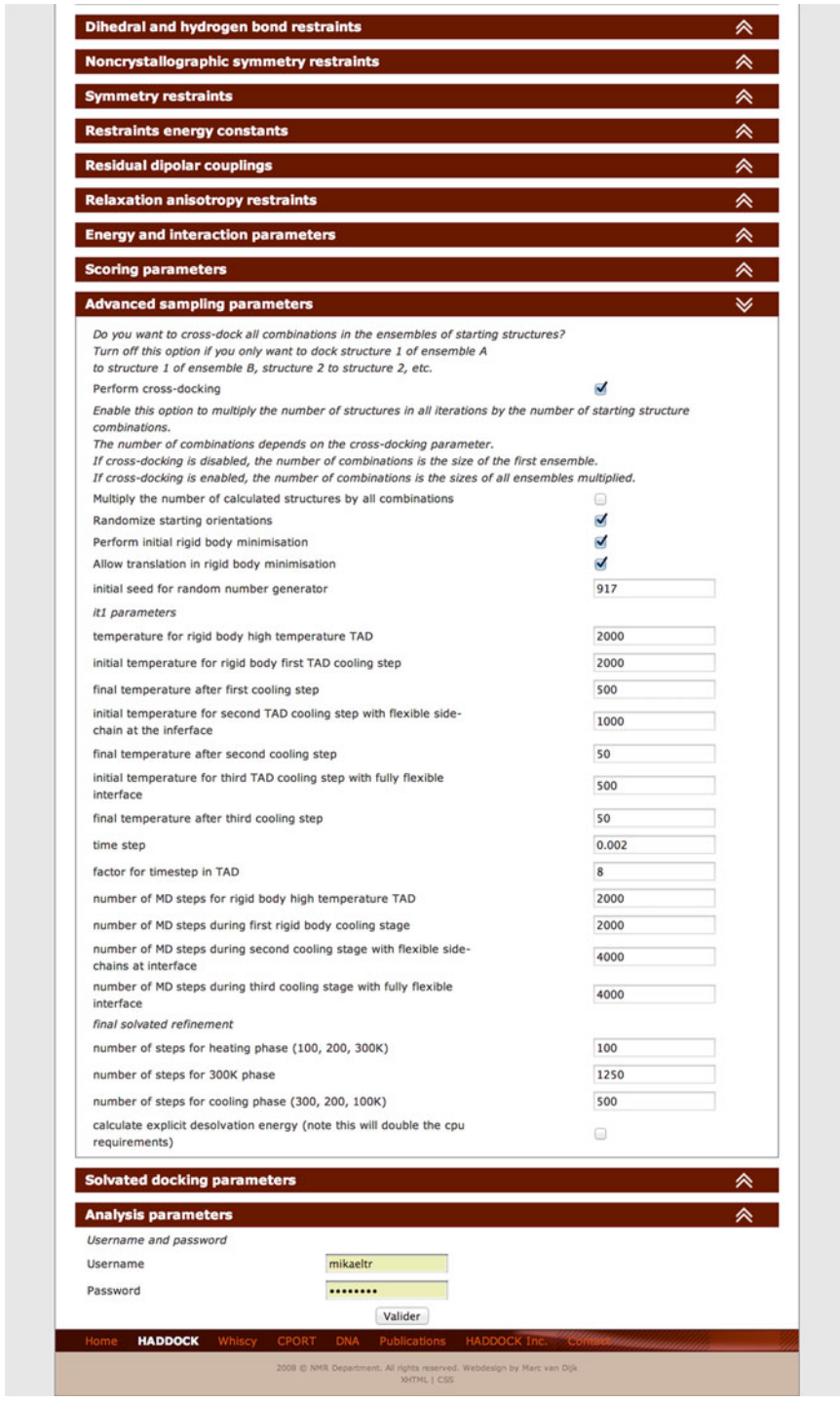


Fig. 2 Overview of the HADDOCK web server Guru interface. The expanded sections include input parameters that need to be changed to perform a protein-peptide docking run. The sections concerned are *Parameters for clustering* and *Advanced sampling parameters*

Table 1
Input data used for the protein-peptide docking run

Protein (Calpain Domain VI)	
Active residues	6, 9, 12, 13, 28, 31, 32, 33, 35, 36, 38, 39, 69, 73, 76, 77, 80, 81, 84, 131
Passive residues	None
Fully flexible segments	None
Peptide (Calpastatin inhibitory domain C)	
Active residues	None
Passive residues	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
Fully flexible segments	11-Jan
C- and N-termini	Uncharged

11. Here, the peptide corresponds to a disordered segment of the capstatin protein and should thus be considered as noncharged at the Cter and Nter. For this, uncheck the two boxes, respectively, *The C-terminus of your protein is negatively charged* and *The N-terminus of your protein is negatively charged* (*see* **Note 6**).
12. In our example, no input data other than the list of active residues on the protein receptor will be used.
13. In the *Sampling parameters* section, we increase the *Number of structures for rigid body docking* (it0) from 1,000 to 6,000. In that way, each conformation is sampled 2,000 times in the rigid body stage. We also increase the *Number of structures for semi-flexible refinement* (it1) and the *Number of structures for the explicit solvent refinement* (water) to 400 structures.
14. Go to the *Parameters for clustering* section and change the *RMSD Cutoff for clustering* from 7.5 to 5.0 to make up for the smaller size of protein-peptide interfaces.
15. In the *Advanced sampling parameters* section, the default numbers of MD steps are multiplied by a factor 4 to increase the sampling. Therefore, the *number of MD steps for rigid body high temperature TAD*, the *number of MD steps during first rigid body cooling stage*, the *number of MD steps during second cooling stage with flexible side-chains at interface* and *the number of MD steps during third cooling stage with fully flexible interface* are respectively set to 2,000/2,000/4,000/4,000.

16. You can now fill in your *Username* and *Password* at the bottom of the submission page and click on the *Submit Query* button. After few seconds you will be redirected to a page reporting the status of your job, first the outcome of the validation steps performed by the HADDOCK web server, then a link to the result page and the possibility to download a unique self-contained file to resubmit your job (provided here with the default name *haddockparam.web*). On the result page, you can monitor the progress of your docking run. When finished, it will later display the final results, which consist in generic analyses of the models. An email to confirm the processing of your job is sent to your registration email address.
17. Within typically a couple of hours, depending on the web server load, you will receive another email reporting the final status of your job. If successful, a result page as depicted in Fig. 3 will be available at the link given in the e-mail. On this page, you will find the name of your docking run as well as a link to download it as a gzipped tar file. A link to the unique file containing input data and parameters is again provided.
18. In this page, you will find the number of clusters created by HADDOCK and how many structures coming from the *water* steps have been clustered. By default, only the 200 models with the lowest HADDOCK scores are analysed, therefore only half of the refined models are clustered. In our example, 15 clusters are created, gathering 66.5 % of the top 200 models. For an easier visualization of the results, only the ten best clusters based on the average HADDOCK score of its top four models are displayed in the summary page. You can find information and analyses of the last cluster in the gzipped tar file. For each cluster, information relative to the HADDOCK score of the top four models, the cluster size and different statistics and energy values are reported (*see* **Note 7**).
19. At last, a graphical representation of different CAPRI assessment criteria with respect to the HADDOCK score is provided for the ten best clusters in the *Results analysis* section as shown in Fig. 4. The first three plots show the HADDOCK score versus the interface-ligand-RMSD (i-l-RMSD), the i-RMSD and the l-RMSD, respectively (*see* **Note 8**). The next plot displays the HADDOCK score versus the fraction of common contacts (FCC) (*see* **Note 9**). The last three plots show the van der Waals, electrostatics, and AIRs energy versus i-RMSD.
20. It is possible to manually compare a reference structure with the best models of each cluster generated by HADDOCK. The 3D structures of these models are located in the root of the docking run you downloaded as a gzipped tar file. Their name follows the following syntax: *cluster2_1.pdb*.

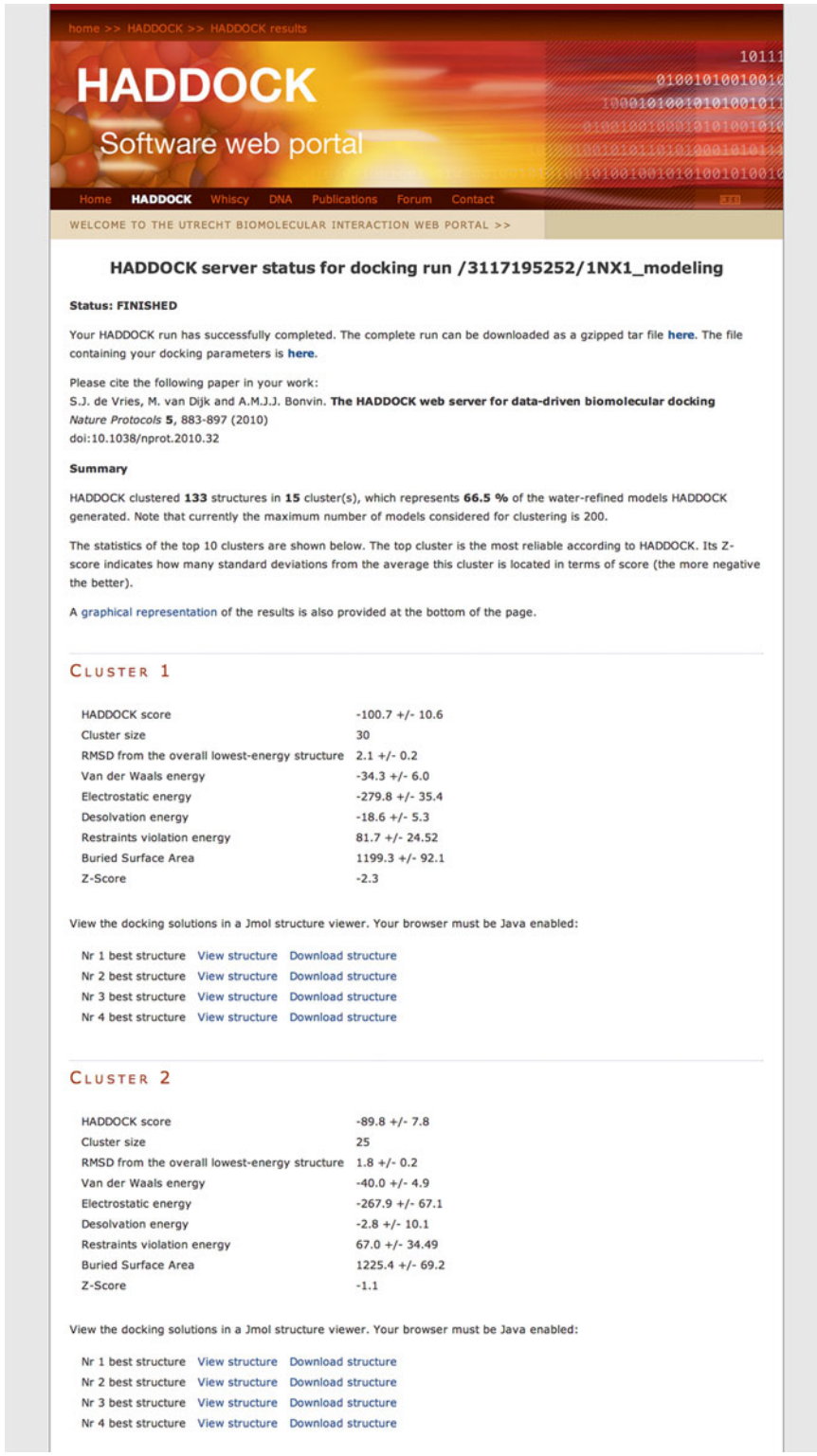


Fig. 3 Example view of a result page of the HADDOCK web server. Links toward the complete run and a HADDOCK-formatted summary of your input parameters can be found. Moreover, a brief summary of the clustering performances is shown with a focus on the first two clusters (according to HADDOCK score average of the top-four structures) analytical information

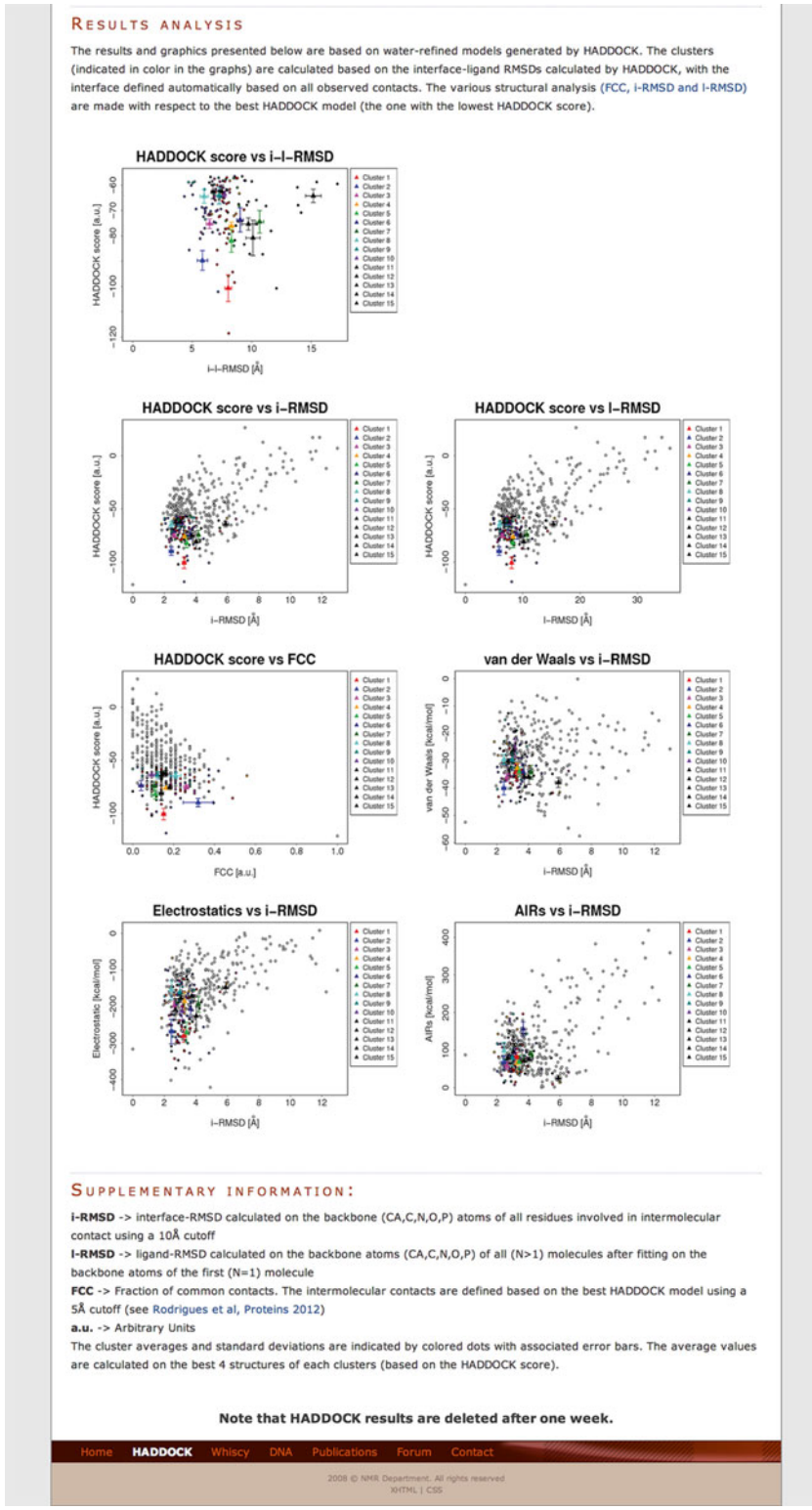


Fig. 4 Results analysis section of a result page of the HADDOCK web server. Several graphics with the main energetic parameters plotted with respect to the HADDOCK score are shown and separated according to the cluster number of each structure

This file is for instance the best model according to its HADDOCK score in the second cluster given by HADDOCK.

You can use ProFit to get precise values of RMSD. PyMol is useful as well since it has its own fitting algorithm and will give you a RMSD value as well as a visual feedback of the differences between the clustered models and the reference structure. Keep in mind that your reference structure has to be formatted in the same way that the PDB models generated by HADDOCK. ProFit considers only structures with an identical number of atoms.

4 Case Studies

The settings we described before have been used to test HADDOCK against a large benchmark of 62 protein-peptide complexes for which an unbound form of the protein was available. The challenge was then double here: model successfully the peptide's conformation at the correct interface and reproduce the bound form of the protein. We analysed the quality of the models generated by HADDOCK but also our capacity to rank efficiently the correct predictions among the top HADDOCK score models. HADDOCK successfully generated acceptable models (*see* Subheading 2.3.3 for definition of acceptable models) for about 70 % of the tested cases (Fig. 5a). Among these (*see* **Note 9**), at least one acceptable model or better is found in the top 20 models in 76 % of the cases. But after clustering, 50 % of the cases contain at least an acceptable structure in the best cluster and this quickly reaches 75 % if the top three clusters are considered (Fig. 5b).

We illustrated the HADDOCK protein-peptide docking protocol in Subheading 3.1.2 with the modeling of the calpain/calpastatin complex, starting from the unbound structure of the calpain and an ensemble of three conformations for the disordered region of the calpastatin. In the last step of HADDOCK protocol (refinement water step), this docking run generated 25 final acceptable models (i-RMSD $\leq$ 2 Å), 18 of which ended in a cluster and 7 were not clustered. Among the 18 clustered models, 15 come from the 2nd best cluster according to HADDOCK and two structures are the 1st and 4th models based on their HADDOCK score. The 2nd best cluster given by HADDOCK is the cluster for which the average HADDOCK score of its top four models is the second lowest among all the clusters. To get a precise idea, the best cluster has an average HADDOCK score for its four best models of -100.7 ± 10.6, as opposed to -89.8 ± 7.8 for the 2nd best cluster. Considering the standard deviations those tow clusters are rather close. The representative best four models of the top-ranking cluster have, on average, an i-RMSD of 3.9 ± 0.3 Å when compared to the crystal structure whereas the best four models of the second best cluster have an average i-RMSD of 1.9 ± 0.3 Å.

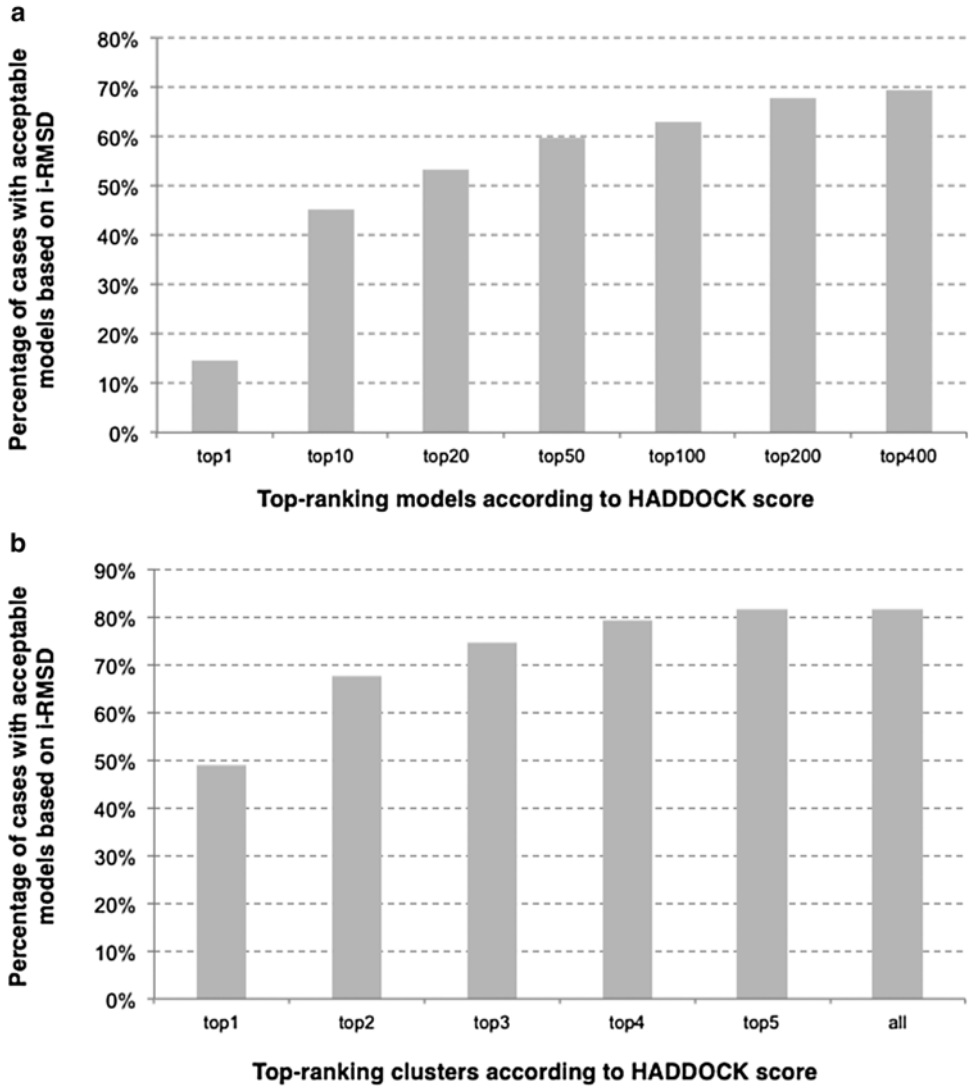


Fig. 5 (**a**) Success rate of unbound/unbound docking as a function of the number of top models considered. (**b**) Clustering performance of HADDOCK in unbound/unbound docking onto acceptable cases (with at least one acceptable model) as a function of the number of clusters considered

The peptide starting conformations and the resulting best model in term of i-RMSD from the reference complex are shown in Fig. 6. Statistics of the two clusters is presented in Table 2. We voluntarily chose this case to illustrate that the correct solution is not always on top and various clusters should be examined, especially when their scores are rather close. Ideally, it would be best to have some independent data at hand to validate the generated models. The models can also serve at starting point for the design of experiments to test the predictions, for example by mutagenesis. It is often the synergistic combination of modeling and experiment that allows to answer challenging biological questions.

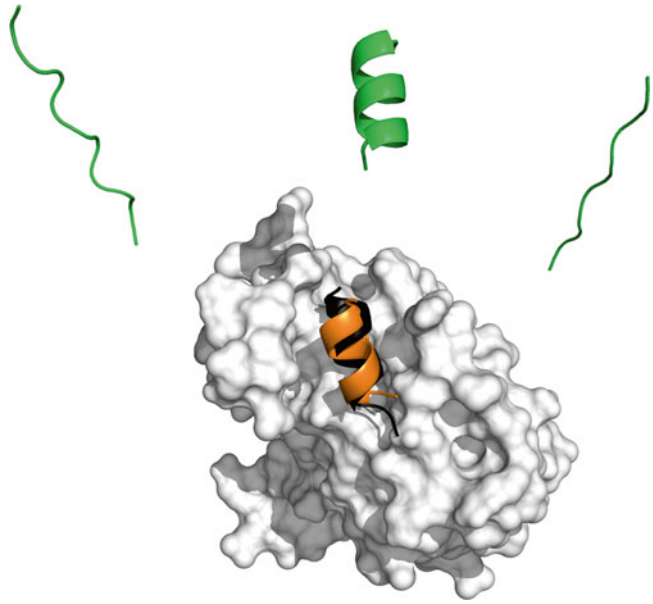


Fig. 6 Summary of HADDOCK protocol illustrated by the docking run between the calpain domain VI and the calpastatin inhibitory domain C. In *green*, the three starting conformations provided to HADDOCK. In *white*, the protein (calpain) rendered as a surface, in *black* the crystal conformation of the peptide as found in the bound complex PDB file and in *purple*, a near-native model corresponding to HADDOCK's third best ranked structure (here the first model of the second cluster)

Table 2
Comparison of two best clusters from HADDOCK for 1NX1 modelling run

	Cluster 1	Cluster 2
HADDOCK score (average)	−100.7 ± 10.6	−89.8 ± 7.8
Cluster size	30	28
RMSD from the overall lowest-energy structure (Å)	2.1 ± 0.2	1.8 ± 0.2
Z-score	−2.3	−1.1
i-RMSD from reference structure for best four structures (Å)	3.9 ± 0.3	1.9 ± 0.3

5 Notes

1. Note that the server does support N-acetylated and C-amino termini. This can however not be specified in the web form, but must be done by editing the coordinates file of the peptide molecule and adding residues at the N- and C-termini, respectively, named ACE/CTN (for example adding a GLY at the termini and rename it to ACE/CTN, respectively; HADDOCK will take care of removing/adding the necessary atoms).

2. This feature will be available in the next release of the web server and is available in the local installation of HADDOCK upon request.
3. For very short peptides, this value might be further decreased.
4. The PDB files provided to HADDOCK have to be correctly formatted to avoid any issues during the simulation process. Any chainID and/or segID should be removed from the input PDBs and there should be no overlap in residue numbering. This can be done for example using the PDB cleaner website (http://www.igs.cnrs-mrs.fr/Caspr2/magicPDB.cgi) [37]. Missing atoms in the PDB files are not problematic since HADDOCK will rebuild them based on the topology files of the force field.
5. Defining the largest molecule as first molecule for docking is important for the final clustering because the structures are first fitted on the interface residues of the first molecule and then the RMSD is calculated on the interface residues of the second molecule. The interface residues are defined from an analysis of contacts in the generated models (at it1 and water, respectively). Defining the largest molecule first should thus result in a better fitting.
6. The charge state of the termini has to be properly set depending on the system under study: naturally occurring peptide with charged termini or peptide fragment extracted from a protein (typically loop or intrinsically disordered region), which should be uncharged … (*see* also **Note 1**).
7. The Z-score indicates how many standard deviations from the average a cluster is located in terms of its HADDOCK score. So the more negative the better.
8. All reported RMSDs are calculated with respect to the lowest scoring model (the best model according to the HADDOCK score). The i-l-RMSD, which is used for clustering, is calculated on the interface backbone atoms of all chains except the first one after fitting on the backbone atom of the interface of the first molecule. The i-RMSD is calculated by fitting on the backbone atoms of all the residues involved in intermolecular contacts within a cutoff of 10 Å. The l-RMSD is obtained by first fitting on the backbone atoms of the first molecule and then calculating the RMSD on the backbone atoms of the remaining chains.
9. The FCC stands for Fraction of Common Contacts and is calculated by comparing the lists of contacts at the interface between the protein and the peptide chain in the reference structure and the model structure. A contact is defined when two residues from different chains of the complex are closer than 5 Å from each other. The FCC is then the percentage of common residue pairs shared between a model and the reference structure.
10. We define a successful case a case for which at least one acceptable model is present in the final 400 models generated.

References

1. Tzakos AG, Fuchs P, van Nuland NA et al (2004) NMR and molecular dynamics studies of an autoimmune myelin basic protein peptide and its antagonist: structural implications for the MHC II (I-Au)-peptide complex from docking calculations. Eur J Biochem 271: 3399–3413
2. Musi V, Birdsall B, Fernandez-Ballester G et al (2006) New approaches to high-throughput structure characterization of SH3 complexes: the example of Myosin-3 and Myosin-5 SH3 domains from S. cerevisiae. Protein Sci 15: 795–807
3. Huang BX, Kim H-Y (2006) Interdomain conformational changes in Akt activation revealed by chemical cross-linking and tandem mass spectrometry. Mol Cell Proteomics 5:1045–1053
4. Casares S, Ab E, Eshuis H et al (2007) The high-resolution NMR structure of the R21A Spc-SH3:P41 complex: understanding the determinants of binding affinity by comparison with Abl-SH3. BMC Struct Biol 7:22
5. Gelis I, Bonvin AM, Keramisanou D et al (2007) Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. Cell 131:756–769
6. Schneider T, Kruse T, Wimmer R et al (2010) Plectasin, a fungal defensin, targets the bacterial cell wall precursor Lipid II. Science 328: 1168–1172
7. Wodak SJ, Janin J (1978) Computer analysis of protein-protein interaction. J Mol Biol 124:323–342
8. Strynadka NCJ, Eisenstein M, Katchalski-Katzir E et al (1996) Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. Nat Struct Mol Biol 3:233–239
9. Petsalaki E, Russell RB (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. Curr Opin Biotechnol 19:344–350
10. Stein A, Aloy P (2008) Contextual specificity in peptide-mediated protein interactions. PLoS One 3:e2524
11. London N, Movshovitz-Attias D, Schueler-Furman O (2010) The structural basis of peptide-protein binding strategies. Structure 18:188–199
12. London N, Raveh B, Schueler-Furman O (2013) Peptide docking and structure-based characterization of peptide binding: from knowledge to know-how. Curr Opin Struct Biol 23:894–902
13. Petsalaki E, Stark A, Garcia-Urdiales E, Russell RB (2009) Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput Biol 5:e1000335
14. Antes I (2010) DynaDock: a new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. Proteins 78:1084–1104
15. Raveh B, London N, Schueler-Furman O (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins 78:2029–2040
16. Ben-Shimon A, Eisenstein M (2010) Computational mapping of anchoring spots on protein surfaces. J Mol Biol 402:259–277
17. Dagliyan O, Proctor EA, D'Auria KM et al (2011) Structural and dynamic determinants of protein-peptide recognition. Structure 19: 1837–1845
18. Raveh B, London N, Zimmerman L, Schueler-Furman O (2011) Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. PLoS One 6:e18934
19. Donsky E, Wolfson HJ (2011) PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors. Bioinformatics 27: 2836–2842
20. Lavi A, Ngan CH, Movshovitz-Attias D et al (2013) Detection of peptide-binding sites on protein surfaces: the first step toward the modeling and targeting of peptide-mediated interactions. Proteins 81:2096–2105
21. Verschueren E, Vanhee P, Rousseau F et al (2013) Protein-peptide complex prediction through fragment interaction patterns. Structure 21:789–797
22. De Vries SJ, van Dijk AD, Krzeminski M et al (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins 69:726–733
23. Trellet M, Melquiond ASJ, Bonvin AMJJ (2013) A unified conformational selection and induced fit approach to protein-peptide docking. PLoS One 8:e58769
24. Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737
25. 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
26. Brunger AT (2007) Version 1.2 of the crystallography and NMR system. Nat Protoc 2: 2728–2733
27. Jorgensen WL, Tirado-Rives J (1988) The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 110:1657–1666

28. Moreira IS, Fernandes PA, Ramos MJ (2010) Protein-protein docking dealing with the unknown. J Comput Chem 31:317–342
29. Lensink MF, Wodak SJ (2013) Docking, scoring, and affinity prediction in CAPRI. Proteins 81:2082–2095
30. Diella F, Haslam N, Chica C et al (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13:6580–6603
31. Van Dijk ADJ, Boelens R, Bonvin AMJJ (2005) Data-driven docking for the study of biomolecular complexes. FEBS J 272:293–312
32. Melquiond ASJ, Bonvin AMJJ (2010) Data-driven docking: using external information to spark the biomolecular rendez-vous. In: Protein-protein complexes: analysis, modelling and drug design. Edited by M. Zacharrias, Imperial College Press, London, p 183–209
33. Jorgensen WL, Chandrasekhar J, Madura JD et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
34. Janin J, Henrick K, Moult J et al (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52:2–9
35. Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78:3073–3084
36. Schrodinger L (2010) The PyMOL molecular graphics system, version 1.3r1
37. Claude J-B, Suhre K, Notredame C et al (2004) CaspR: a web server for automated molecular replacement using homology modelling. Nucleic Acids Res 32:W606–W609

Chapter 11

Computational Approaches to Developing Short Cyclic Peptide Modulators of Protein–Protein Interactions

Fergal J. Duffy, Marc Devocelle, and Denis C. Shields

Abstract

Cyclic peptides are a promising class of bioactive molecules potentially capable of modulating "difficult" targets, such as protein–protein interactions. Cyclic peptides have long been used as therapeutics derived from natural product derivatives, but remain an underexplored class of compounds from the perspective of rational drug design, possibly due to the known weaknesses of peptide drugs in general.

While cyclic peptides are non"druglike" by the accepted empirical rules, their unique structure may lend itself to both membrane permeability and proteolytic resistance—the main barriers to oral delivery. The constrained shape of cyclic peptides also lends itself better to virtual screening approaches, and new tools and successes in this area have been recently noted. An increasing number of strategies are available, both to generate and screen cyclic peptide libraries, and best practises and current successes are described within.

This chapter will describe various computational strategies for virtual screening cyclic peptides, along with known implementations and applications. We will explore the generation and screening of diverse combinatorial virtual libraries, incorporating a range of cyclization strategies and structural modifications. More advanced approaches covered include evolutionary algorithms designed to aid in screening large structural libraries, machine learning approaches, and harnessing bioinformatics resources to bias cyclic peptide virtual libraries towards known bioactive structures.

Key words Cyclic peptide, Protein–protein interaction, Rational drug design, Virtual screening

1 Introduction to Cyclic Peptides

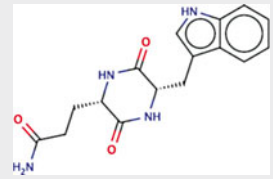
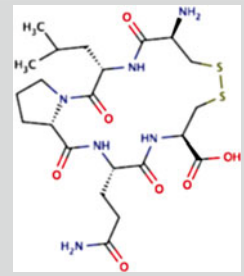
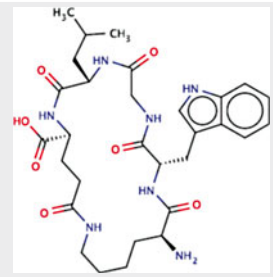
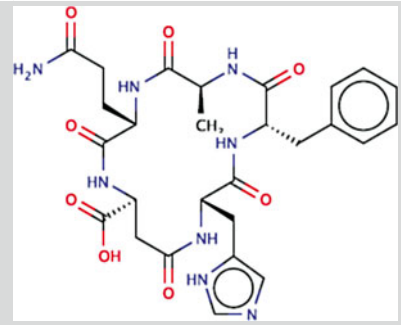
Cyclic peptides are derivatives of linear peptides where the linear peptide has been pinned into a macrocycle by the addition of a chemical bond. This bond can be between amino acid side chains, or the peptide N and C-termini, or a combination of them. Cyclic peptides are of interest as a class of molecules for their ability to mimic specific, high-affinity binding of certain known linear peptides, while potentially avoiding the drawbacks of linear peptides, which include poor oral bioavailability, poor membrane permeability, vulnerability to proteolytic degradation, and lack of a rigid three-dimensional structure.

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_11,

2 Cyclic Peptide Structures

Cyclic peptides are formed through an extra cyclizing bond between peptide termini or side chains. There are many possible ways of doing this, described in Table 1. These strategies include disulphide-bonding, where the thiol side chain of two cysteines are bonded, which are common in natural proteins, both interchain (attaching two protein chains) or intrachain (to constrain a portion of a protein chain in a particular structure); and head–tail bonding,

Table 1
Table showing the principle strategies for cyclizing peptides

Constraint type	Description	Example structure
Head–Tail bond	An amino-acid N-terminus bound to an amino-acid C-terminus	
Disulphide bond	Two cysteine side chains disulphide bound together	
Side-Chain to Side-Chain Bond	An amino acid with an amine group side chain (Lysine) bound to an amino acid with a carboxyl side chain (Aspartic and Glutamic Acid). Also includes depsipeptide bonds between amino acids with a hydroxyl side chains (Serine, Tyrosine, or Threonine) bound to amino acids with carboxyl side chains	
Side-Chain to N-Terminus or C-Terminus	Side-Chain to N-terminus bonds consist of an N-terminal amine bound to a side-chain carboxyl group. Side-Chain to C-terminus bonds consist of a C-terminal carboxyl bound to a side-chain amine	

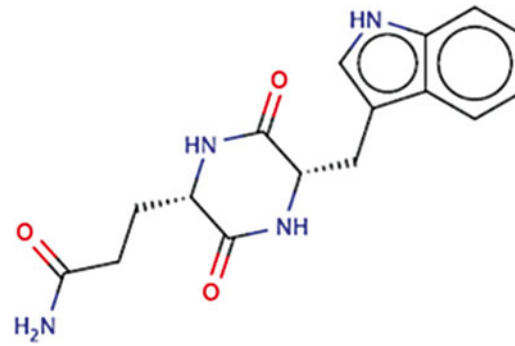


Fig. 1 A Diketopiperazine: Tryptophan–Glutamine head–tail cyclized

where the peptide N-terminus forms a peptide bond with its C-terminus, effectively removing the peptide termini.

Different types of constraints are suited to different uses: Disulphide bonds are generally the cheapest and easiest cyclic peptides to synthesize; however, they may not be suitable for intracellular targets, due to the reducing environment of the cytosol attacking the S-S bond. Head–tail bonding forces the peptide chain into a tight ring compared with other methods, but is more synthetically difficult.

Generally, a cyclic peptide must be at least four amino acids in length to be practically synthesizable, otherwise the steric strain of forcing the structure into a cyclic shape will be too large. For example, a disulphide bonded Cysteine-Alanine-Cysteine peptide would be extremely difficult to synthesize, however Cysteine-Alanine-Alanine-Cysteine would be achievable.

One exception to this is two amino acid head–tail bonded peptides, known as diketopiperazines, where the two peptide bonds form a lactam ring structure. The symmetry between the amino acid backbones allows this tight ring to be formed. Figure 1 shows an example structure. Diketopiperazines are known to have a broad range of biological activities and the unusual lactam ring is a promising drug discovery scaffold [1]. The diketopiperazine lactam ring is similar to the β-lactam ring present in penicillin and cephalosporin antibiotics, which could be considered a class of cyclized peptide, as they are biosynthesized from a starting tri-peptide of L-Cysteine, D-Valine, and L-α-aminoadipic acid, before being modified into their bicyclic β-lactam active form [2].

3 Cyclic Peptides' Role in Drug Discovery

Traditionally, drugs are small molecules which bind into a deep protein cavity, affecting the protein's natural function. However, finding small molecules that bind the large, relatively flat surfaces involved in protein interfaces is usually difficult [3].

Small molecule drugs are typically planar molecules of low molecular weight with simple stereochemistry, and it has been suggested that larger and more complex molecules may be required to effectively target protein–protein interactions [4].

Very large molecules may also act as therapeutics, such as antibodies [5] and therapeutic proteins [6]. Proteins and antibodies can be highly effective molecules displaying extremely strong binding affinities to protein surfaces in-vivo. However, these therapeutics are limited by their size, which makes oral delivery impossible, hinders bioavailability and solubility, and increases cost [7]. This points to the desirability of a middle class of molecule that combines the binding ability of proteins with the bioavailability and stability of small molecules. Oligopeptides (those with roughly 2–20 amino acids) seem like a promising choice as they are chemically identical to natural proteins; however, their short length means that they often lack protein secondary structural elements such as α -helices and β -sheets. This lack of a defined three-dimensional structure can negatively affect peptide stability, and binding affinity [8], in part due to the entropic cost of fixing a flexible molecule into a defined shape on binding to a protein, as well as increasing peptides vulnerability to proteolytic cleavage.

Therefore, a class of structurally stable medium size molecules is desirable. This can be accomplished by the use of molecules based around a large ring structure, known as macrocycles [9, 10]. Macrocycles are generally defined as molecules that contain a ring structure of larger than 8–12 atoms, which covers all practically synthesizable cyclic peptides. Cyclic peptides represent an important class of macrocycle, combining the structural constraint of the macrocyclic ring while sharing the natural building blocks of proteins. Other approaches to macrocyclic drugs include those produced from natural products [11], synthetic peptidomimetic macrocycles [12], and stapled peptides [13].

4 Properties of Cyclic Peptides

4.1 Conformational Constraint

The advantageous properties of cyclic peptides principally come from their relatively fixed three-dimensional shape compared to linear peptides. A regular peptide will have three rotatable bonds for each amino acid along the peptide backbone (N - Cα, Cα - C, C - N). By contrast, the backbone flexibility of a cyclic peptide is drastically reduced (although not totally eliminated).

The key theoretical advantage of conformational constraint is improved specificity. All natural peptides share an identical backbone structure, and chemical variety is provided by side chain groups, meaning that all natural proteins and peptides are chemically very similar. Since protein–protein and protein–ligand binding involves shape complementarity to a protein surface or fitting a

defined hydrophobic pocket, a flexible peptide may shape itself to fit many possible substrates. While this increases the likelihood of affecting the target protein of interest, it also increases the likelihood of off-target effects, which are not generally desirable. 21 % of failures in Phase III clinical trials between 2007 and 2012 were caused by safety issues [14], which carry huge costs, in terms of both money and human effort. A more conformationally constrained molecule can therefore be better suited to drug discovery efforts.

The affinity with which a molecule binds to its target is defined by the enthalpy and entropy of binding. Enthalpy is the term that reflects the strength of the interactions of a ligand with its target—hydrogen bonding, salt bridges, hydrophobic desolvation, and Van der Waals' forces all contribute to this energy, and it is the more intuitive force to understand. Entropy reflects the preference of both ligand and target structures to be in a more "disordered," low-energy state. It also reflects the critical contribution to binding energy that comes from disturbing the network of water molecules around a binding site, which is still poorly understood [15]. It is entropically unfavorable to force a flexible molecule into a rigidly defined shape, such as that required to fit into a protein binding site. Thermodynamically, binding is favored when the combination of enthalpy and entropy changes experienced when a molecule binds to a protein leads to a lower energy state, releasing energy. Binding can be principally influenced by enthalpy, entropy, or a combination of both, and approved drugs exist that are both primarily enthalpic and entropic binders [16]. A molecule that is chemically constrained into a specific shape avoids the entropic penalty that a similar flexible molecule must pay on binding to a target, and therefore, constrained molecules can make higher affinity binding partners. It is possible for a structurally constrained molecule that does not completely fill a binding pocket to bind more strongly to a protein pocket than a more flexible molecule that can rotate to completely fill the pocket, due to avoiding the entropic penalty of desolvating water molecules, and adding more flexible functional groups to the molecule [17].

4.2 Bioavailability

One of the biggest, if not the biggest single, issue facing peptide drugs is bioavailability. The human body has evolved a complex set of biological machinery for producing, controlling, and regulating peptides, and using peptides to regulate biological processes. The centrality of peptides to biological function is both the reason for peptides' huge power to affect biological processes, and hence the therapeutic interest in peptides, and also their Achilles' heel for use as drugs, because unless the peptides are delivered directly to their target, the body can recognize and limit peptide distribution very effectively. Strategies to evade these control mechanisms have been dealt with in detail elsewhere [18], so here we will focus on methods most applicable to cyclic peptides.

The favored delivery method for drugs is by the oral route [19], because it does not require constant medical supervision, and is associated with high rates of patient compliance with their treatment, as compared with alternatives, such as injection. The drawback being that drugs must be absorbed into the bloodstream through the gut to be distributed around the body, and, if the drug is targeted at the brain, cross the blood–brain barrier also.

Bioavailability generally refers to the fraction of the administered drug that reaches the circulatory system, as a fraction of the dose administered dose. Peptide drugs generally exhibit very poor bioavailability, due to two factors: the difficulty of crossing biological membranes, such as the membranes and junctions of the epithelial intestinal cells, and being degraded by the large cohort of specialized peptidase and protease enzymes residing in the gut. Both of these factors can be combated somewhat with the use of cyclic peptides, some of which are orally bioavailable, but there is not as of yet a general strategy for the production of orally bioavailable cyclic peptide drugs.

5 Membrane Permeability

Difficulty in crossing hydrophobic biological membranes is associated with molecular size, number of hydrogen-bond donors and acceptors, and the log octanol-water partition coefficient of the molecule ($\log P_{(o/w)}$), summarized by Lipinski in a set of empirical rules [20]. Cyclic peptides are usually large molecules that fall outside these accepted guidelines for an orally bioavailable molecule, therefore they must explore other strategies for membrane permeability, compared with small molecule drugs.

Focusing on permeability through the intestinal membrane, there are a few routes that a molecule can take: passively penetrating and diffusing transcellularly through the intestinal epithelial cells; passing paracellularly through the tight junctions that link the cells; and carrier mediated transport, where the molecule is encapsulated by a cell vesicle that can travel through the cell using the cell's evolved secretion machinery.

One theorized method of developing membrane-penetrating cyclic peptides is to design a peptide that can internally satisfy its own hydrogen bonds, effectively masking these hydrophilic groups from the hydrophobic membrane. Rezai et al. [21] used this hypothesis to successfully develop membrane permeating cyclic peptides based on a cyclo[Leu-Leu-Leu-Leu-Pro-Tyr] sequence, predicted to potentially adopt an internally hydrogen bonded configuration. They also showed that membrane diffusion rates corresponded with the degree of intramolecular hydrogen bonding. It has been also observed, inspired by the natural cyclic peptide therapeutic ciclosporin [22] that selective *N*-methylation of

backbone amide groups in cyclic peptides can improve membrane permeability by "hiding" hydrophilic amides, and this has been used to develop cyclic hexapeptides with cell permeabilities similar to testosterone [23]. Combining these two insights, peptides have been designed that are partially internally hydrogen bonded, with the remaining free backbone amides *N*-methylated to give an even greater degree of membrane permeability [24]. There are drawbacks to this method: to employ this strategy a candidate cyclic peptide must adhere to a restricted conformational space, not necessarily suited to binding to a biological target. Another drawback is that the effect of *N*-methylation on the peptide backbone can affect the affinity and selectivity of the peptide, although it has also been seen to increase specificity through conformational constraint [25, 26].

An alternative approach to improving membrane permeability is covalent attachment of a polyethylene glycol (PEG) group to the peptide. PEG is considered a biocompatible attachment, suitable for improving pharmacological properties of peptides [27]. It has been successfully used to add oral bioavailability to insulin [28], and Chen et al. [29] have created PEGylated cyclic peptides based on the RGD (Arginine-Glycine-Aspartic acid) integrin binding motif that showed improved pharmacokinetics over nonmodified peptides.

There has also been work done on peptide delivery by harnessing the body's own peptide transport systems: the PEPT1 and PEPT2 transporters are able to naturally take up di- and tripeptides, and a variety of different hydrophilic peptidomimetic drugs [30]. Certain macromolecules may also move through the cell via receptor mediated endocytosis, and peptide conjugation to vitamin B_{12} has been shown to be capable of moving the peptide through the gut lining [31].

6 Proteolytic Resistance

Proteases, or peptidases, break down peptides by binding a specific recognition site. Proteases act by binding to a specific sequence, or set of sequences along the peptide backbone, and breaking the peptide backbone at a specified location. There are a very large number of known proteases, the MEROPS [32] protease database lists 703 known and putative peptidases in human alone. One class of proteases are the exopeptidases, which cleave either the N-terminal or C-terminal amino acid from a peptide chain. Cyclic peptides, since they lack a natural C or N-terminal, are naturally protected from these peptidases.

The catalytic action of a protease depends on the precise layout of the residues involved in catalysis at the enzyme active site. For example, in the case of serine proteases such as chymotrypsin, there are three amino acids, histidine, serine, and aspartic acid involved

catalyzing the breakup of the peptide bond, which are precisely spaced in the active site [33]. The peptide must fit into the protease active site in the correct orientation for proteolytic degradation to occur, and the specificity of a particular enzyme is determined by how well it recruits the peptide to the active site. Recruiting the peptide is done with a specific peptide binding site that directs the peptide to the active site. Conformational restraint, or nonnatural amino acids can prevent proper recognition of peptides by the proteases [34, 35]. The incorporation of nonnatural, D-enantiomer amino acids is commonly used in cyclic peptides, and has been shown to prevent protease recognition in both linear [36] and cyclic peptides [37]. Bacteria have even evolved cyclic peptides as natural protease inhibitors [38].

Despite this, cyclic peptides are not guaranteed to be resistant to the whole spectrum of proteases. Another option to prevent proteolytic degradation is protecting the peptide by chemically modifying it to be cleaved into the active form in the gut: i.e. a prodrug (reviewed by Wang et al. [39]). Alternatively, drug delivery can be designed to go in between, rather than through the gut epithelial cells where there are few proteases [40], which may be assisted by using a tight-junction modulator compound [41].

The body's proteolytic system may also be worked around using a specially formulated slow-release system. Here, the peptide is packaged in a synthetic substance which is slowly degraded by proteases. This degradation releases the contained peptides, which are active for a short time before they are, in turn, degraded. However, the synthetic slow release formulation lengthens the delivery period to maintain a constant low dose of active peptide in the bloodstream, despite the rapid clearance of the peptide.

Recently, Amiram et al. [42] have described a slow release peptide formulation that can be completely biologically synthesized in an *E. coli* expression system, which is effective for 120 times longer than an injection of the native peptide drug, and avoids the difficult and expensive step of manufacturing synthetic microparticles to act as a slow release system.

6.1 Cyclic Peptides and Protein–Protein and Protein–Peptide Interactions

Protein–protein interactions (PPIs) pose difficulties for traditional small molecules, due to the large, shallow interfaces that typify these interactions. Historically, small molecule drugs have been developed as analogues of cellular metabolites or hormones that sit in the well-defined hydrophobic pockets that are the protein's active sites, modulating signaling or enzymatic activity. This has raised questions over the suitability of traditional small molecules to modulate protein–protein interactions, due to the fundamentally different nature of protein–protein interaction surfaces, and protein binding pocket chemistry [43]. Despite this, small molecule protein–protein interaction inhibitors have been identified, reviewed here [3], those with structural information available are curated in the 2P2I database [44].

A typical protein–protein interaction surface is roughly 1,600 Å^2, and there are known complexes with surfaces areas five times larger [45], which presents a very large surface for a cyclic peptide to cover. However, the binding energy for a protein–protein interaction is not evenly spread over the contact surface, and generally is focused on a few "hot-spots" that provide most of the binding energy [46]. Additionally, protein–protein surfaces are not fixed surfaces, but have dynamic shapes, characterized by transient pockets and a degree of flexibility [47]. Protein–protein x-ray structures can be misleading here, as they can only show a static interaction surface, but it is an important consideration in drug design. Cyclic peptides are at an advantage compared to small molecules when designing modulators of protein–protein interactions: they are constructed from the same natural building blocks as the protein's biological binding partner; they are typically larger than average small molecules, which allows them to cover more of the hot-spot features; and they can conformationally mimic key features or epitopes of one half of a protein binding surface to out-compete the natural protein binding partner. This approach has been used to develop cyclic peptide mimics of reverse turns (or β-turns as they are sometimes known) [48, 49]. These turn structures have been recognized as ligands for over one hundred G-protein coupled receptors [50]. Fasan et al. [51] have used an α-helical cyclic peptide to inhibit the interaction of the p53 tumor suppressor protein with its regulatory partner HDM2, when small molecule approaches did not work.

Domain–motif interactions, where a globular protein binds to a short sequence (3–15 amino acids) of a disordered protein are an interesting subclass of protein–protein interactions. These short sequences are known as Short Linear Motifs (SLiMs), also known as Eukaryotik Binding Motifs (ELMs), that can act as domain binding locations, protein targeting signals, post-translational modification sites, or cleavage sites. The ELM [52] database records currently known examples. These might also be seen as protein–peptide interactions, rather than protein–protein interactions, due to the small size of the SLiM sequence binding to the protein. Linear motif interactions are smaller than most protein–protein interactions making them more tractable for designing drugs than typical protein–protein interactions. One of the best-known classes of bioactives derived from a known SLiM motif are cyclic peptides mimicking the RGD motif, which is selective for the αV family of integrins. This has led to the development of Cilingetide [53], a drug under investigation for treatment of angiogenesis. Cyclic peptides have also been developed incorporating the RGD-related related NGR motif, for targeting anti-tumor compounds to tumors [54]. Peptide compounds have also been developed that target the SH3 domain [55].

Protein–peptide interactions, of which SLiMs are a subset, comprise an estimated 15–40 % of the interactions in the cell [56].

This makes them an extremely attractive template for cyclic peptide drug development. The PepX [57] database curates a set of protein-binding peptides. Cyclic peptides cannot always be used to mimic a linear peptide, as linear peptides often bind in an extended linear conformation [58] incompatible with the geometry of a cyclic peptide. Nevertheless, the set of protein binding peptides contains promising targets for cyclic peptides, such as protease inhibitors [59].

7 Currently used Cyclic Peptides/Cyclic Peptides as Drugs

By searching the ChEMBL [60] database of bioactive druglike molecules for approved drug structures, we can identify cyclic peptide like drugs. Here we define "cyclic peptide-like" as those compounds with at least a dipeptide in a macrocycle as a substructural component (i.e. those matching the SMARTS [61] pattern N;r;!r3;!r4;!r5;!r6;!r7]CC(=O)[N;r;!r3;!r4;!r5;!r6;!r7]CC=O. Drugs are given a United Stated Adopted Name (USAN) which is based on a rough nomenclature that consists of a series of "stems," which are usually suffixes, but can be prefixes or infixes that relate drugs to broad chemical families based on structure or activity. The current list of USAN stems can be found at http://www.ama-assn.org/ama1/pub/upload/mm/365/stem-list-cumulative.pdf.

Table 2 summarizes the activity of these drugs as inferred from their name. From this table, it is clear that cyclic peptide antibacterials and antifungals dominate the list. The term "peptide drugs" covers compounds with a wide variety of activity, and includes compounds such as Octreotide, a somostatin mimic [62], Linaclotide, a peptide agonist of guanylate cyclase 2C used for treating abdominal pain for IBS sufferers [63], Davalintide, an amylin-mimetic peptide to reduce food intake [64], and Cilengitide, an angiogenesis inhibitor [53]. Despite the bias of cyclic peptide structures towards antibiotic action, it is clear that cyclic peptide drugs exist with a wide range of biological activities.

The size distribution of peptide drugs is shown in Fig. 2. Figure 2a shows the typical molecular weight distribution of all drug molecules—it can clearly be seen that drug molecular weights between 200 and 600 predominate—consistent with known guidelines for oral drug availability. Figure 2b shows the weight distribution for approved "cyclic peptide-like" drugs, as defined above. It is clear that cyclic peptide drugs are a great deal larger than typical drugs, with cyclic peptide drug molecular weights lying primarily between 1,000 and 2,000. Figure 2b also includes some disulphide-bonded proteins that have been approved as drugs, such as Insulin (m.w: 5916) and Mirostepin (m.w: 8848).

Table 2
Activities of cyclic peptide drugs inferred from the stem of the United States Adopted Name

Activity type	USAN stems	Number of cyclic peptide drugs
Antibacterials	-planin, -mycin, -myxin, -vancin, -tracin, -cetin, -cidin, -cycline, -ganan, -tricin	31
Peptide drug	-tide	26
Oxytocin antagonists and derivatives	-siban, -tocin	9
Vasoconstrictors	-pressin	6
Antifungals	-fungin	5
Immunosuppresants	-sporin, -dar	5
Depsipeptide derivitives	-depsin	1
Prehormones or hormone-release stimulating peptides	-relin	1
Enzyme Inhibitor and growth hormone derivative	som- -stat	1
Tachykinin receptor antagonists	-tant	1
Antivirals	-vir-	1

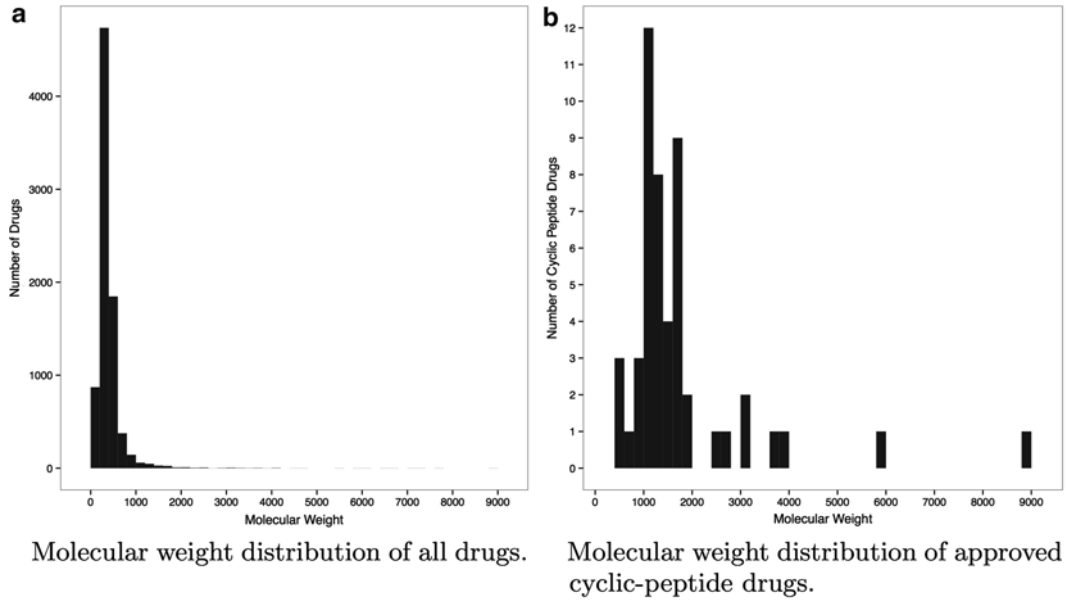


Fig. 2 Histrograms showing the molecular weight distributions of (**a**) all drugs and (**b**) approved cyclic peptide drugs only

To further illustrate the properties of cyclic peptide drugs, Figure 3 shows example cyclic-peptide drug structures, one from each activity type with more than one representative in Table 2. Figure 3 illustrates the variation in bioactive cyclic peptides, all of them have the typical central peptide-backbone macrocycle, but it varies in size, from 5 backbone amino acids in the case of Octreotide Fig. 3c and Lypressin Fig. 3d to 12 backbone amino acids in the case of Ciclosporin (Fig. 3f).

While the cyclic peptide ring is more or less the same between peptides, it can play an important role in binding: for example, it has been shown [65] that there exist several hydrogen bond contacts between the ciclosporin backbone and human cyclophilin A (its therapeutic target). This suggests that the cyclic peptide backbone is more than an inactive scaffold, given biological relevance by its side chain groups.

7.1 Biological Methods for Cyclic Peptide Screening

To contrast with computational cyclic peptide screening methods, it is useful to briefly introduce the main methods for identifying diverse and complex bioactive peptide and cyclic-peptide structures.

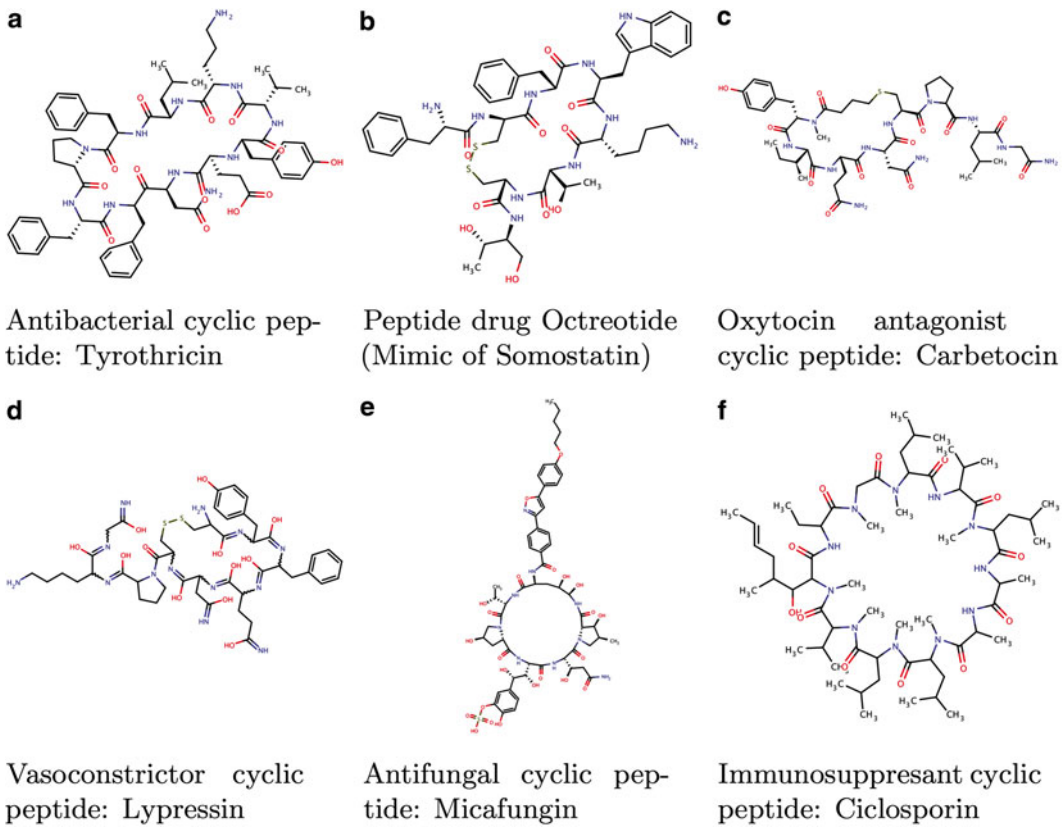


Fig. 3 Example cyclic structures for each activity class of cyclic peptides with more than one representative in Table 2

7.2 Genetically Encoded De-Novo Peptide Libraries

Phage display is a technology that takes advantage of certain members of the filamentous phage virus family, most commonly the M13 phage that have the ability to display nonnative peptides on its surface coat proteins to generate and assay large peptide libraries for binding to a biological macromolecule [66]. Phage display represents a biological method of rapidly generating large random peptide and cyclic peptide libraries. To create a phage-display peptide library, random oligonucleotides are inserted into the coding sequence of one of the coat proteins. The phage is inserted into bacteria, usually *E. coli*, where the phage begins replicating and is released, displaying the random peptide on its surface. Phages are selected by binding to an immobilized target protein, with the nonbinders being washed away. This process is repeated several times to identify strong binders and the binding phage can then be sequenced to retrieve the binding peptide sequence. Phage display enables the display of libraries of 10^{10} peptides simultaneously [67], with typical peptide sizes between 5 and 20 residues [68], and has been used to successfully identify high-affinity disulphide-bonded cyclic peptides [69].

Another method of genetically encoding peptides is via the SICLOPPS [70] technique (split intein-mediated circular ligation of peptides and proteins), which allows production of any head–tail bonded cyclic peptides inside a eukaryotic cell. This technique has been used to express and screen a library of head–tail cyclic peptides in yeast cells that yielded cyclic peptides that specifically reduce the toxicity of human α-synuclein [71].

8 Cyclic Peptides Derived from Natural Sources

The primary advantage of cyclic peptides derived from natural sources is the idea that evolution has already done the work of selecting a set of bioactive peptide scaffolds that can be taken advantage of for drug discovery purposes. Natural sources contain a rich diversity of cyclic peptide, and cyclic peptide like structures. Natural cyclic peptides come from two principal sources—they can be synthesized, like proteins, from DNA, or they can be nonribosomal peptide natural products, synthesized by specialized nonribosomal peptide synthetases in microorganisms like bacteria and fungi which can incorporate a great variety of nonnatural amino acids and post-translational modifications and possess a vast chemical diversity [11]. Nonribosomal cyclic peptides are the principal source of cyclic peptide antibiotic structures such as Tyrocidine [72] and Daptomycin [73]. The cyanotoxins, hugely potent natural toxins produced by cyanobacteria, have many cyclic peptide examples, such as microcystins [74], which inhibit protein phosphatases type 1 and 2A, and nodularins [75]. Nonribosomal cyclic peptides also include anti-cancer drugs, such as the epothilones [76].

Compared to the nonribosomal natural product cyclic peptides, bioactive genetically encoded cyclic peptides are somewhat underexplored. There are many studies involving cyclic peptide analogues of a particular protein loop or peptide motif, such as the previously mentioned RGD cyclic peptides, which mimic an integrin recognition motif [26], but relatively few taking an existing cyclic portion of a protein "as-is". One example is the serine protease inhibitor cyclic peptides based on a disulphide-bonded reactive site loop of the Bowman-Birk protease inhibitors [77].

9 Virtual Screening of Cyclic Peptides

9.1 Virtual Screening Methods

Virtual screening refers to a set of computational methods that aim to identify active molecules for a biological target based on similarity to known active ligands, or by complementarity to a binding surface. Virtual screening methods can be broken down into two main categories: Ligand-based and structure-based screening.

Ligand-based screening is based around the observation that molecules similar to already known ligands will have similar biological activity. Therefore, potentially active molecules for a particular target can be identified by their similarity to the known ligand(s). Ligand-based screening methods include fingerprinting methods, pharmacophore matching, and shape-based matching. Common ligand-based screening methods are described in Table 3. Ligand-based screening methods can be extremely quick, especially those that reduce a molecule to a bit-string of 0s and 1s representing its chemical properties, as comparing bit-strings is very fast computationally. Bit-strings can be precalculated for large libraries of candidate compounds, allowing them to be easily rescreened against many true ligand structures. Despite the lack of structural information on the target protein, ligand-based methods have been shown to be just as accurate as structure-based methods [78], although this may be down to an imperfect understanding of how to design a docking scoring function that works well across diverse target types [79].

Structure-based screening is based on exploiting the known three-dimensional structure of the target and the topology of its ligand binding surface to design or choose possible active molecules, often by attempting to "dock" the prospective ligand molecule into a predefined binding site on the target. Docking generally consists of two steps, iteratively repeated: pose generation, also known as the search stage, and scoring. Pose generation is the act of computationally positioning the ligand within the defined binding site, when it is then scored, using a specialized scoring algorithm, to predict how good the pose is. This is generally repeated until a set limit of time or number of rounds of posing and scoring has been met, and the best score, or set of scores is returned, along

Table 3
Common ligand-based screening methods

Method	Description	Examples
Molecular Fingerprints	The molecule is represented by a bit-string where each bit represents either the presence or absence of a chemical fingerprint (Structure-based) or the bit-string is based on assigning a numerical value to the atomic and bonding properties of linear substructures of the molecule and passing the results through a hash function to create a bit-string (Hash-based). Comparing two molecules is done by calculating the number of shared "on" bits in the fingerprints and dividing that by the total "on" bits in both keys (the Tanimoto score)	MACCs keys [80] (Structure based), Daylight fingerprints [81](Hash-based)
Shape Matching	The similarity of two molecules is compared based on their three-dimensional shape, either by aligning 3D structures of the molecules and calculating the root mean squared deviation (RMSD), or by comparing statistical measures of molecular shape	USR [82], USRCAT [83], ROCS [84], PhaseShape [85]
Pharmacophore Matching	The molecule is broken into a set of points or volumes representing a particular chemical feature, such as a hydrogen bond donor/acceptor, +/− charge, lipophilic regions and aromatic groups. A query pharmacophore model can be built from features of a known ligand known to be important for binding, and the matching score is based on a feature volume overlap score between the query pharmacophore and the candidate ligand features	Pharao [86], MOE [87], Pharmer [88], Catalyst [89]

with predicted binding conformations. Table 4 gives a short synopsis of common scoring algorithms, and Table 5 explains the pose generation algorithms. Docking approaches can also be divided into rigid and flexible approaches—rigid docking is very fast, but less accurate. Rigid body docking has been used by Mosca et al [90] to accurately identify interacting proteins in the *Saccharomyces cerevisiae* interactome. Generally, when using virtual screening to find protein ligands, a flexible ligand—static protein, or static protein backbone model is used. Yuriev at al [91] have reviewed different docking approaches and challenges in detail.

One of the advantages of the virtual screening approach is the availability of many high-quality software packages available, both commercial, such as MOE [87], OEChem [100], the Schrödinger [101] suite of programs, Accelrys Discovery Studio [89] and open-source toolkits such as RDKit [102], OpenBabel [103], and the Chemistry Development Kit [104] (CDK). A group promoting open-source virtual screening tools, The Blue Obelisk [105] maintains a list of open source screening tools. Commercial virtual screening approaches generally provide a complete graphical environment, with a graphical workbench ready to immediately screen, while the open-source equivalents are often programming toolkits

Table 4
Description of various docking scoring algorithms

Scoring algorithm	Description	Example implementations
Empirical	Count up the number of favorable interactions between ligand and target, or calculate the change in solvent accessible surface area to rank docking poses	Sybyl FlexXScore [92], Autodock Vina [93]
Molecular Mechanical	Use a molecular mechanical force field, such as AMBER to estimate binding affinities based on Van der Waals, hydrogen bond, and charged contacts	DOCK [94]
Knowledge Based	Assesses docking score based on the statistical similarity of docked conformations to known protein ligand structures, such as those in the Protein Data Bank [95]	eHITs [96], Sybyl PMF [92]

Note that many docking programs

Table 5
Description of various docking pose generations algorithms

Pose generation algorithm	Description	Example implementations
Systematic Sampling	The computational space is systematically explored by rotating flexible bonds. This may be followed by sampling a diverse subset of the generated structures	MOLSDOCK [97]
Incremental Construction	The molecule is assembled within the constraints of the defined docking site, from fragments of the input molecule	DOCK [94], E-novo [98]
Genetic Algorithms	The ligand conformations in the docking site are represented as a "gene" and using the docking score as the fitness function, they are mutated, recombined, and rescored iteratively	Autodock4 [99]
Monte Carlo	Ligand poses are randomly modified and locally optimized	Autodock Vina [93]

that require the user to construct their workflow using a computer scripting language such as Python. There have been efforts to make more user-friendly open-source screening tools, such as the Knime [106] workbench.

10 Screening Virtual Cyclic Peptide Libraries

10.1 Combinatorial Library Generation

In order to begin a virtual screening campaign, there are two basic requirements: a target, which is generally the surface of a biological macromolecule, usually a protein, but there are other options including DNA and RNA structures; and a library of candidate compounds.

There are several sources of diverse preprepared small molecule virtual libraries, such as those curated by the ChEMBL and ZINC [107] databases, and pharmaceutical companies generally curate their own compound collections. For cyclic peptide compounds, the most straightforward method of assembling a library is by combinatorial generation, where a number of basic building blocks are assembled into a set of compounds, such as amino acid building blocks for cyclic peptides. This is a little explored approach, although Burns et al [108] have successfully used docking of cyclic peptide virtual libraries to find RNA binding partners.

A key advantage of virtual compound generation is the ease of including exotic amino acids and modifications that may be difficult or expensive to synthesize. It is then only necessary to synthesize compounds that are among the top hits. A huge chemical diversity of amino acid structures are commercially available, for example on the ZINC database, but may be cost prohibitive to use in high-throughput screening, due to synthesis costs. Virtual libraries allow basic validation of possible compounds before any complex chemistry takes place. We have developed CycloPs [109], software designed for the generation of virtual libraries of cyclic peptides, which can incorporate a variety of cyclic peptide constraint strategies, as well as user-defined amino acid structures, (allowing, for example, the inclusion of amino acids including post-translational modifications in the library), and the ability to filter out cyclic peptides likely to be difficult to synthesize.

10.1.1 Structural Optimization

In general, virtual screening methods can be two- or three-dimensional. 2D approaches represent the molecule as a mathematical graph structure of atoms joined by bonds, and calculate molecular similarity based on substructures in these graphs, or by the various possible paths through the graph. In contrast, three-dimensional virtual screening approaches, such as pharmacophore matching, use the actual predicted three-dimensional shape of the molecule to score hits. For this, the three-dimensional shape of the compound in solution must be predicted. Due to the conformational restraint of cyclic peptides, accurate solvation structures may be predicted: an example being the work of Goldtzvik et al [110], who used the DEEPSAM structure prediction algorithm from the Tinker [111] molecular modeling package to accurately predict the solution structures of a set of five small cyclic peptides. However, the potential conformational change upon binding of a cyclic peptide to its target means that it is useful to predict a range of likely conformations for a cyclic peptide to be used in any rigid-body virtual screening step, rather than attempting to predict one single shape. Molecules with large numbers of rotatable bonds are difficult to model computationally, both in terms of calculating all possible three-dimensional conformations, and in identifying the biologically relevant ones. However the constrained central ring

structure of cyclic peptides makes these calculations considerably faster and less error prone than the corresponding linear peptide, and is a key reason for the feasibility of virtual screening of cyclic peptides in contrast to linear peptides.

Conformer generation is a standard step in virtual screening, and most of the software packages mentioned above will have built-in routines to accomplish this task. Care must be taken when selecting a conformer generation algorithm as some conformer generation software designed for small molecules including Confab [112] generate conformers by systematically rotating flexible bonds, and are not capable of varying macrocyclic rings. Suitable cyclic peptide software for generating cyclic peptide conformers include loop prediction software in the Protein Local Optimisation Program [113], which been successfully used to accurately predict the solution structures of cyclic hexapeptides [21].

Also, in their recent assessment of the quality of various currently available conformer generation software packages, Ebejer et al [114] recommend an approach using the RDKit [102] cheminformatics library, that combines stochastic conformer generation and subsequent optimization that can generate diverse, low-energy conformations, including varying ring structures.

10.1.2 Screening

Cyclic peptides are larger, and more three-dimensional than drug-like small molecules, which are typically planar molecules without complex stereochemistry or structure [4]. For this reason, it seems more appropriate to use three-dimensional screening methods, despite the fact that, in general, two-dimensional virtual screening methods have proved as effective as three-dimensional methods for small molecules [115]. Cyclic peptides are often very structurally similar, with a large shared peptide backbone, and a defined set of side chain groups that reduces the power of substructural searches to discriminate between structures. For example, a fingerprint-based approach will not be able to discriminate between two cyclic peptides with a different sequence, but the same amino acid composition (such as CGVPRRC and CRVGPRC), despite potentially very different activities. Methods to use include docking, or three-dimensional pharmacophore matching. Pharmacophore matching is a ligand-based screen, with key pharmacophore points taken from a 3D structure of a known ligand interacting with the protein target, but is also possible to include structural information by the use of exclusion volumes, where the candidate molecule must match the key pharmacophore features of the known ligand, while staying out of the exclusion volumes, which are used to avoid hits that would have steric clashes with the protein target.

10.2 Validating Peptide Hits

Virtual screening is not capable of proving biological activity, so it is usually desirable to assay the top hits of a virtual screening campaign. With the advances in peptide synthesis techniques since the

invention of solid-phase peptide synthesis in 1963 [116, 117], and consequent fall in price of peptide synthesis, it is now possible to order custom peptides at a reasonable cost from many suppliers.

10.2.1 Peptide Arrays

The SPOT synthesis technique [118] has allowed the development of peptide arrays, which allows the synthesis of thousands of peptides on membrane sheets, enabling high-throughput follow-on screens. Peptide arrays were originally developed for use in antibody epitope mapping, but they are flexible enough to be used for many applications. The use of peptide arrays for studying protein–protein interactions has been reviewed by Katz [119] et al, with the same techniques translating over to protein–peptide interactions. The basic technique involves incubating the array with a chemically or fluorescently tagged binding partner, or incubation with a protein of interest followed by a fluorescently tagged secondary antibody, before removing the unbound substrate and visualizing the results. Results from peptide arrays are semi-quantitative—i.e. they can distinguish between strong binding, weak binding, and no binding, but do not provide a precise measure of binding affinity. Inserting control peptides of known binding affinity onto the array will allow estimation of binding affinity by comparing the signal strength of known and unknown binders.

10.3 Virtual Screening vs. High-Throughput Screening

The main competitor to virtual screening is high-throughput screening. High-throughput screening involved automated compound handling using variations on the traditional 96 well plate—often with thousands of wells, each of which will contain one or more individual assays. Current approaches can screen up to 100,000 compounds a day [120], and can use extremely tiny volumes of reactants, at a low cost per molecule screened. However, the equipment itself is expensive, usually found only in industry, with a few academic exceptions [121].

The purpose of both methods is to filter down a large library of compounds into a shortlist of active or "lead-like" compounds which can be used as the basis for more rational design or optimization methods. The key disadvantage of virtual screening compared with high-throughput screening is that virtual screening can only predict which compounds are likely to be active, but high-throughput screening provides direct, physical evidence of activity. The main advantages of virtual screening are the cost of computational power, which is rapidly plummeting in an age of increasing computer performance, and harnessing graphical processing units (GPUs) for very fast parallel processing, which can result in a 260-fold increase in screening speed compared to a traditional CPU [122]. Virtual screening is benefiting from the vast resources that have been poured into improving general purpose computing power for all users, in contrast to the specialized equipment required for high-throughput screening.

10.4 Using Evolutionary Algorithms to Screen Large Cyclic Peptide Libraries

A major stumbling block in computationally screening combinatorial cyclic peptide libraries is the issue of the exponential explosion in the number of peptide structures. Figure 2 shows that a typical therapeutic cyclic peptide has a molecular weight of between 600 and 2,000 Da. With the mean amino acid weight being 120 Da, this corresponds to a typical therapeutic cyclic peptide size between 5 and 16 amino acids, ignoring potential post-translational modifications. To fully explore the set of head–tail 10-mer cyclic peptides combinatorially would require generating over 10^{13} cyclic peptide structures—a prohibitively large amount.

There are several options to attack the problem of combinatorial explosion, such preselecting a restricted library of amino acids, limiting the number of variable positions (for example, varying 4 positions on an 8-mer cyclic peptide). The choices of amino acids and variable position would ideally be guided by the biological or chemical properties of the protein interface of interest.

A more advanced approach to screening large combinatorial spaces is the use of evolutionary algorithms. Evolutionary algorithms comprises a set of computational techniques that can efficiently explore very large problem spaces: instead of trying to comprehensively test all possible solutions, evolutionary algorithms seek to iteratively improve a population of candidate solutions, in a manner analogous to natural evolution. This approach has the advantage of avoiding premature optimization to local minima by the incorporation of mutation, and using a selection algorithm that chooses high-fitness genes, but not necessarily the very highest fitness genes.

Evolutionary algorithms (also known as genetic algorithms) are often used in virtual screening as a method of efficiently exploring ligand flexibility in docking [123–125]. The idea of using a evolutionary algorithm approach to screen combinatorial libraries is not new [126], and this approach has been previously applied to designing *de-novo* molecules from fragments [127, 128], or by pseudo-retrosynthesis, where a molecule is broken up into building blocks which can then be recombined [129]. This approach is equally well suited to peptide design [130–132].

10.5 Applying Evolutionary Algorithms to Cyclic Peptides

An *evolutionary algorithm* starts with a random population of "genes" that undergo *selection* based on a user-defined *fitness function*. After the selection step, the population undergoes *recombination* and *mutation* to create the next generation. The process repeats itself, either for a number of predefined rounds or until the evolutionary fitness converges to a stable value.

Table 6 describes the different method implementations that can be used for the selection, mutation and recombination stages of the evolutionary algorithm. In the case of cyclic peptides, it is convenient to encode cyclic peptides as a sequence string, which can be converted into a chemical structure using a tool such as

Table 6
Methods of applying evolutionary algorithms to cyclic peptide discovery

Evolutionary process	Description	Implementations
Selection	Individuals from a population are selected based on their fitness. To avoid local minima, it is generally desirable to not simply select the top scoring peptides, but to choose a diverse panel of peptides with above-average scores	*Proportional selection*: Peptides are selected, with a likelihood of selection weighted by their fitnesses *Linear rank selection*: Peptides are ranked according to their fitness, and selected in rank order *Binary tournament selection*: Peptides are assigned random pairings, the paired peptide with the higher fitness is selected *Q-tournament selection*: All peptides participate in "Q" number of tournaments, and the peptides with the most wins are selected
Recombination	To produce the next generation of peptides, after fitness evaluation and selection, selected peptides are shuffled and recombined in various ways	*Single, double, and multipoint crossover*: Two parent peptides are cleaved at one, two, or *n* randomly chosen points, and alternating parts of each sequence are used to create the child peptides *Distance bisector crossover*: Two parent peptides are split at the halfway point, and recombined *Uniform Crossover*: Each position within the parent sequences is assigned a random probability score. If this score exceeds a certain threshold, the amino acids are swapped *Unchanged*: Peptides are not recombined. This method can be used in combination with any of the above methods
Mutation	Amino acids within the peptide sequence can be randomly mutated to another amino acid	*Basic mutation*: Each amino acid in the peptide has a small % chance of being randomly replaced with another *AA Class mutation*: Each amino -acid in the peptide has a small % chance of being replaced with an amino acid from another chemical class (polar, nonpolar, positive charge, negative charge)

Based on algorithms described in [135–137]. Note that this is not an exhaustive list of all appropriate algorithms

CycloPs [109], for use in the fitness function. The fitness function can be almost any biologically relevant score, such as the output from the virtual screening techniques described above. A suitable fitness function is one that provides a meaningful, numerical, measurement of how "good" each peptide is which can be used to

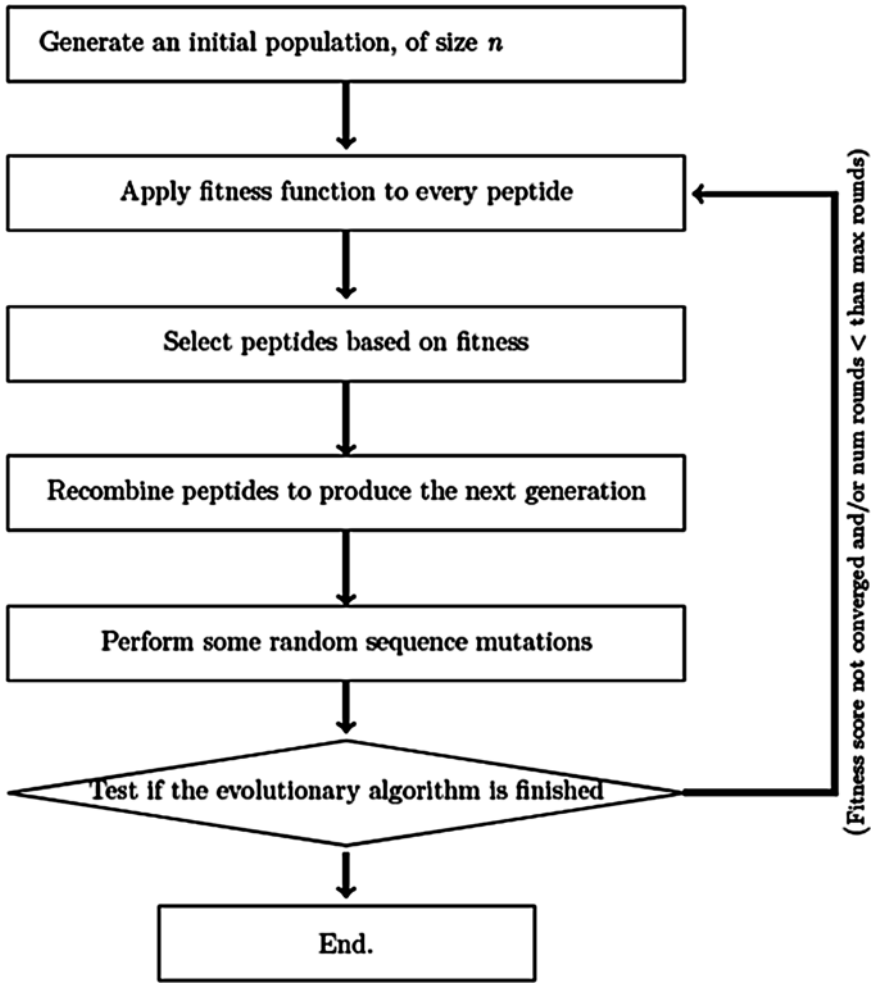


Fig. 4 Evolutionary algorithm procedure for cyclic peptide screening

compare peptides across the whole virtual population of peptides. Figure 4 outlines the basic procedure for running an evolutionary algorithm.

Thus, genetic algorithms for cyclic peptides are very similar to genetic algorithms applied to any set of features represented as a linear string, with the exception that recombination events typically are limited to double reciprocal recombinations in order to keep the size constraints of the peptides within a controlled range.

10.5.1 Selecting the Fitness Function

Generally, the most application-specific and computationally expensive step in an evolutionary algorithm will be the calculation of the fitness function after each round of selection [133]. Therefore, choosing the most appropriate function is vital if the algorithm is to converge on an optimal solution in an acceptable amount of time.

The randomness inherent the evolutionary algorithm allows the algorithm to avoid getting stuck in the local minima that characterize the solution spaces of real, complex problems. However, evolutionary algorithms do have limits. While they are efficient at exploring large problem spaces and avoiding local minima, it is very unlikely that an evolutionary algorithm will consistently discover the optimal solution to a problem. Evolutionary algorithms have the power to provide *approximate* solutions to problems that may be very difficult or impossible to exactly solve (such as finding the most potent drug for a particular target). An appropriate fitness function should be an accurate measure of how good any particular solution is to a problem.

The natural fitness functions for screening cyclic peptides are the same programs used for virtual screening of cyclic peptide libraries, including the pharmacophore screening [86–89], shape matching [82–85], and docking [92–99] programs previously mentioned. Deciding upon the most suitable approach to calculating the fitness is likely to be a process of trial and error, balancing the computational time required for each calculation with the desired generation size, and assessing which virtual screening implementation is most suited to the particular problem under investigation, as the performance of a docking program, for example, is highly dependent on the individual binding site [134].

11 Bioinformatic Discovery of Bioactive Cyclic Peptides

11.1 Biological Cyclic Peptide Libraries

An alternative approach to combinatorial virtual libraries is assembling cyclic peptide libraries from biological sources, by exploiting the diversity of cyclic peptides that have developed through evolution in the genome, or as natural products.

Natural cyclic peptides provide a useful source of guidance for virtual screening. Focusing libraries on structures based on those found in nature is a method of stacking the deck of peptides towards bioactivity while retaining manageable numbers of structures.

Harnessing protein structural information, either generated in-house, or available from public structural biology resources such as the Protein Data Bank has been used to find "self-inhibitory" peptides, where a peptide derived from a protein–protein interface inhibits the formation of that interface, and it has been observed that many protein–protein interaction surfaces are dominated by short segments of peptides [138]. More recently, this approach was used to identify peptides that inhibit viral membrane fusion [139]. While these studies principally examined short linear peptides, the same principles can be used to identify cyclic peptides, either by cyclizing bioactive linear peptides to improve bioavailability, or by searching for bioactive peptides with natural cyclic shapes, derived from loop or turn regions of protein secondary structure.

Well-known examples of the use of cyclic peptides to mimic protein loops are the RGD peptides [26]. The RGD tripeptide motif is a cell attachment β-turn motif found in numerous proteins, and cyclic peptides containing this motif have been shown to inhibit integrin αVβ3 activity, which plays an important role in tumor metastasis.

Along with structural bioinformatics approaches, harnessing evolutionary protein sequence data can be used to identify highly conserved short peptide sequences (implying biological relevance) likely to participate in a protein–protein interaction of interest. This type of bioinformatics approach has been previously successfully used to identify peptides from signaling rich juxtamembrane regions that have the ability to modulate platelet function [140]. This sort of analysis provided a rich set of template sequence which may be developed into bioactive motifs, from which libraries of cyclic peptides may be derived.

11.2 Machine Learning Approaches

Machine learning techniques involve developing a computational screen based on known data, where a computer develops a model based on generalizing from the known data in order to be able to accurately classify any new data. Supervised learning is form of machine learning where known data known to belong to a certain "class" (e.g. binder/nonbinder) is computationally processed to infer a computational model that can then classify further examples. Supervised learning algorithms include approaches such as artificial neural networks and support vector machines [141].

These techniques have been successful for a wide variety of peptide classification tasks, including signal peptide prediction [142], predicting novel antibacterial peptides [143], improving the ability of flexible peptide docking to discern binding peptides [144] and classifying peptides into binders and nonbinders based on quantitative structure–activity relationship (QSAR) descriptors for the peptides [145].

However, despite the power and success of these methods, effectively using these methods to predict bioactive peptides requires a large amount of peptide activity data to act as a training set [143], which must be determined in vitro. This requires significant laboratory work to be done prior to any computational screening.

12 Conclusions

Despite their promise for use in applications not well suited to traditional small molecules, virtual screening of cyclic peptides, and peptides in general, is not a well-explored area. This is possibly due to the known difficulties in computationally modeling peptide structures, and the known drawbacks of peptides as drugs.

Cyclic peptides are computationally more tractable than linear peptides, and present the possibility of overcoming some of the drawbacks of linear peptides.

There have been successes and proofs-of-concept showing the power and utility of virtual screening applied to cyclic peptides. Recently Norris et al [146] have shown the ability of docking to predict the affinity of angiotensin converting enzyme (ACE) inhibitory dipeptides, but did not consider larger peptides due the large number of rotational bonds. Cyclic peptides may somewhat avoid these issues, and Arbor et al [147] have successfully created a virtual library of cyclic tetrapeptides that closely mimic known three-dimensional structures of reverse-turns. There have been new developments in peptide docking with the introduction of Rosetta FlexPepDock [148], which has been shown to be able to retrieve near-native peptide conformations in a variety of docking experiments. It is, however, significantly more computationally expensive than other docking approaches. London et al [149] have used this approach to test peptides binding to Bcl-2, and validated their results using peptide arrays. Mandal et al have used docking to model the interaction of conformationally constrained phosphopeptides to the SH2 domain of the signal transducer and activator of transcription 3 (Stat3) protein—involved in aberrant growth in malignant tumor cells [150].

In contrast to the lack of cyclic peptide virtual screening studies, there exist numerous studies using virtual screening to identify nonpeptidic bioactives based on pharmacophores derived from bioactive peptides and cyclic peptides, including Urotensin II receptor antagonists [151], the somostatin receptor and thrombin receptor mimetics [152]. These pharmacophore models have been shown to have the power to produce true hits, and must be seen as attractive for cyclic peptide screening. There are also many examples of high-throughput screening using peptide libraries, such as screening cyclic peptide antibiotics [153], of peptide integrin inhibitors [154], and binding to human leukocyte antigen (HLA) class I molecules [155]. Most of these structures are well characterized, with crystal structures including binding partners available, and these structures are also accessible to virtual screening approaches.

Cyclic peptides sit in a niche between typical small molecules and larger peptides and antibodies, with some of the potential advantages and disadvantages of both. Virtual screening has not quite reached its potential, likely due to our incomplete knowledge of the fundamental nature of ligand binding [156], and must be used with an awareness of its fundamental limitations, but the pharmacophore matching and conformational prediction techniques have reached a point where their application to cyclic peptides has shown its power. Perhaps due to the general distaste for peptide drugs, nonpeptidic compounds have seemed to be the historical first choice for developing therapeutics based on a natural

bioactive peptide. This implies a possible amount of low-hanging fruit for developing cyclic peptide analogues instead. There is a wealth of peptide–protein activity data available that can be harnessed, and virtual screening is a fast way of getting started.

Acknowledgements

The authors thank Science Foundation Ireland (grant 08 IN.1 B1864) for funding this work.

References

1. Martins MB, Carvalho I (2007) Diketopiperazines: biological activity and synthesis. Tetrahedron 63:9923–9932
2. Brakhage AA (1998) Molecular regulation of beta-lactam biosynthesis in filamentous fungi. Microbiol Mol Biol Rev 62:547–585
3. Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature 450: 1001–1009
4. Huigens RW et al (2013) A ring-distortion strategy to construct stereochemically complex and structurally diverse compounds from natural products. Nat Chem. doi:10.1038/nchem.1549
5. Beck A, Wurch T, Bailly C, Corvaia N (2010) Strategies and challenges for the next generation of therapeutic antibodies. Nat Rev Immunol 10:345–352
6. Leader B, Baca QJ, Golan DE (2008) Protein therapeutics: a summary and pharmacological classification. Nat Rev Drug Discov 7:21–39
7. Chames P, Van Regenmortel M, Weiss E, Baty D (2009) Therapeutic antibodies: successes, limitations and hopes for the future. Br J Pharmacol 157:220–233
8. Roxin Á, Zheng G (2012) Flexible or fixed: a comparative review of linear and cyclic cancer-targeting peptides. Future Med Chem 4: 1601–1618
9. Driggers EM, Hale SP, Lee J, Terrett NK (2008) The exploration of macrocycles for drug discovery–an underexploited structural class. Nat Rev Drug Discov 7:608–624
10. Kotz J (2012) Bringing macrocycles full circle. Sci Exch 5
11. Schwarzer D, Finking R, Marahiel MA (2003) Nonribosomal peptides: from genes to products. 275–287. 10.1039/b111145k
12. Mullard A (2012) Protein–protein interaction inhibitors get into the groove. Nat Rev Drug Discov 11:173–175
13. Verdine GL, Hilinski GJ (2012) Stapled peptides for intracellular drug targets. Methods Enzymol 503:3–33, Elsevier Inc
14. Arrowsmith J (2011) Trial watch: phase III and submission failures: 2007-2010. Nat Rev Drug Discov 10:87
15. Snyder PW et al (2011) Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. Proc Natl Acad Sci U S A 108: 17889–17894
16. Freire E (2008) Do enthalpy and entropy distinguish first in class from best in class? Drug Discov Today 13:869–874
17. Biela A et al (2012) Ligand Binding Stepwise Disrupts Water Network in Thrombin: Enthalpic and Entropic Changes Reveal Classical Hydrophobic Effect. J Med Chem 55:6094–6110
18. Hamman JH, Enslin GM, Kotzé AF (2005) Oral delivery of peptide drugs: barriers and developments. BioDrugs 19:165–177
19. Ranade V (1991) Drug delivery systems 5A. Oral drug delivery. J Clin Pharmacol 31:2–16
20. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23:3–25
21. Rezai T, Yu B, Millhauser GL, Jacobson MP, Lokey RS (2006) Testing the conformational hypothesis of passive membrane permeability using synthetic cyclic peptide diastereomers. J Am Chem Soc 128:2510–2511
22. Biron E et al (2008) Improving oral bioavailability of peptides by multiple N-methylation: somatostatin analogues. Angew Chem Int Ed Engl 47:2595–2599
23. Ovadia O et al (2011) The effect of multiple N-methylation on intestinal permeability of cyclic hexapeptides. Mol Pharm 8:479–487

24. White TR et al (2011) On-resin N-methylation of cyclic peptides for discovery of orally bioavailable scaffolds. Nat Chem Biol 7: 810–817
25. Doedens L et al (2010) Multiple N-methylation of MT-II backbone amide bonds leads to melanocortin receptor subtype hMC1R selectivity: pharmacological and conformational studies. J Am Chem Soc 132:8115–8128
26. Dechantsreiter MA et al (1999) N-Methylated cyclic RGD peptides as highly active and selective alpha(V)beta(3) integrin antagonists. J Med Chem 42:3033–3040
27. Roberts MJ, Bentley MD, Harris JM (2012) Chemistry for peptide and protein PEGylation. Adv Drug Deliv Rev 64:116–127
28. Cefalu WT (2004) Concept, Strategies, and Feasibility of Noninvasive Insulin Delivery. Diabetes Care 27:239–246
29. Chen X, Park R, Shahinian AH, Bading JR, Conti PS (2004) Pharmacokinetics and tumor retention of 125I-labeled RGD peptide are improved by PEGylation. Nucl Med Biol 31:11–19
30. Rubio-Aliaga I, Daniel H (2002) Mammalian peptide transporters as targets for drug delivery. Trends Pharmacol Sci 23:434–440
31. Habberfield A (1996) Vitamin B12-mediated uptake of erythropoietin and granulocyte colony stimulating factor in vitro and in vivo. Int J Pharm 145:1–8
32. Rawlings ND, Morton FR, Kok CY, Kong J, Barrett AJ (2008) MEROPS: the peptidase database. Nucleic Acids Res 36:D320–D325
33. Hedstrom L (2002) Serine protease mechanism and specificity. Chem Rev 102:4501–4524
34. Rozek A, Powers J-PS, Friedrich CL, Hancock REW (2003) Structure-based design of an indolicidin peptide analogue with increased protease stability. Biochemistry 42: 14130–14138
35. Getz JA, Rice JJ, Daugherty PS (2011) Protease-resistant peptide ligands from a knottin scaffold library. ACS Chem Biol 6:837–844
36. Guichard G et al (1994) Antigenic mimicry of natural L-peptides with retro-inverso-peptidomimetics. Proc Natl Acad Sci U S A 91:9765–9769
37. Fernandez-Lopez S et al (2001) Antibacterial agents based on the cyclic D,L-alpha-peptide architecture. Nature 412:452–455
38. Young TS et al (2011) Evolution of cyclic peptide protease inhibitors. Proc Natl Acad Sci U S A 108:11052–11056
39. Wang W, Jiang J, Ballard CE, Wang B (1999) Prodrug approaches to the improved delivery of peptide drugs. Curr Pharm Des 5:265–287
40. T Borchardt R, Jeffrey A, Siahaan T, Gangwar S, Pauletti G (1997) Improvement of oral peptide bioavailability: Peptidomimetics and prodrug strategies. Adv Drug Deliv Rev 27:235–256
41. Ward P, Tippin T, Thakker D (2000) Enhancing paracellular permeability by modulating epithelial tight junctions. Pharm Sci Technolo Today 3:346–358
42. Amiram M, Luginbuhl KM, Li X, Feinglos MN, Chilkoti A (2013) Injectable protease-operated depots of glucagon-like peptide-1 provide extended and tunable glucose control. Proc Natl Acad Sci U S A. doi:10.1073/pnas.1214518110
43. Whitty A, Kumaravel G (2006) Between a rock and a hard place? Nat Chem Biol 2:112–118
44. Betzi S et al (2007) Protein protein interaction inhibition (2P2I) combining high throughput and virtual screening: Application to the HIV-1 Nef protein. Proc Natl Acad Sci U S A 104:19256–19261
45. Lo Conte L, Chothia C, Janin J (1999) The atomic structure of protein-protein recognition sites. J Mol Biol 285:2177–2198
46. London N, Movshovitz-Attias D, Schueler-Furman O (2010) The structural basis of peptide-protein binding strategies. Structure 18:188–199
47. Metz A et al (2012) Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein-protein interface. J Chem Inf Model 52:120–133
48. Arbor S, Kao J, Wu Y, Marshall GR (2008) c[D-pro-Pro-D-pro-N-methyl-Ala] adopts a rigid conformation that serves as a scaffold to mimic reverse-turns. Biopolymers 90: 384–393
49. Larregola M, Lequin O, Karoyan P, Guianvarc'h D, Lavielle S (2011) beta-Amino acids containing peptides and click-cyclized peptide as beta-turn mimics: a comparative study with "conventional" lactam- and disulfide-bridged hexapeptides. J Pept Sci 17:632–643
50. Tyndall JD, Pfeiffer B, Abbenante G, Fairlie DP (2005) Over one hundred peptide-activated G protein-coupled receptors recognize ligands with turn structure. Chem Rev 105:793–826
51. Fasan R et al (2004) Using aβ-Hairpin To Mimic anα-Helix: Cyclic Peptidomimetic Inhibitors of the p53–HDM2 Protein–Protein Interaction. Angew Chemie 116: 2161–2164
52. Gould CM et al (2010) ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res 38:D167–D180

53. Reardon DA et al (2008) Randomized phase II study of cilengitide, an integrin-targeting arginine-glycine-aspartic acid peptide, in recurrent glioblastoma multiforme. J Clin Oncol 26:5610–5617
54. Colombo G et al (2002) Structure-activity relationships of linear and cyclic peptides containing the NGR tumor-homing motif. J Biol Chem 277:47891–47897
55. Gril B et al (2007) Grb2-SH3 ligand inhibits the growth of HER2+ cancer cells and has antitumor effects in human cancer xenografts alone and in combination with docetaxel. Int J Cancer 121:407–415
56. Petsalaki E, Russell RB (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. Curr Opin Biotechnol 19:344–350
57. Vanhee P et al (2010) PepX: a structural database of non-redundant protein-peptide complexes. Nucleic Acids Res 38:D545–D551
58. Stanfield RL, Wilson IA (1995) Protein-peptide interactions. Curr Opin Struct Biol 5:103–113
59. Luckett S et al (1999) High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. J Mol Biol 290:525–533
60. Gaulton A et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107
61. Systems D. C. I. (2011) SMARTS—A language for describing molecular patterns. (2008)
62. Lamberts SW, van der Lely AJ, de Herder WW, Hofland LJ (1996) Octreotide. N Engl J Med 334:246–254
63. Andresen V et al (2007) Effect of 5 days linaclotide on transit and bowel function in females with constipation-predominant irritable bowel syndrome. Gastroenterology 133:761–768
64. Mack CM et al (2010) Davalintide (AC2307), a novel amylin-mimetic peptide: enhanced pharmacological properties over native amylin to reduce food intake and body weight. Int J Obes (Lond) 34(385–95)
65. Kallen J, Mikol V, Taylor P, Walkinshaw MD (1998) X-ray structures and analysis of 11 cyclosporin derivatives complexed with cyclophilin A. J Mol Biol 283:435–449
66. Pande J, Szewczyk MM, Grover AK (2010) Phage display : Concept, innovations, applications and future. Biotechnol Adv 28:849–858
67. Hoogenboom HR et al (1998) Antibody phage display technology and its applications. Immunotechnology 4:1–20
68. Willats WGT (2002) Phage display: practicalities and prospects. Plant Mol Biol 50(6): 837–854
69. McLafferty MA, Kent RB, Ladner RC, Markland W (1993) M13 bacteriophage displaying disulfide-constrained microproteins. Gene 128:29–36
70. Horswill AR, Benkovic SJ (2005) Cyclic peptides, a chemical genetics tool for biologists. Cell Cycle 4:552–555
71. Kritzer JA et al (2009) Rapid selection of cyclic peptides that reduce -synuclein toxicity in yeast and animal models. Nat Chem Biol 5:655–663
72. Gale EF, Taylor ES (1946) Action of tyrocidine and detergents in liberating amino acids from bacterial cells. Nature 157:549
73. Arbeit RD, Maki D, Tally FP, Campanaro E, Eisenstein BI (2004) The safety and efficacy of daptomycin for the treatment of complicated skin and skin-structure infections. Clin Infect Dis 38:1673–1681
74. Dawson R (1998) the toxicology of microcystins. Toxicon 36:953–962
75. Namikoshi M et al (1994) New nodularins: a general method for structure assignment. J Org Chem 59:2349–2357
76. Goodin S, Kane MP, Rubin EH (2004) Epothilones: mechanism of action and biologic activity. J Clin Oncol 22:2015–2025
77. Domingo GJ, Leatherbarrow RJ, Freeman N, Patel S, Weir M (1995) Synthesis of a mixture of cyclic peptides based on the Bowman-Birk reactive site loop to screen for serine protease inhibitors. Int J Pept Protein Res 46:79–87
78. Evers A, Hessler G, Matter H, Klabunde T (2005) Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols. J Med Chem 48:5448–5465
79. Warren GL et al (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49:5912–5931
80. Anderson S (1984) Graphical representation of molecules and substructure-search queries in MACCStm. J Mol Graph 2:83–90
81. Daylight Chemical Information Systems (2012) Daylight Toolkit www.daylight.com
82. Ballester PJ, Richards WG (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. J Comput Chem 28:1711–1723
83. Schreyer AM, Blundell T (2012) USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. J Cheminform 4:27
84. GRANT JA, GALLARDO MA, PICKUP BT (1996) A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. J Comput Chem 17:1653–1666

85. Sastry GM, Dixon SL, Sherman W (2011) Rapid shape-based ligand alignment and virtual screening method based on atom/feature-pair similarities and volume overlap scoring. J Chem Inf Model 51:2455–2466
86. Taminau J, Thijs G, De Winter H (2008) Pharao: pharmacophore alignment and optimization. J Mol Graph Model 27:161–169
87. Chemical Computing Group (2012) Molecule operating environment (MOE) http://www.chemcomp.com/index.htm
88. Koes DR, Camacho CJ (2011) Pharmer: efficient and exact pharmacophore search. J Chem Inf Model 51:1307–1314
89. Inc, A. S. Discovery Studio Modelling Environment (2012) http://accelrys.com/products/discovery-studio/
90. Mosca R, Pons C, Fernández-Recio J, Aloy P (2009) Pushing structural information into the yeast interactome by high-throughput protein docking experiments. PLoS Comput Biol 5:e1000490
91. Yuriev E, Agostino M, Ramsland PA (2009) Challenges and advances in computational docking: 2009 in review. J Mol Recognit 24:149–164
92. Tripos International (2010) Sybyl-X. St. Louis, Missouri. Retrieved from http://www.certara.com/products/molmod/sybyl-x.
93. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461
94. Lang PT et al (2009) DOCK 6: combining techniques to model RNA-small molecule complexes. RNA 15:1219–1230
95. Berman HM et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
96. Zsoldos Z, Reid D, Simon A, Sadjad SB, Johnson AP (2007) eHiTS: a new fast, exhaustive flexible ligand docking system. J Mol Graph Model 26:198–212
97. Viji SN, Prasad PA, Gautham N (2009) Protein-ligand docking using mutually orthogonal Latin squares (MOLSDOCK). J Chem Inf Model 49:2687–2694
98. Pearce BC, Langley DR, Kang J, Huang H, Kulkarni A (2009) E-novo: an automated workflow for efficient structure-based lead optimization. J Chem Inf Model 49:1797–1809
99. Morris GM et al (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791
100. OpenEye Scientific Software (2010) OEChem. Retrieved from http://www.eyesopen.com/oechem-tk
101. Schrodinger LLC (2012). Schrodinger. https://www.schrodinger.com/
102. Landrum G. RDKit: Open-source cheminformatics. at http://www.rdkit.org
103. O'Boyle NM et al (2011) Open Babel: An open chemical toolbox. J Cheminform 3:33
104. Steinbeck C et al (2006) Recent developments of the Chemistry Development Kit (CDK)—An open-source Java library for chemo- and bioinformatics. Curr Pharm Des 12:2111–2120
105. Guha R et al (2006) The Blue Obelisk-interoperability in chemical informatics. J Chem Inf Model 46:991–998
106. Mazanetz MP, Marmon RJ, Reisser CBT, Morao I (2012) Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. Curr Top Med Chem 12:1965–1979
107. Irwin JJ, Shoichet BK (2005) ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
108. Burns VA, Bobay BG, Basso A, Cavanagh J, Melander C (2008) Targeting RNA with cysteine-constrained peptides. Bioorg Med Chem Lett 18:565–567
109. Duffy FJ et al (2011) CycloPs: generating virtual libraries of cyclized and constrained peptides including nonnatural amino acids. J Chem Inf Model 51:829–836
110. Goldtzvik Y, Goldstein M, Benny Gerber R (2013) On the crystallographic accuracy of structure prediction by implicit water models: Tests for cyclic peptides. Chem Phys 415:168–172
111. Ponder JW (2013) Tinker: Software tools for molecular design. http://dasher.wustl.edu/ffe/
112. O'Boyle NM, Vandermeersch T, Flynn CJ, Maguire AR, Hutchison GR (2011) Confab—Systematic generation of diverse low-energy conformers. J Cheminform 3:8
113. Jacobson MP et al (2004) A hierarchical approach to all-atom protein loop prediction. Proteins Struct Funct Genet 55:351–367
114. Ebejer JP, Morris GM, Deane CM (2012) Freely Available Conformer Generation Methods: How Good Are They? J Chem Inf Model. doi:10.1021/ci2004658
115. Venkatraman V, Pérez-Nueno VI, Mavridis L, Ritchie DW (2010) Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. J Chem Inf Model 50:2079–2093
116. Merrifield RB (1963) Solid Phase Peptide Synthesis 1. Synthesis of a Tetrapeptide. J Am Chem Soc 85:2149
117. Coin I, Beyermann M, Bienert M (2007) Solid-phase peptide synthesis: from standard procedures to the synthesis of difficult sequences. Nat Protoc 2:3247–3256

118. Frank R (2002) The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports–principles and applications. J Immunol Methods 267:13–26
119. Katz C et al (2011) Studying protein-protein interactions using peptide arrays. Chem Soc Rev 40:2131–2145
120. Hann MM, Oprea TI (2004) Pursuing the leadlikeness concept in pharmaceutical research. Curr Opin Chem Biol 8:255–263
121. Dove A (2007) High-throughput screening goes to school. Nat Methods 4:523–532
122. Guerrero G, Pérez-Sánchez H, Wenzel W, Cecilia J, García, J (2011) In 5th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2011) (Rocha, M., Rodríguez, J. C., Fdez-Riverola, F. & Valencia, A.) 93:63–69 (Springer Berlin Heidelberg)
123. Oshiro CM, Kuntz ID, Dixon JS (1995) Flexible ligand docking using a genetic algorithm. J Comput Aided Mol Des 9:113–130
124. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
125. Morris GM et al (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
126. Sheridan RP, Kearsley SK (1995) Using a Genetic Algorithm To Suggest Combinatorial Libraries. J Chem Inf Model 35:310–320
127. Westhead DR et al (1995) PRO-LIGAND: an approach to de novo molecular design. 3. A genetic algorithm for structure refinement. J Comput Aided Mol Des 9:139–148
128. Schneider G, Lee ML, Stahl M, Schneider P (2000) De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. J Comput Aided Mol Des 14:487–494
129. Schneider G et al (2009) Voyages to the (un) known: adaptive design of bioactive compounds. Trends Biotechnol 27:18–26
130. Belda I et al (2005) ENPDA: an evolutionary structure-based de novo peptide design algorithm. J Comput Aided Mol Des 19:585–601
131. Hohm T, Limbourg P, Hoffmann D (2006) A multiobjective evolutionary method for the design of peptidic mimotopes. J Comput Biol 13:113–125
132. Knapp B, Giczi V, Ribarics R, Schreiner W (2011) PeptX: using genetic algorithms to optimize peptides for MHC binding. BMC Bioinformatics 12:241
133. Jin Y (2003) A comprehensive survey of fitness approximation in evolutionary computation. Soft Comput 9:3–12
134. Sousa SF, Fernandes PA, Ramos MJ (2006) Protein-ligand docking: current status and future challenges. Proteins 65:15–26
135. Baker JE (1987) Reducing bias and inefficiency in the selection algorithm. Proc Second Int Conf Genet algorithms 14–21
136. Holland JH (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT Press http://books.google.ie/books?id=cyV7nQEACAAJ
137. Back T (1998) Selective pressure in evolutionary algorithms: a characterization of selection mechanisms. Proc First IEEE Conf Evol Comput IEEE World Congr Comput Intell 57–62 doi:10.1109/ICEC.1994.350042
138. London N, Raveh B, Movshovitz-Attias D, Schueler-Furman O (2010) Can self-inhibitory peptides be derived from the interfaces of globular protein-protein interactions? Proteins 78:3140–3149
139. Xu Y (2012) Rahman, N. a B. D., Othman, R., Hu, P. & Huang, M. Computational identification of self-inhibitory peptides from envelope proteins. Proteins 80:2154–2168
140. Edwards RJ et al (2007) Bioinformatic discovery of novel bioactive peptides. Nat Chem Biol 3:108–112
141. Kotsiantis S (2007) Supervised Machine Learning: A Review of Classification Techniques. Inform 31
142. Nielsen H, Brunak S, von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng 12:3–9
143. Fjell CD et al (2009) Identification of novel antibacterial peptides by chemoinformatics and machine learning. J Med Chem 52:2006–2015
144. Khan W, Duffy F, Pollastri G, Shields DC, Mooney C (2013) Predicting Binding within Disordered Protein Regions to Structurally Characterised Peptide-Binding Domains. PLoS One 8:e72838
145. Cherkasov A (2005) Inductive Descriptors: 10 Successful Years in QSAR. Curr Comput Aided-Drug Des 1:21–42
146. Norris R, Casey F, FitzGerald RJ, Shields D, Mooney C (2012) Predictive modelling of angiotensin converting enzyme inhibitory dipeptides. Food Chem 133:1349–1354
147. Arbor S, Marshall GR (2009) A virtual library of constrained cyclic tetrapeptides that mimics all four side-chain orientations for over half the reverse turns in the protein data bank. J Comput Mol Des 23:87–95
148. Raveh B, London N, Zimmerman L, Schueler-Furman O (2011) Rosetta FlexPepDock ab-initio: simultaneous folding,

docking and refinement of peptides onto their receptors. PLoS One 6:e18934
149. London N, Gullá S, Keating AE, Schueler-Furman O (2012) In silico and in vitro elucidation of BH3 binding specificity toward Bcl-2. Biochemistry 51:5841–5850
150. Mandal PK et al (2009) Conformationally constrained peptidomimetic inhibitors of signal transducer and activator of transcription. 3: Evaluation and molecular modeling. J Med Chem 52:2429–2442
151. Flohr S et al (2002) Identification of Nonpeptidic Urotensin II Receptor Antagonists by Virtual Screening Based on a Pharmacophore Model Derived from Structure – Activity Relationships and Nuclear Magnetic Resonance Studies on Urotensin II. J Med Chem 45:1799–1805
152. Alexopoulos K et al (2001) Design, synthesis, and modeling of novel cyclic thrombin receptor-derived peptide analogues of the Ser42-Phe-Leu-Leu-Arg46 motif sequence with fixed conformations of pharmacophoric groups: importance of a Phe/Arg/NH2 cluster for receptor activation and im. J Med Chem 44:328–339
153. Xiao Q, Pei D (2007) High-throughput synthesis and screening of cyclic peptide antibiotics. J Med Chem 50:3132–3137
154. Lee Y, Kang D-K, Chang S-I, Han MH, Kang I-C (2004) High-throughput screening of novel peptide inhibitors of an integrin receptor from the hexapeptide library by using a protein microarray chip. J Biomol Screen 9:687–694
155. Harndahl M et al (2009) Peptide binding to HLA class I molecules: homogenous, high-throughput screening, and affinity assays. J Biomol Screen 14:173–180
156. Schneider G (2010) Virtual screening: an endless staircase? Nat Rev Drug Discov 9:273–276

Chapter 12

A Use of Homology Modeling and Molecular Docking Methods: To Explore Binding Mechanisms of Nonylphenol and Bisphenol A with Antioxidant Enzymes

Mannu Jayakanthan, Rajamanickam Jubendradass, Shereen Cynthia D'Cruz, and Premendu P. Mathur

Abstract

Bisphenol A (BPA) and nonylphenol (NP) are phenolic compounds used widely by the industries. BPA and NP are endocrine disruptors possessing estrogenic properties. Several studies have reported that BPA and NP induce oxidative stress in various organs or cell types in animals, by inhibiting the activities of antioxidant enzymes like catalase, superoxide dismutase, glutathione peroxidase, and glutathione reductase. However, it is not understood how BPA and NP interact with these enzymes and inhibit their functions. Hence, it would be significant to check, whether binding sites are available for NP and BPA in antioxidant enzymes. In the present study three-dimensional structures of antioxidant enzymes, catalase, superoxide dismutase, glutathione peroxidase, and glutathione reductase were modeled and docked with BPA and NP. Docking studies revealed that BPA and NP have binding pockets in the antioxidant enzymes. Among the antioxidant enzymes, Catalase was maximally inhibited by BPA and superoxide was maximally inhibited by NP.

Key words Catalase, Superoxide dismutase, Glutathione peroxidase, Glutathione reductase

1 Introduction

Molecular docking is an automated computational technique employed in computer-aided drug design for identification of potent bioactive agents. It operates by identifying the best binding mode of given ligand with its macromolecular target and evaluates the binding affinity, results of which were used in ranking the best interacting ligands and in selecting the promising bioactive compounds. Success of a molecular docking algorithm depends on the implication of efficient search method, which is used for exploring the potential energy landscape of ligands for finding their optimum configuration, accompanied with the proper scoring scheme for evaluating the binding modes of the ligand with their targets. In rigid docking methods, search algorithm employs the rotation and translational functions to

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_12,

locate low energy configurations, while in flexible docking methodology, sampling of various conformations of ligand is performed by altering their torsion angles. Besides, the concept of flexibility is also applied in a method called induced fit docking for searching the possible best interacting conformations of amino acids. Molecular docking is widely used for screening the libraries of compounds, for selection of best interacting hits, and in the design of novel leads based on the available drug molecules [1, 2].

In the recent years, there has been much concern regarding the adverse effects of various environmental contaminants on human health. With the advent of industrialization, economic development, and urbanization drastic changes occurred in the lifestyle and surroundings of humans, which resulted in the extensive production and use of substances that could facilitate life. As a result, many potentially hazardous chemicals got released into the environment at an alarming rate and exposure to these chemicals has become inevitable. These chemicals released into the environment turned out to be one of the leading causative factors for the high incidence of various pathological conditions [3, 4]. Of the various chemicals, bisphenol (BPA) and Nonylphenol (NP) are phenolic compounds possessing estrogenic property. BPA is a plastic monomer used in the manufacture of polycarbonate plastics and epoxy resins. Polycarbonate plastics are lightweight, tough, and optically clear plastics, which are used to make various consumer products such as baby bottles, water bottles, toys, and medical equipment, whereas epoxy resins find application as protective coatings in dental sealants and food and beverage containers [5–7]. NP is the final degradation product of alkylphenol polyethoxylates, which are widely used in the preparation of lubricating oil additives, resins, plasticizers, surface-active agents, detergents, paints, cosmetics, and other formulated products. Humans are constantly exposed to BPA and NP by contaminated water and food products. Lipophilic nature of BPA and NP leads to its accumulation in animal tissues [8–10]. BPA and NP have been associated with various abnormal functions of vital system such as endocrine, reproductive, and immune systems in wild life and humans. The toxicity of BPA and NP has been widely reported due to disturbance of prooxidant/ antioxidant balance of cells [11–13]. Antioxidants are located throughout the cells and provide protection against ROS. Superoxide dismutase (SOD), catalase, glutathione peroxidase, and glutathione reductase are powerful enzymatic antioxidant enzymes in maintaining the fine balance between the pro-oxidants/antioxidants in cells. SOD is considered as the first line of defense against oxyradicals in cells by catalyzing dismutation of O_2^- anion to hydrogen peroxides and molecular oxygen. Catalase is present in the peroxisomes of nearly all aerobic cells and serves to protect the cells from toxic effects of hydrogen peroxide by catalyzing decomposition of H_2O_2 into water and oxygen. Glutathione peroxidase and glutathione reductase also protect the

cells from highly toxic hydrogen peroxides by converting it into water and oxygen [14–16].

Several studies have shown that BPA and NP can decrease the activities of antioxidant enzymes in various organs [17–22]. However, it is not known whether BPA and NP can directly interact with antioxidant enzymes and impair their functions. Hence it would be interesting to see if such binding pockets are available in antioxidant enzymes. Therefore, in the present study three-dimensional structures of SOD, catalase, glutathione peroxidase, and glutathione reductase are modeled and the binding affinities of these antioxidant enzymes with BPA and NP were compared. Such data would help to identify whether NP and BPA can directly interact with the proteins and affect their functions.

2 Materials

2.1 Homology Modeling of Antioxidant Enzymes

1. UniProt database (http://www.uniprot.org/), to access protein sequence and functional information.
2. NCBI-BLASTp, helps to identify the similar sequences from biological sequence databases.
3. RCSB Protein Data Bank, a database of experimental macromolecular structures (*see* **Note 1**).
4. Align123, a pairwise alignment program implemented in Discovery Studio 3.1 (Accelrys Software, http://accelrys.com/) (*see* **Note 2**).
5. MODELLER, a program for protein structure modeling implemented in Discovery Studio 3.1 (*see* **Note 3**).
6. PROCHECK is a tool for analyzing stereochemical quality of a protein structure (*see* **Note 4**).
7. ProSA-web is used for calculating overall and local model quality score for a specific protein structure (*see* **Note 5**).

2.2 Molecular Docking of Antioxidant Enzymes

1. Glide: a module of Maestro 9.1 software (Schrödinger Software Suite 2011) for ligand-receptor docking.
2. Protein Preparation Wizard, an interface in Maestro 9.1 for initial preparation of protein for docking study.
3. LigPrep, a module of Maestro 9.1 for ligand preparation in docking study.
4. PubChem, a compound database, contains validated chemical structure information.
5. Grid Generation Wizard, an interface in Maestro 9.1 for generation of macromolecular grid.
6. Glide extra precision docking, a method in Maestro 9.1 for ligand-receptor docking.

3 Methods

3.1 Homology Modeling of Antioxidant Enzymes

The homology modeling of antioxidant enzymes was carried out in the following six different steps (Fig. 1).

1. The amino acid sequences of SOD (P07632), catalase (P04762), glutathione peroxidase (P04041), and glutathione reductase (P70619) were retrieved from UniProt database (http://www.uniprot.org/).
2. NCBI-BLASTp was used to search for suitable template structures in the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/). Experimental structures such as human SOD (PDB ID: 2C9V), human erythrocyte catalase (PDB ID: 1QQW), bovine glutathione peroxidise (PDB ID: 1GP1), and human glutathione reductase (PDB ID: 1XAN) were chosen as suitable templates for homology modeling of SOD, catalase, glutathione peroxidase, and glutathione reductase, respectively. These template structures were downloaded from PDB database.
3. Sequence alignment of input model sequences and template structures were carried out using Align123 program.
4. MODELLER automodel was used to build the homology models. We have generated five models for each of the protein

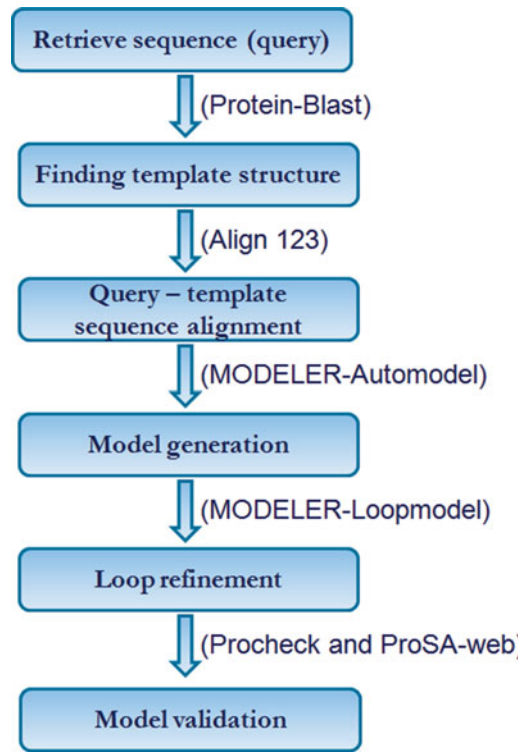


Fig. 1 Workflow for homology modeling: Steps involved in homology modeling of antioxidant enzymes

sequences. These models were generated along with cofactors such as metal ions "copper (Cu) and zinc (Zn)" for SOD; prosthetic group "heme" for catalase; and prosthetic group "FAD" for glutathione reductase. The functional water molecule, which coordinated with copper ion, was copied from the template into the rat SOD.

5. Loop refinement, in which, among the five models generated for each enzyme, the lowest DOPE score model was selected for loop refinement purposes. The loop refinement was carried out using MODELLER loop model in Discovery Studio 3.1.
6. Finally, structure validation was carried out using SAVES-PROCHECK and ProSA-web software tools [23, 24] by submitting modeled protein structures (*see* **Note 6**).

3.2 Molecular Docking of Antioxidant Enzymes

The docking study was carried out in four different steps (Fig. 2).

1. Preparations of protein, in which, the modeled structures of SOD (*see* **Note 7**), catalase (*see* **Note 8**), glutathione peroxidase (*see* **Note 9**), and glutathione reductase (*see* **Note 10**) were optimized for structure using "Protein Preparation Wizard" of the Maestro 9.1 software (Schrödinger Software Suite 2011). The structures were processed to assign bond order and hydrogens. The functional water molecule of rat SOD was maintained in the cavity site. The structures were optimized using exhaustive sampling method, in which orientation of hydroxyl groups, amide groups of Asn and Gln, and imidazole ring of His residues were optimized and energy was minimized by applying OPLS_2005 force field.
2. Preparation of ligands using LigPrep module, the ligand structures such as bisphenol A (CID:6623) and 4-nonylphenol

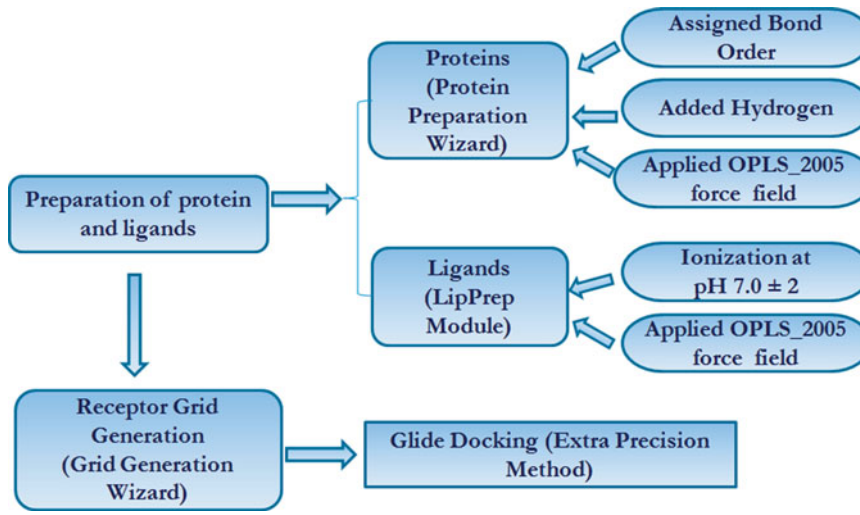


Fig. 2 Workflow for molecular docking: Steps involved in molecular docking of antioxidant enzymes

Table 1
List of active residues used for Glide grid generations

Protein	Defined active site residues	References
Superoxide dismutase [Cu-Zn]	His64, Arg144	[25, 26]
Catalase	Asn148	[27]
Glutathione peroxidase 1	Sec47	[28]
Glutathione reductase	His413	[29, 30]

(CID:1752) were obtained from PubChem compound database. These ligands were used to generate possible ionization state at pH 7.0 ± 2.0 by applying OPLS_2005 force field and one low energy ring conformation per ligand to further process into docking study.

3. The generation of "Grids" to represent the volume of the protein receptor that can be used to search for ligand docking. This step was carried out using "Grid Generation Wizard." We have set grid box on the centroid of the active site residues that are mentioned in Table 1.
4. Finally, glide docking was carried out, in which, glide extra precision method was applied to the generated grids to run docking schema (*see* **Notes 11–16**).

4 Notes

1. Homology modeling is a method to predict the three-dimensional structure of a protein sequence when the experimental structure is not available. Since the lack of experimental structures for protein sequences of rat SOD [Cu-Zn], catalase, glutathione peroxidase 1, and glutathione reductase were not available in PDB, homology modeling was used to predict its 3D structures.
2. The minimum requirement for a homology modeling is the availability of 30 % sequence identity between the input sequence and the template structure. The alignment results produced 82.5 % identity and 89 % similarity for SOD, 89 % identity and 95.8 % similarity for catalase, 83.7 % identity and 88.3 % similarity for glutathione peroxidase 1, and 84.3 % identity and 78.1 % similarity for glutathione reductase with their respective template structures. Thus the alignment results were reliable for predicting the three-dimensional structures of the proteins.
3. The models generated using the alignment results were validated. Three criteria were adopted for protein structure validation. One was checking stereochemical quality of using

Ramachandran Plot, second was checking *Z*-score (indicates overall model quality) of the modeled structure and comparison with the *Z*-score of the experimentally determined protein structures of current PDB. Finally, local model quality was plotted by means of finding energies for each of the amino acid position.

4. The Ramachandran plot for each of the modeled structures (Fig. 3, Table 2) shows that more that 91 % of the residues is present in the most favored region of the plot. In this validation, no amino acid residues were present in generously allowed region or disallowed region of the plot. This confirms

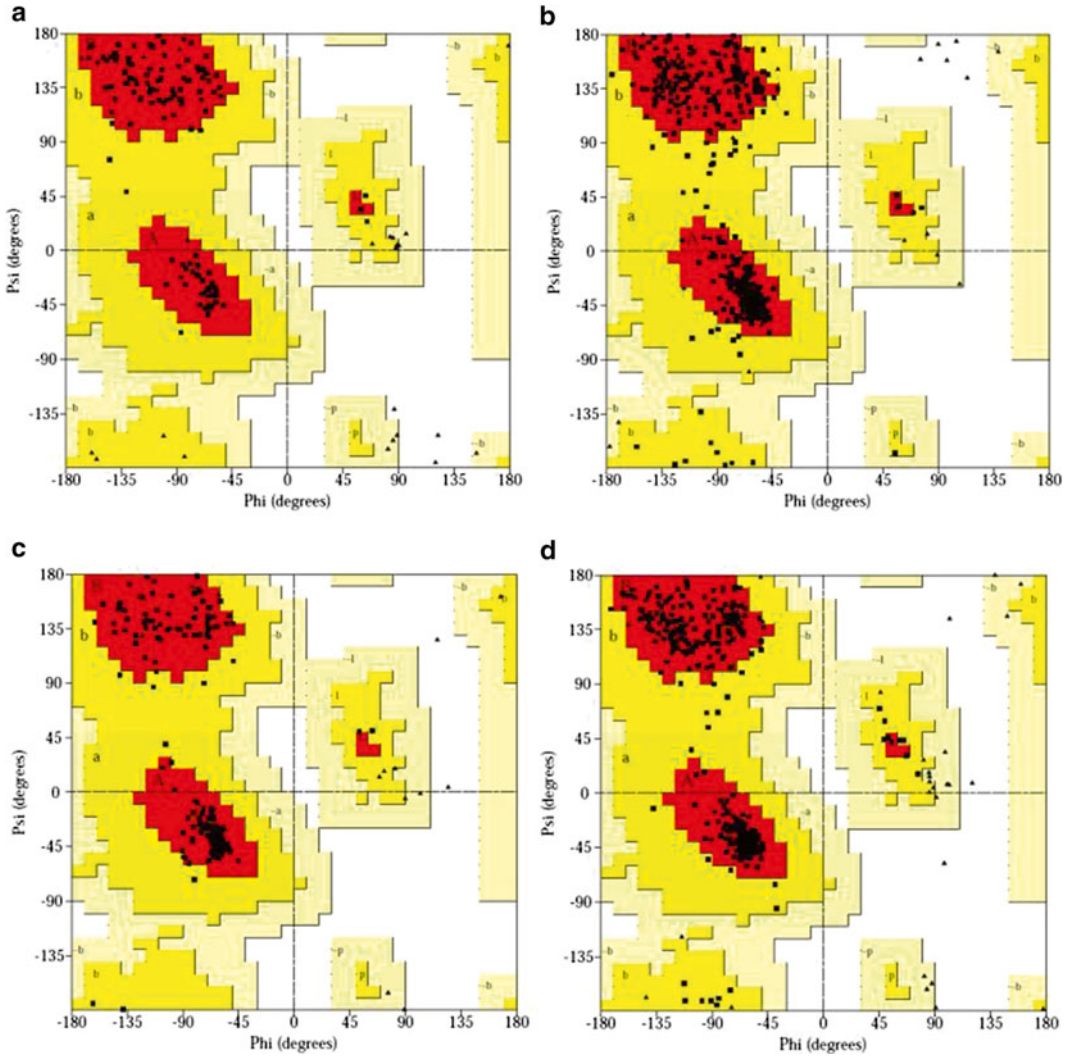


Fig. 3 Illustration of Ramachandran plot of the 3-D models (**a**) superoxide dismutase [Cu-Zn], (**b**) catalase, (**c**) glutathione peroxidase 1, and (**d**) glutathione reductase. These plots represent that none of the residues present in disallowed region (*white area*) except proline and glycine (shown as *triangles*). The plots were developed using SAVES-PROCHECK tool

Table 2
Statistical data on Ramachandran Plot and *Z*-Score values of modeled ROS scavenging enzymes

Name of the protein	Percentage of residues in most favored regions [A,B,L]	Percentage of residues in additional allowed regions [a,b,l,p]	Percentage of residues in generously allowed regions [~a,~b,~l,~p]	Percentage of residues in disallowed regions	ProSA-web *Z*-score
Superoxide dismutase [Cu-Zn]	94.2	5.8	0.0	0.0	–6.92
Catalase	90.6	9.4	0.0	0.0	–9.44
Glutathione peroxidase 1	91.4	8.6	0.0	0.0	–6.14
Glutathione reductase	92.1	7.9	0.0	0.0	–8.86

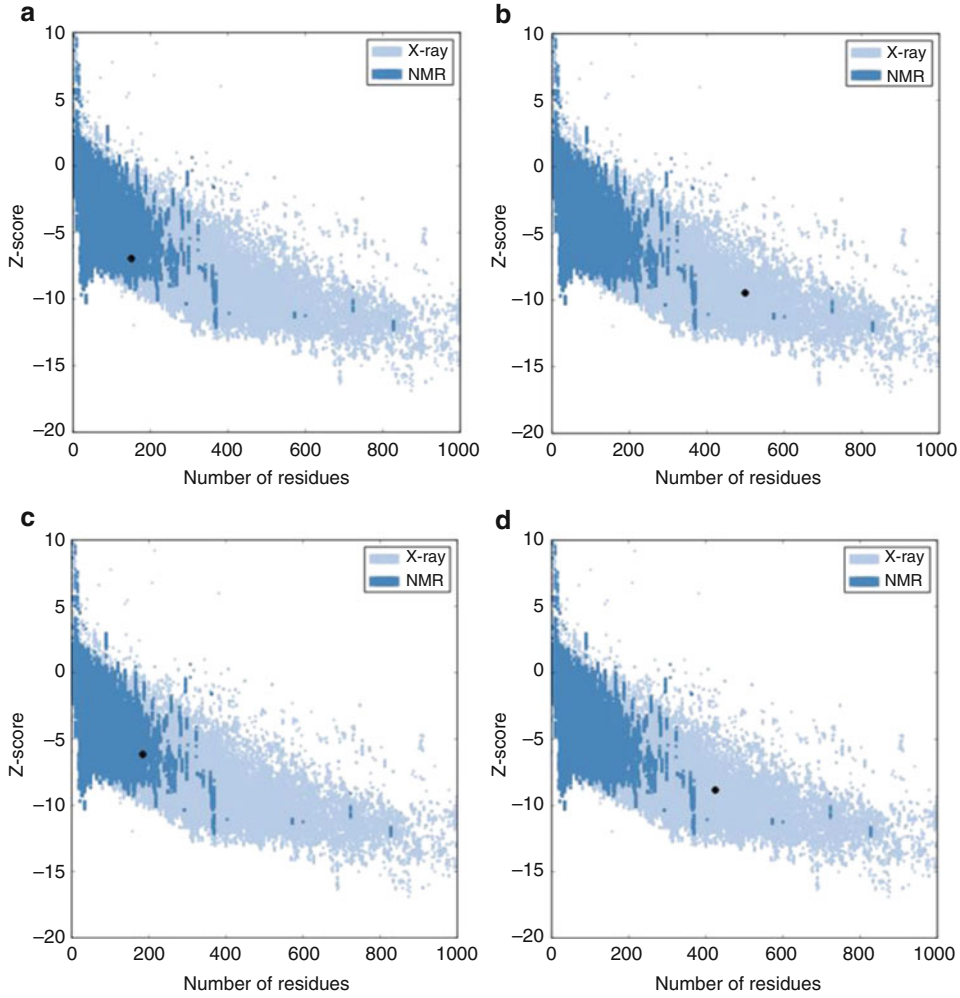


Fig. 4 Validation of overall model quality of the 3-D models (**a**) superoxide dismutase [Cu-Zn], (**b**) catalase, (**c**) glutathione peroxidase 1, and (**d**) glutathione reductase using ProSa-web service. The *black spot* indicates that the input modeled structures are within the range of *Z*-score values of experimental structures with respect to number of residues

that the predicted model is similar to the standard experimental structures.

5. ProSA-web was used to calculate *Z*-score values of the modeled structures (Table 2) and displayed in a plot by incorporating the number of amino acid residues against *Z*-score values of the experimental structures (Fig. 4). The calculated *Z*-score values for each input modeled structures were depicted in the dark black spot in the plot. These plots confirm that the calculated *Z*-score values for the modeled structures were within the range characteristic to the similar size of experimental structures (viz., NMR and X-ray).
6. Similarly, the plots for local model quality were developed by plotting the values of knowledge-based energy against sequence positions to depict erroneous parts of a model (Fig. 5).

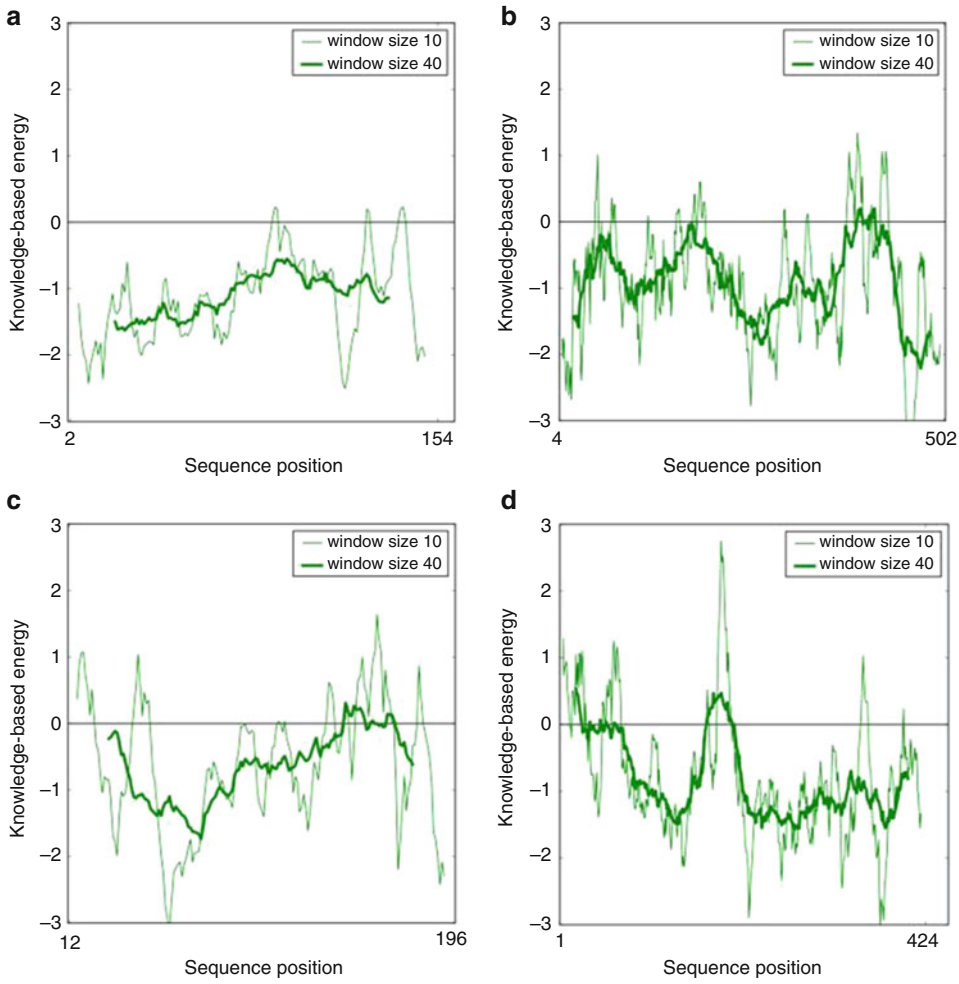


Fig. 5 Validation of local model quality of the 3-D models (**a**) superoxide dismutase [Cu-Zn], (**b**) catalase, (**c**) glutathione peroxidase 1, and (**d**) glutathione reductase using ProSa-web service. *Negative energy values* or *equal to zero* correspond to error free region of the structure

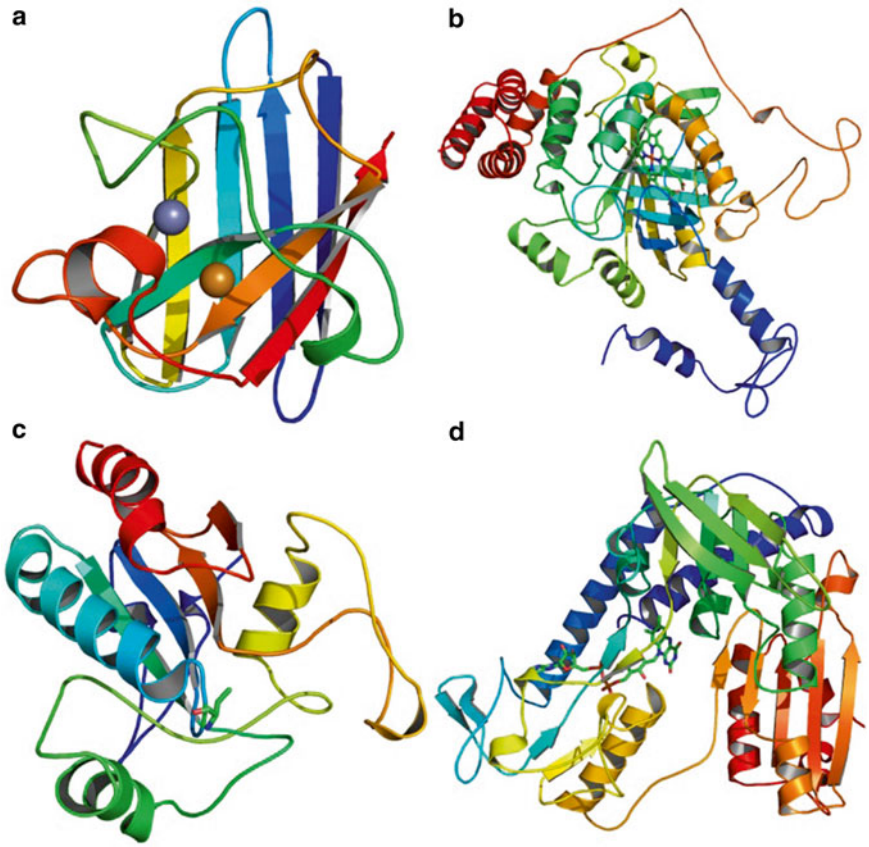


Fig. 6 Cartoon representations of modeled structures of (**a**) superoxide dismutase [Cu-Zn], (**b**) catalase, (**c**) glutathione peroxidase 1, and (**d**) glutathione reductase. PyMOL software was used to depict the modeled structures. The cofactors were represented in *stick model*. These cofactors are (**a**) zinc (shown as *gray ball*) and copper (shown as *orange ball*) (**b**) Heme prosthetic group (**d**) FAD. Whereas (**c**) contains selenocysteine amino acid residue

In general, negative values correspond to the reliable part of the model. These plots show the negative values of energy particularly in the windows size of 40 (i.e., the average energy value for 40 amino acid residues at position *i*) which confirm that the modeled structures are reliable. Hence, these validated protein structures (Fig. 6) were further used for molecular docking studies to interpret the interaction results of BPA and NP.

7. The modeled structures of antioxidant enzymes were taken for docking studies with bisphenol A (BPA). The molecular docking of BPA with superoxide dismutase (SOD) reveals the formation of three different types of interactions: hydrogen bond, Van der Waals interactions, and Pi interactions. His64, Glu133, and Lys137 form hydrogen bonds, His49, Pro63, and His72 form Van der Waals interactions, and Lys137 forms Pi interaction with BPA (Table 3). In addition, Zn, Cu, and water

Table 3
The binding residues of ROS enzymes involved in interactions with bisphenol A and 4-noylphenol

Protein complexes	H-bond interaction	vdW interaction	Pi interaction	Docking score	Glide energy (kcal/mol)
Bisphenol A					
Superoxide dismutase [Cu-Zn]	HIS64, GLU133, LYS137	HIS49, PRO63, HIS72	LYS137	–4.120	–25.785
Catalase	TRP303, THR445	PRO304, PHE198, PRO151, VAL302, ILE152, HIS194, VAL450, HIS235, HIS305, LYS306	–	–5.536	–32.103
Glutathione peroxidase 1	GLY48, GLN82	THR49, TYR147, SER159	ARG179	–3.217	–22.695
Glutathione reductase	TYR27, SER416	ILE384, PRO351,VAL356, PHE349, PRO414	–	–4.601	–30.632
4-nonylphenol					
Superoxide dismutase [Cu-Zn]	HIS64, LYS137	PRO63, PHE65, ASN66, PRO67, SER69, HIS81, ARG116	–	–5.052	–25.917
Catalase	TRP303	THR150, PRO151, ILE152, HIS194, GLN195, PHE198, ARG203, TYR215, HIS235, VAL302, PRO304, HIS305, LYS306, PHE446, VAL450	LYS306	-2.383	-23.217
Glutathione peroxidase 1	ARG179	GLY48, THR49, LEU141, MET142, THR143, ASP144, TYR147, SER159, TRP160, ASN161	ARG52, ARG179	0.612	–22.636
Glutathione reductase	GLU419	PHE349, PRO351, MET352, TYR353, ILE384, GLY385, MET389, HIS413, PRO414, THR415, SER416, SER417, GLU418	–	–1.804	–29.438

molecules coordinate at the active site cavity of superoxide dismutase to bind with BPA (Fig. 7A (a, b)).

8. The docking of BPA with catalase forms a stable binding complex by producing a docking score of –5.536 and glide energy of –32.1.3 kcal/ mol. There are two hydrogen bonds formed by Trp303 and Thr445 residues along with Van der Waals interactions by Pro304, Phe198, Pro151, Val302, Ile152, His194, Val450, His235, His305, and Lys306. The 2D representation of this docked complex (Fig. 7B (a)) shows clearly that the binding of BPA is more stable due to the involvement of large number of interacting residues. The prosthetic group, heme, is completely buried inside the core of the protein and also next to the cavity site (Fig. 7B (b)). The highest docking score and glide energy also implies that bisphenol A has more stable interactions towards catalase when compared to SOD.
9. The binding site amino acid residues Gly48, Thr49, Gln82, Tyr147, Ser159, and Arg179 are involved in the interactions of bisphenol A with glutathione peroxidase. The selenocysteine residue (Sec47) coordinates with BPA at the protein cavity site (Fig. 7C (a, b)). The docking complex of glutathione peroxidase with BPA shows low docking score and low glide energy as compared to other antioxidant enzymes (Table 3). Thus, stability of the binding complex could also be less.
10. The docking of BPA with glutathione reductase results in the formation of two hydrogen bonds with Tyr27 and Ser416. In addition, Van der Waals interactions were formed with Ile384, Pro351, Val356, Phe349, and Pro414 amino acid residues to further establish the binding complex (Fig. 7D (a, b)). The docking score and glide energy of the complex were –3.217 and –22.695 kcal/mol, respectively (Table 3). The glide energy shows less variation compared to catalase enzyme. Hence, both glutathione reductase and catalase are the enzymes which are maximally inhibited following BPA exposure. In our previous experimental findings, we also found that the percentage inhibition of catalase activity following BPA exposure was maximum as compared to other antioxidant enzymes [18]. Hence we imply that BPA has strong interaction with catalase and inhibit its activities.
11. Similarly, the inhibitory activity of NP on antioxidant enzymes were checked by docking studies. The molecular docking of NP

Fig. 7 (continued) represents Van der Waals interactions residues; *Cyan circle* represents water molecule; *Gray circle* represents zinc metal; *Blue halo* around residues represent solvent accessible area. *Green* and *blue dotted lines* represent hydrogen bond interactions with amino acid main chain and side chain residues respectively. *Orange line* represents Pi interactions. Discovery Studio 3.1 was used to depict 2D diagram. In 3-D representations, proteins are represented in carton model. The binding cite cavity is represented in *violet color surface model* with bisphenol A in *stick representation* and colored as per element (*Cyan*-Carbon; *Red*-Oxygen; *Blue*-Nitrogen). PyMOL software was used to depict 3-D representations

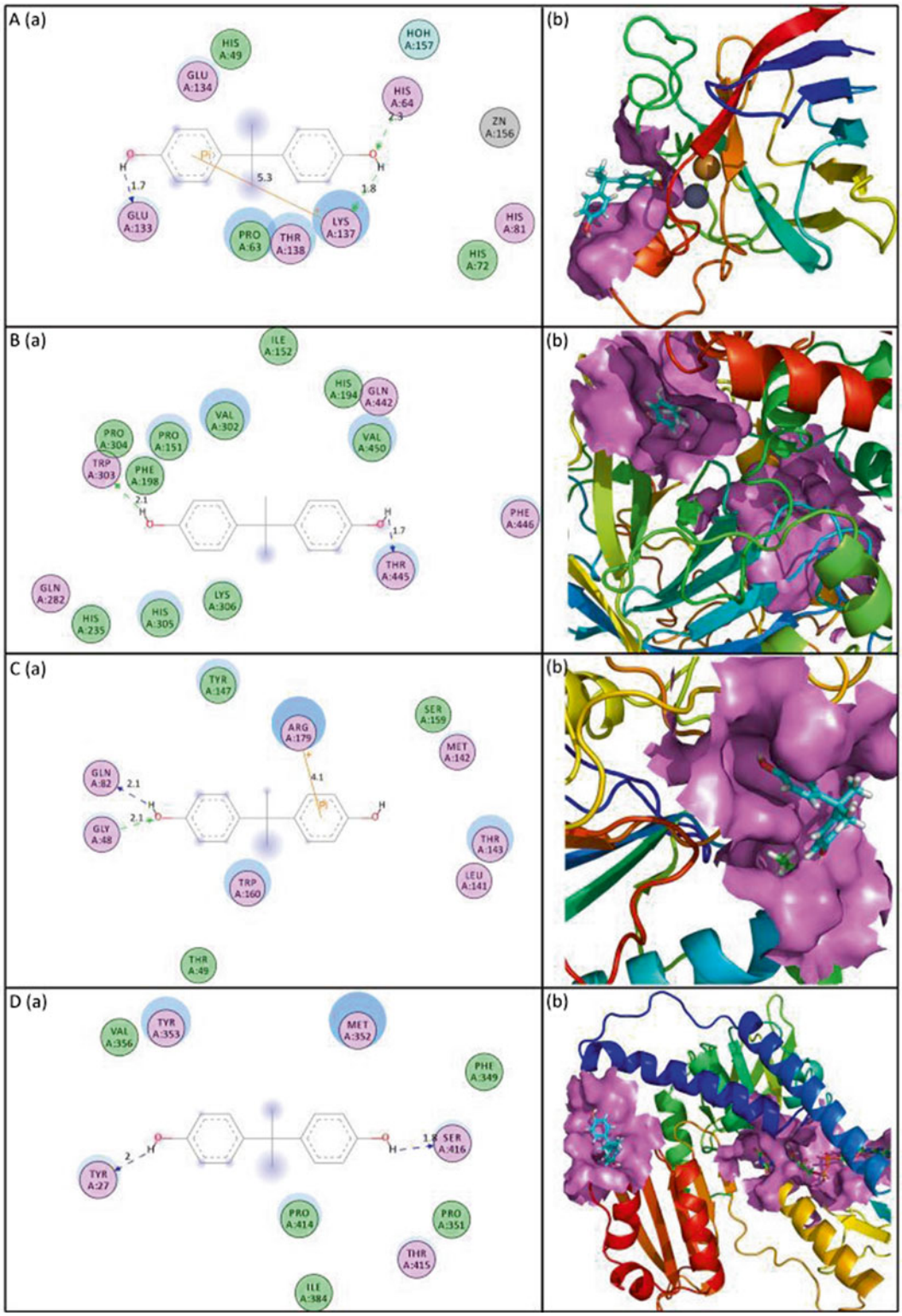


Fig. 7 Illustration of 2-D (a) and 3-D (b) representations of docked complex of bisphenol A with (**A**) superoxide dismutase [Cu-Zn], (**B**) catalase, (**C**) glutathione peroxidase 1, and (**D**) glutathione reductase. In 2-D representations, *Pick circle* represents residues involved in hydrogen bond, polar or charged interactions; *green circle*

with SOD reveals the formation of two different types of interactions: hydrogen bond and Van der Waals interactions. The amino acid residues His64 and Lys137 form hydrogen bonds. Pro63, Phe65, Asn66, Pro67, Ser69, His81, and Arg116 form Van der Waals interaction as shown in Fig. 8A (a, b)). The docking score was −5.052 and glide energy was −25.917 kcal/mol. Docking score and glide are very high with SOD when compared with other antioxidant enzymes.

12. The molecular docking of NP with catalase shows hydrogen bond formation with Trp303, one Pi cation interaction with Lys306 and Van der Waals interactions with Thr150, Pro151, Ile152, His194, Gln195, Phe198, Arg203, Tyr215, His235, Val302, Pro304, His305, Lys306, Phe446, and Val450 (Fig. 8B (a)) The docking score is comparatively low to SOD docking complex.
13. The binding site amino acid residues Gly48, Thr49, Arg52, Leu141, Met142, Thr143, Asp144, Tyr147, Ser159, Trp160, Asn161, and Arg179 are involved in the interactions of NP with glutathione peroxidase. The selenocysteine residue (Sec47) coordinates with NP at the protein cavity site (Fig. 8C (a, b)).
14. The docking complex of glutathione peroxidase with NP shows low docking score and low glide energy as compared to other antioxidant enzymes. Thus, stability of the binding complex could also be less. NP forms single hydrogen bond with Glu419 of glutathione reductase, in addition to the Van der Waals interactions formed by Phe349, Pro351, Met352, Tyr353, Ile384, Gly385, Met389, His413, Pro414, Thr415, Ser416, Ser417, and Glu418 (Fig. 8D (a, b)). The docking score of the complex is −1.804. This low docking score indicates that stability of the binding of NP with glutathione reductase is very less.
15. The highest docking score and glide energy revealed that NP has more stable interactions towards SOD when compared to other enzymes. Several studies have demonstrated the induction of oxidative stress in various organs following NP exposure by inhibition of antioxidant enzymes [12, 20, 21]. In our previous study, we observed a significant decline in the activities of SOD, catalase, glutathione peroxidase, and glutathione reductase in the pancreas and liver of NP-treated rats when compared to the corresponding group of control animals [22]. Of these we found that maximum inhibition was shown by SOD. Hence we imply that NP has strong interaction with SOD and inhibit its activities.
16. We conclude that antioxidant enzymes viz. superoxide dismutase, catalase, glutathione peroxidase, and glutathione reductase have favorable binding pockets for interactions with NP and BPA. Catalase was maximally inhibited by BPA, and similarly superoxide dismutase could be maximally inhibited by NP.

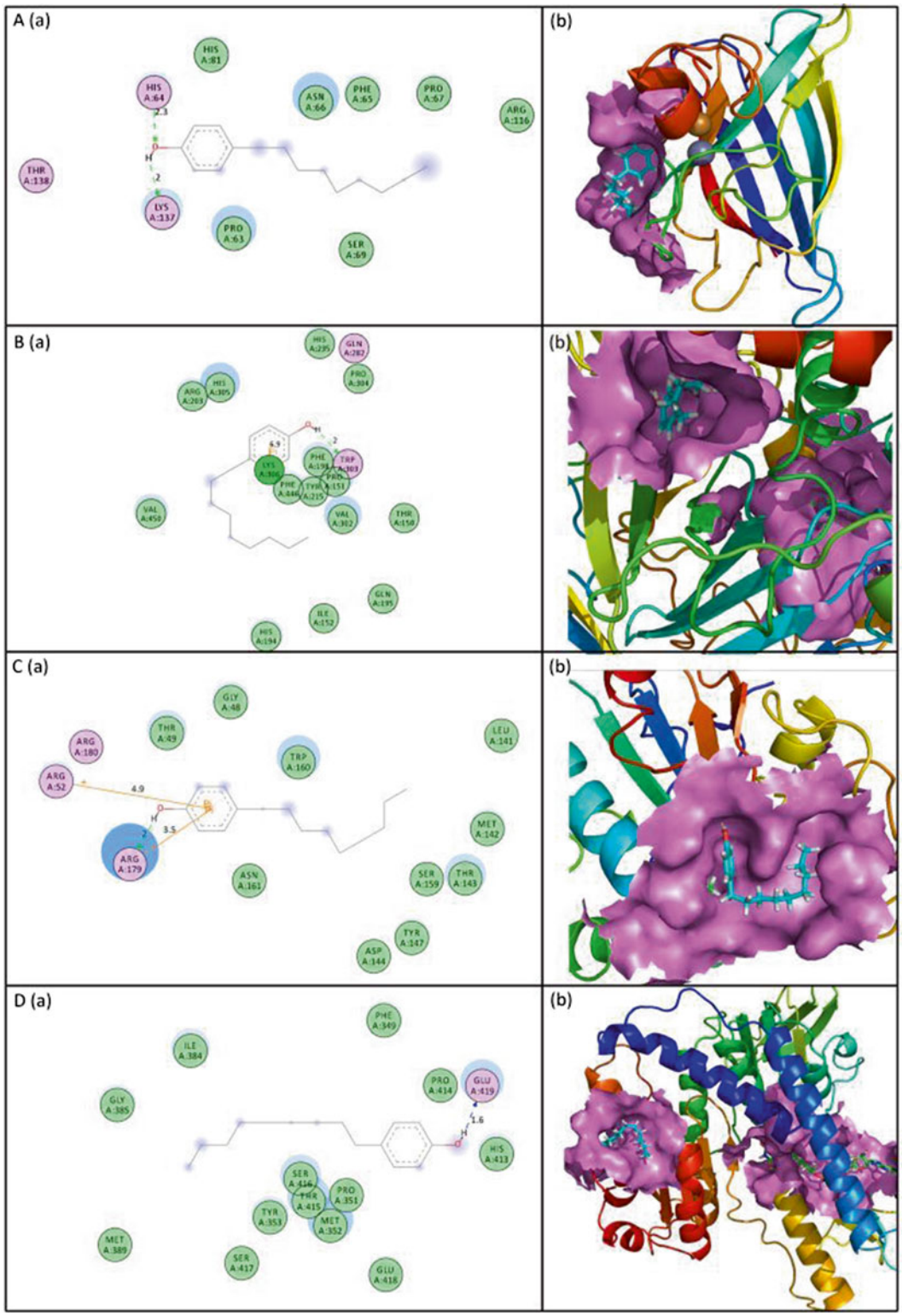


Fig. 8 Illustration of 2-D (a) and 3-D (b) representations of docked complex of 4-nonylphenol with (**A**) superoxide dismutase [Cu-Zn], (**B**) catalase, (**C**) glutathione peroxidase 1, and (**D**) glutathione reductase. (See legend of Fig. 5 for residue color code and structural representations)

Acknowledgements

P. P. Mathur acknowledges the receipt of financial support from the University Grants Commission, New Delhi (F. No. 32-600/2006), Department of Science and Technology, Government of India under the projects (SP/SO/B-65/99) and DST-FIST. R. Jubendradass acknowledges Department of Science and Technology, New Delhi for an Inspire fellowship. Shereen Cynthia D'Cruz acknowledges the Indian Council of Medical Research, New Delhi, India for a Senior Research Fellowship. The authors also thank the staff of the Centre for Bioinformatics, Pondicherry University, Pondicherry, for providing various facilities.

Declaration of interest:
The authors report no conflicts of interest.

References

1. Morris GM, Lim-Wilby M (2008) Molecular docking. In: Andreas K (ed) Molecular modeling of proteins, Methods in molecular biology. Humana, Totowa, NY, pp 365–382
2. Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 32:335–373
3. Clapp RW, Jacobs MM, Loechler EL (2008) Environmental and occupational causes of cancer: new evidence 2005–2007. Rev Environ Health 23:1–37
4. Irigaray P, Newby JA, Clapp R et al (2007) Lifestyle-related factors and environmental agents causing cancer: an overview. Biomed Pharmacother 61:640–658
5. Vandenberg LN, Hauser R, Marcus M et al (2007) Human exposure to bisphenol A (BPA). Reprod Toxicol 24:139–177
6. Howdeshell KL, Peterman PH, Judy BM et al (2003) Bisphenol A is released from used polycarbonate animal cages into water at room temperature. Environ Health Perspect 111: 1180–1187
7. Kang JH, Kito K, Kondo F (2003) Factors influencing the migration of bisphenol A from cans. J Food Prot 66:1444–1447
8. Foran CM, Bennett ER, Benson WH (2000) Exposure to environmentally relevant concentrations of different nonylphenol formulations in Japanese medaka. Mar Environ Res 50:135–139
9. Guenther K, Heinke V, Thiele B et al (2002) Endocrine disrupting nonylphenols are ubiquitous in food. Environ Sci Technol 36:1676–1680
10. Saito I, Onuki A, Seto H (2004) Indoor air pollution by alkylphenols in Tokyo. Indoor Air 14:325–332
11. Hanioka N, Jinno H, Tanaka-Kagawa T et al (2000) Interaction of bisphenol A with rat hepatic cytochrome P450 enzymes. Chemosphere 41:973–978
12. Chitra KC, Mathur PP (2004) Vitamin E prevents nonylphenol-induced oxidative stress in testis of rats. Indian J Exp Biol 42:220–223
13. Yasemin SK, Recep A (2010) Taurine prevents nonylphenol-induced oxidative stress in rats. J Anim Vet Adv 9:37–43
14. Kanner J, German JB, Kinsella JE (1987) Initiation of lipid peroxidation in biological systems. Crit Rev Food Sci Nutr 25:317–364
15. Betteridge DJ (2000) What is oxidative stress? Metabolism 49:3–8
16. Sies H (1997) Oxidative stress: oxidants and antioxidants. Exp Physiol 82:291–295
17. Chitra KC, Rao KR, Mathur PP (2003) Effect of bisphenol A and co-administration of bisphenol A and vitamin C on epididymis of adult rats: a histological and biochemical study. Asian J Androl 5:203–208
18. Bindhumol V, Chitra KC, Mathur PP (2003) Bisphenol A induces reactive oxygen species generation in the liver of male rats. Toxicology 188:117–124
19. Kabuto H, Amakawa M, Shishibori T (2004) Exposure to bisphenol A during embryonic/fetal life and infancy increases oxidative injury and causes underdevelopment of the brain and testis in mice. Life Sci 74:2931–2940
20. Mao Z, Zheng YL, Zhang YQ (2010) Behavioral impairment and oxidative damage induced by chronic application of nonylphenol. Int J Mol Sci 12:114–127
21. Aydogan M, Korkmaz A, Barlas N et al (2008) The effect of vitamin C on bisphenol A, nonylphenol

and octylphenol induced brain damages of male rats. Toxicology 249:35–39
22. Chitra KC, Latchoumycandane C, Mathur PP (2002) Effect of nonylphenol on the antioxidant system in epididymal sperm of rats. Arch Toxicol 76:545–551
23. Laskowski RA, MacArthur MW, Moss DS et al (1993) PROCHECK – a program to check the stereochemical quality of protein structures. J Appl Cryst 26:283–291
24. Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35:W407–W410
25. Strange RW, Antonyuk SV, Hough MA et al (2006) Variable metallation of human superoxide dismutase: atomic resolution crystal structures of Cu-Zn, Zn-Zn and as-isolated wild-type enzymes. J Mol Biol 356: 1152–1162
26. Hart PJ, Balbirnie MM, Ogihara NL et al (1999) A structure-based mechanism for copper-zinc superoxide dismutase. Biochemistry 38:2167–2178
27. Ko TP, Safo MK, Musayev FN et al (2000) Structure of human erythrocyte catalase. Acta Crystallogr D Biol Crystallogr 56:241–245
28. Epp O, Ladenstein R, Wendel A (1983) The refined structure of the selenoenzyme glutathione peroxidase at 0.2-nm resolution. Eur J Biochem 133:51–69
29. Karplus PA, Schulz GE (1989) Substrate binding and catalysis by glutathione reductase as derived from refined enzyme: substrate crystal structures at 2 A resolution. J Mol Biol 210:163–180
30. Pai EF, Schulz GE (1983) The catalytic mechanism of glutathione reductase as derived from x-ray diffraction analyses of reaction intermediates. J Biol Chem 258:1752–1757

Chapter 13

Computational Peptide Vaccinology

Johannes Söllner

Abstract

Immunoinformatics focuses on modeling immune responses for better understanding of the immune system and in many cases for proposing agents able to modify the immune system. The most classical of these agents are vaccines derived from living organisms such as smallpox or polio. More modern vaccines comprise recombinant proteins, protein domains, and in some cases peptides. Generating a vaccine from peptides however requires technologies and concepts very different from classical vaccinology. Immunoinformatics therefore provides the computational tools to propose peptides suitable for formulation into vaccines. This chapter introduces the essential biological concepts affecting design and efficacy of peptide vaccines and discusses current methods and workflows applied to design successful peptide vaccines using computers.

Key words Peptide, Vaccine, Epitope, B-cell, T-cell, Workflow, Prediction, Immunity

1 Introduction to Relevant Biology and Immunity

1.1 What Is Immunity?

The primary importance of an immune system is to differentiate self from non-self or dangerous from non-dangerous. Mostly self-tissues should be left alone as long as they are not malignantly transformed (cancers) and intruders should be attacked and removed to prevent their spread and utilization of body resources. In the case of pathogenic organisms such as worms, viruses, and bacteria co-evolution can generate various mechanisms of the pathogens to prevent successful immune reactions. Knowledge of these mechanisms can facilitate design of successful drugs such as vaccines. Innate immunity is capable of quickly recognizing and attacking intruders. However the full potential of the immune system is leveraged by extending the involved mechanisms by adaptive immunity, learning from previous immune reactions. Specifically two types of lymphocytes are critical for adaptive immunity, B-cells producing antibodies (the humoral immune response) and T-cells with their T-cell receptors for cellular immunity. Both are selected from a naïve repertoire to match non-self structures (shapes).

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_13,

Typically (but like everything in biology with exceptions) antibodies attack extracellular targets and are also important lines of defense in the mucus covering mucous membranes, while T-cells attack infected or mutated cells. The latter works because most cell types continuously present fragments (peptides) of all proteins found in their interior on protein complexes called MHCs (major histocompatibility complex) type I. These peptides are provided by the proteasome and result from the turnover of intracellular proteins present in the cytoplasm. Peptides are transported into the ER (endoplasmatic reticulum) where they are loaded onto MHC molecules, which are then presented on the cell surface. Importantly, MHCs are coded by (for vertebrate standards) highly polymorphic genes so specificity of these MHCs for their peptide ligands can be very different among individuals. This can lead to presentation of different peptides in different individuals, and hence also immune reactions of different intensity or against different portions of a pathogen (or cancer related) protein. Typically certain regional subpopulations of a species (or ethnicities in the case of humans) are enriched with certain MHC alleles, resulting in different vaccine efficacies in different parts of the world [1]. This variability is one of the major hurdles for peptide-based vaccines.

Length of MHC presented peptides can vary (in some cases also loops are possible), but typically ranges from 8 to 11 amino acids, depending on host species and MHC allele. Each MHC thereby has some and usually very significant tolerance in peptide binding patterns, but preferring certain amino acids in certain positions (sometimes called anchor positions if they contribute particularly to binding energy). Based on structure and peptide specificity currently at least 12 MHC class I supertypes are accepted, the members of which show overlapping peptide binding specificities [2]. In addition to class I MHCs also class II is important. These molecules are related, but structurally different and allow significantly longer peptides to be presented. Class II ligand peptides do not stimulate cytotoxic T-cells as class I ligands do, but rather a different class of T-cells called helper cells. These are in turn crucial to stimulate cytotoxic T-cells. The bottom line is, class II ligands are very important to obtain cellular immunity. However, also class II MHC alleles are fairly variable regarding their ligand specificity among individuals of a species. To generate T-cell based immune responses in an particular population of animals or human first the presence and specificities of extant MHC molecules needs to be determined or approximated.

While class I peptides (ligands for class I MHCs) normally have to be present in the cytoplasm of a cell to be presented on an MHC, a certain subclass of dendritic cells (a form of antigen-presenting cell) also presents fragments of proteins taken up from the environment to stimulate naïve T-cells, a mechanism referred to as cross-presentation [3]. This is important as it means also

peptides (or other antigens) contained in non-live vaccines and directed to these dendritic cells can lead to the development of immune reactions against intracellular targets (cell based immunity). It is the biological requirement for peptide based vaccines and strongly linked to the composition and delivery of non-life vaccines to target these cells.

1.2 What Is an Epitope?

The part of an immunogen recognized by either an antibody or a T-cell receptor is called an epitope, the complimentary part on the antibody (or T-cell receptor) the paratope. In the case of T-cell epitopes the receptor interacts not only with the antigen, but with the MHC/peptide complex. However the complexed peptide determines specificity of the reaction. In the case of B-cell epitopes continuous and discontinuous epitopes are differentiated. In continuous epitopes amino acids forming the contact interface to the antibody are directly linked to each other through peptide bonds, in other words a single peptide forms the epitope. In discontinuous epitopes the interface is formed by amino acids in close spacial contact but not directly linked through a continuous backbone, so not a single peptide. According to some estimates up to 90 % of all epitopes may be discontinuous [4]. Ultimately this is tricky to say as also discontinuous epitopes can have dominant continuous segments and the number of antibody/antigen structures resolved as 3D structures is fairly limited and generally restricted to regions of little flexibility, so by trend few flexible loops. However, current evidence suggests the discontinuous scenario is clearly more frequent.

Also, antibodies generated against peptides tend to be less affine than those generated against the native (complete) antigen. This may have several reasons: one may be that a peptide used for immunization has more degrees of freedom (is less restricted) and may therefore take on various shapes unlikely to be observed in the native antigen. Stimulated antibodies are therefore not as likely to match the native (more constrained) peptide well. Among the solutions to this practical problem two are of particular practical importance, (1) mimotopes and (2) circular peptides, both to be discussed later.

1.3 Antigenicity Versus Immunogenicity

Antigenicity describes the potential of a molecule (called an *antigen* in this context) to give rise to the development of specific antibodies complimentary (matching) the antigen's surface or alternatively the rise of T-cells equipped with T-cell receptors matching specific peptides/fragments derived from the antigen. Antigenicity at least for antibodies thereby is assumed to be largely species neutral. In the case of T-cell epitopes it mostly depends on abundance of peptides and MHC affinity (determining MHC/peptide complex stability) but also cleavage patterns of the proteasome (which is biased) and transport efficiency into the ER for MHC loading. In other words, at least for antibody epitopes it should be similar in a camel

and a whale, because it depends on basic physicochemical properties such as solvent accessibility of the epitope or potential as protein–protein binding interface. In the case of T-cell ligand determining specificity of MHC alleles on the other hand can be similar or different between MHC supertypes, between species and individuals. Also for antibodies antigenicity is not fully species neutral as differences in available antibody repertoire may exist (based on the genetic diversity of building blocks in the genetic process leading to antibody production) and antibodies can have different space requirements to access a particular patch of surface on a target antigen (e.g., a protein). As such B-cell antigenicity can be considered a form of biased protein–protein interaction potential. However, even existence of immature antibodies/T-cell receptors of good affinity does not necessarily lead to the productive generation of antibodies/T-cell populations as similarity to self and immunological history play a role. Substantially self-reactive antibodies and T-cells are normally removed from the repertoire to prevent self-damage. Therefore, antibodies specifically binding, for example, a patch of viral protein surface but also recognizing a human protein with substantial affinity will likely (and hopefully for the host) not be produced by mature B-cells, and likewise for T-cells and T-cell receptors. However, they may be generated in, for example, rabbits, if these lack the self-similarity to the viral protein. Likewise the history of the affected immune system is of importance. Immune systems typically retain the potential to quickly reactivate previous immune responses through memory cells. This is also a basis for vaccines, to protect against future encounters against the same or similar pathogens. Typically these memory populations are reactivated rather than generating new types of antibodies. This is normally good, but can also be problematic if the old population is not sufficiently efficient in removing the eliciting antigen. For example because the original pathogen/virus was too different from the currently matching one, an effect which can be the cause of severe disease syndromes such as dengue hemorrhagic fever upon sequential infection with multiple dengue variants [5]. Also, low-affinity antibodies (or low titer antibodies) can actually exacerbate infection by providing Fc-receptor mediated update routes for certain pathogens [6] or simply lack efficiency in clearing infection. This background and species dependent aspect is called immunogenicity and commonly is used also to comprise antigenicity, in other words the effect of generating an immune response. In cancer, immunogenicity may be most clearly differentiated from antigenicity because cancer tissues derive from the same genetic root as non-malignant tissues, and are as such highly related. Immunogenicity can be boosted and steered by providing additional danger signals, molecules typically found in invading pathogens (micro-organisms or viruses). These molecules are called adjuvants and are typical components of vaccines to reduce the

amount of required antigens (dose sparing) or possibly obtain more humoral or more cellular (or balanced) immune responses, as required. Also class II T-cell epitopes can be considered adjuvants as without these no efficient cytotoxic immune response will be mounted and possibly no memory cells develop. Certain pathogens as an evasion/survival strategy also misdirect immunity into an arm unsuitable to remove the pathogen (for example more into humoral or cellular immunity). This is where adjuvants may play a major role in future therapeutic vaccines (i.e. those not used for prevention but rather therapy of disease), to reprogram immunity.

1.4 What Is a Vaccine?

A vaccine is a pharmaceutical agent to stimulate the immune system of an animal or human to specifically recognize a similar but more dangerous agent. Typically the "similar" agent is a pathogenic organism such as a virus or bacteria, but may also be structures such as proteins or carbohydrates present in malignant cell types such as cancers. These are complimented by additional components to stimulate immune responses called "adjuvants" which control intensity and specific modes of the immune responses.

Scope of a vaccine is either preventive (administered to prevent infection or reduce severity) or therapeutic. Preventive vaccines are typically directed against non-chronic infectious diseases such as influenza, smallpox, or polio whereas therapeutic vaccines are directed against cancers or chronic infections. The latter are typically more complex to develop simply because the natural immune response does not clear the situation, so knowledge on the pathological mechanisms preventing clearance and likely suitable adjuvanting to redirect the immune response are required. Some vaccines may allow both applications, as would be beneficial for example in the case of EBV (Epstein–Barr virus) or malaria vaccines. In the case of cancer vaccines the "similar" agents to determine the specific immune response are typically proteins not present in adult tissues or present at low levels. However cancers may express these, possibly due to their qualities affecting growth, survival or metastasis. In some cases also truly novel recombinant (tumor specific) proteins exist [7]. Alternatives can also involve proteins of different expression levels than in typical host tissues, splice variants, minor mutations or altered posttranslational modification patters. As a result, vaccines against pathogens are often more straight-forward to design as the self-foreign differentiation aspect is not as complicated. Stimulating immune responses against self-structures is always associated with two risks (a) the vaccine may not work (a biological process called "immune tolerance" works to limit self-reactivity) or (b) self-structures of healthy tissues may also be attacked which may lead to autoimmune disease phenotypes. This aspect is also important if a pathogen mimics a host structure (or in some cases contains a homologous protein), an effect naturally driven by co-evolution as less immunogenic

offspring pathogens should succeed more likely. In any case, it is critical to compare similarity of vaccine peptides (or entire proteins) to the proteome of the host to be immunized to avoid similarities.

Currently peptide vaccines, subunit vaccines where all antigens are peptides, are on the fringe of the peptide industry. While of high potential they heavily depend on knowledge regarding the vaccine target and require suitable and modern delivery platforms to be both efficacious and safe [8].

2 B-Cell Epitope Prediction

Numerous methods for prediction of continuous and discontinuous peptides have been developed. From an application perspective and specifically for vaccines and diagnostics however primarily discontinuous epitopes and their predictions are of relevance. The simple reason is that a single peptide can represent a continuous epitope, while a discontinuous epitope would require more complex (non-standard) structures to be synthesized, with the potential exception of mimotopes (see later). From a vaccine perspective the goal is to predict a peptide which will stimulate production of antibodies which will in turn cross-react with the native (complete and structurally intact) protein the peptide intends to simulate. Also continuous epitopes synthesized as short peptides may take on structures different to that in the native protein and methods to approach the original situation can be advantageous to create better mimicry. One way is to include the peptide in a larger protein scaffold or, alternatively, circularize the peptide (for example by chemically linking N- and C-terminus) to stabilize curvature and decrease degrees of movement, thereby increasing potential for antibody affinity [9]. *See* Fig. 1 for an example of a continuous epitope.

Multiple parameters for successful prediction of continuous B-cell determinants (epitopes) have been proposed over the last 40 years, yet prediction efficiency is still often considered suboptimal. Multiple reasons for this deficiency can be identified, not the least being difference between antigenicity (which is normally predicted) and immunogenicity. The other is that they are biologically not as stringently defined as T-cell epitopes, essentially because the surface of a protein (or any other molecule) is continuous with varying degrees of potential protein–protein interaction, while epitopes are often understood as discrete units clearly differentiated from their neighbors. Yet still, certain features are critical and hence often used in B-cell epitope prediction.

2.1 Solvent Accessibility/ Hydrophilicity

By definition B-cell epitopes need to be located on the surface of an antigen. This feature is often approximated by accessibility to solvent molecules, although in fact antibodies are significantly more bulky. A number of methods exist which predict solvent

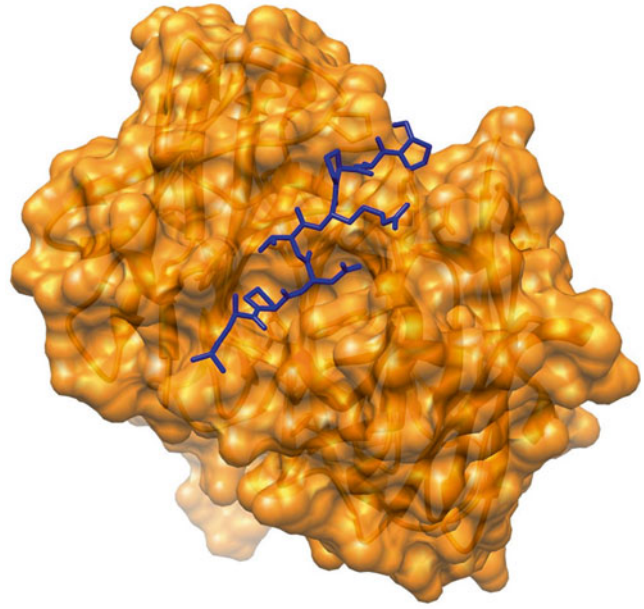


Fig. 1 Visualization of PDB entry 1SM3, complementarity-determining region (CDR) of breast cancer-specific antibody SM3 in complex with its peptide epitope, Mucin sequence Pro-Asp-Thr-Arg-Pro. This is an example for a complex of a short continuous epitope and its antibody. Image prepared with chimera

accessibility from sequence or assess it from available structural models. Good sources for experimentally determined structures are the RCSB PDB [10] and PDBe [11]. If no structures are available the Protein Model Portal [12] can be a source of theoretical models. If none are available, often homology modeling based on structure of closely related proteins or de-novo modeling can be used. Among the latter SWISS-MODEL [13] provides a particularly handy interface which can be run also without a major background in protein modeling. Off-line Modeller [14] or the Rosetta Protein modeling suite [15] are commonly used tools. Based on a suitable 3D model of a protein numerous tools exist to derive solvent accessibility data. DSSP [16] or the CCP4 package (programs SURFACE or AREAIMOL) [17] can be used to obtain accessible surface area, where AREAIMOL also shows relative solvent accessibility of amino acids, thus providing a directly usable metric for accessibility. If no structures are available and no suitable models can be derived, for example because no close homologs have been resolved by NMR or crystallography, sequence based methods for predicting solvent accessibility exist. Specifically Sann [18], NetSurfP [19], and Real-SPINE [20] are fairly recent and publicly accessible methods. PredictProtein [21] is a very easy to use meta-server for obtaining several predictions useful for selecting peptides in vaccinology. Aside of solvent accessibility and

secondary structure it will provide putative trans-membrane domains which are by definition little accessible to solvent. It also includes an overview of evolutionarily constrained (and therefore more conserved) areas and of protein–protein interaction sites which may be energetically favorable for antibody attachment. Also of critical importance to surface accessibility are post-translational protein modifications. Particularly glycosylations, which can contribute very substantially to mass to a protein, essentially masking the protein surface from antibodies as demonstrated in HIV gp160 [22]. Integration of existing knowledge such as from UNIPROT [23] and prediction of glycosylations [24–26] is therefore of critical importance, especially in vaccines against eukaryotic viruses and cancer.

2.2 Secondary Structure

Continuous B-cell epitopes, as linear surface located sections, tend to be rich in coils (surface loops) and are, by trend, low in secondary structures. This is at least true for B-cell epitopes which will stimulate antibodies cross-reactive with the native protein, and the others are of no use to vaccines. Particularly alpha helices should be avoided. Beta-turns have been identified as positive characteristics. However, on a general basis the rationale is to provide peptides structurally flexible in their native context simply because the intention is to raise antibodies able to recognize the native structure. As peptides derived from defined secondary structures (even if solvent accessible) may also assume quite different structures when unconstrained (in a vaccine context) it makes sense to focus on regions where also the native protein provides suitable flexibility, basically to adopt a structure complimentary to the antibody. The situation is different with mimotopes (as mimics of conformational structures), however peptide vaccines unfortunately commonly have to deal with this particular restriction. Secondary structures can either be derived from experimentally determined 3D structures (several of the methods named above include this option) or are predicted from sequence. The PSIPRED server [27] is commonly used for sequence based prediction because of the improvement from inclusion of homology information (conservation between homologs, favoring maintenance of structure). Regions of intrinsic disorder [28] may be of particular interest, providing sufficient degrees of freedom for B-cell epitope selection [9].

2.3 Other Amino-Acid Features

On the level of peptides mainly properties describing preference for solvent accessibility, ease of synthesis and formulation, peptide mobility, residue bulkiness, flexibility, and especially hydrophilicity are used. For these a number of single-residue amino-acid scales exist and have been compiled at ExPASy ProtScale [29]. A lot more have been collected by AAINDEX [30].

It is of high importance however to avoid maximization of these features at all cost. While hydrophilic and flexible peptides may be well accessible small and non-bulky residues also allow little

interaction surface (and thus little affinity), and also specificity of the interaction has to be retained. A serine/threonine/arginine polymer will likely show good antigenicity features but may fail to generate specific antibodies. One way to avoid this is to include sequence information entropy (Shannon entropy) into the evaluation, striving to include diverse residue types.

2.4 Word Frequencies

Several strategies for prediction of B-cell epitopes include preferences for amino-acids or amino-acid words derived from experimentally determined sets of epitopes versus non-epitopes. While the definition of clear-cut non-epitopes (aside of hydrophobic cores) is typically nontrivial, certain enrichments for amino-acids or words have been observed. These partially reflect general features such as hydrophilicity, but may intrinsically also reflect properties akin to required surface for biologically meaningful interactions, biases in antibody repertoire or preferences reflecting constraints in immunogenicity. This feature (based on single amino acid frequencies) was a dominant feature in the work of Kolaskar and Tongaonkar [31] still available in the program *antigenic* in the EMBOSS package and is also used as part of modern predictors. Also words consisting of not directly adjacent/bonded residues have been used [32].

2.5 Variability and Function

Sequence variability is a critical parameter for peptide vaccines (both, B-cell and T-cell based). While lack of highly affine/cross-reactive antibodies can be compensated to a point by choice of suitable adjuvants, altered antigens can abolish effectiveness of a vaccine altogether. This is particularly true for peptide vaccines, due to the restricted size of the immunogen in relationship to mutated area. On the other hand, some degree of variability can indicate evolutionary pressure on a particular epitope. Optimum are regions which are critical to function, so cannot be mutated easily without losing biologically critical functionality. Methods for predicting such functionality including motif based approaches, detection of evolutionarily constrained positions [33], prediction of protein–protein interaction surfaces, and particularly crystal structures showing amino acid contacts with critical receptors are very useful here. However, function of a vaccine target also raises a second aspect. Functionally constrained regions (epitopes/peptides) should be less variable, which is good for a vaccine as typically only few peptides can be included in a formulation and particularly regulatory reasons (depending on vaccine platform, however). But interfering with such functional sites can also be critical for stimulating effective immune responses. In many cases pathogens harbor highly immunogenic epitopes which do not lead to pathogen neutralization. These are called “decoy epitopes,” as they divert the bulk of the immune response to non-critical regions. These immune responses may still harm individual pathogens, for example by initiating a process called opsonization which leads to

antibody-mediated coating of pathogenic organisms. The pathogens (or some of them) may still be infectious, however. Therefore, particularly highly immunogenic proteins should be considered as candidates for decoys [34] and should at least not be the sole focus of a peptide vaccine strategy. If on the other hand critical regions for uptake into host cells or otherwise interaction with the host (such as pathogenicity mechanisms interfering with cell–cell communication) are known these can be interesting targets for a peptide vaccine. While peptide vaccines have substantial shortcomings regarding variability (because, compared to a full protein, each mutated amino acid can substantially affect similarity to the vaccine target) and require substantial technology for delivery, formulation and adjuvanting, they provide the means to target critical epitopes while ignoring undesired regions. Other sub-unit vaccines also include the potential to generate engineered proteins, but not at the same level of limitation to specific epitopes. Thus the strongest drawback of a peptide vaccine is also its biggest advantage: being able to hit precisely one spot, and just there. It cannot be diverted by decoy epitopes or functionally irrelevant regions if designed properly. Certainly this underscores the requirement for good knowledge of the target, including pathogenicity mechanisms on the protein–protein interaction level. Particularly viruses typically show a high degree of tissue tropism, meaning they require specific structures for entry into target cells. Blocking the viral receptors can abolish this process. Similarly bacteria deploy diverse attachment factors called adhesins. Especially for chronically infecting pathogens integration of knowledge and thorough selection and understanding of vaccine targets is a solid basis for effective vaccines. The potential to select (or alter) B-cell epitopes for the specific and effective elimination of host/pathogen interactions is a specific advantage of peptide vaccines over other vaccine approaches, and should therefore be focused on.

To this end selection of peptide B-cell epitopes for vaccines typically requires a trade-off between antigenicity maximization and variability minimization. Biologically/evolutionarily constrained surfaces important in host/pathogen interaction can form an excellent trade-off here while adding the additional benefit of (function) neutralizing antibodies. On this basis a method for prediction of protective epitopes has been published, but is unfortunately not publicly available [35]. Better understanding of the nature of evolutionary pressure on a particular protein region (peptide) in combination with functional constraints may also provide better insights in the optimal selection of certain peptide variants or even prediction of future variants, an aspect of high interest for vaccines [36].

2.6 Combined Methods and Resources

The Immune Epitope and Analysis Resource (IEDB) [37] provides numerous interesting offers for epitope prediction method developers (specifically datasets of high granularity) as well as prediction algorithms useful for peptide vaccine design (B- and T-cell

epitope based, including sequence coverage, population coverage, epitope cluster identification), including a structure-based viewer.

Methods for predicting discontinuous epitopes have been developed and can be used to extract continuous segments (peptides) from predicted discontinuous determinants [38, 39]. Publicly accessible methods for predicting continuous (linear) epitopes commonly are either based on feature plots, especially hydrophilicity scales such as that by Hopp [40] or combine several of the features listed above. The latter are based on machine-learning models such as decision trees, random forests, neural networks or support vector machines (SVMs) trained to discriminate features of known epitopes and non-epitopes based on various experimental datasets and amino-acid/peptide parameters. Almost a plethora of publicly available methods of this type exists [41–46]. The relative merits of continuous B-cell epitope prediction methods has often been questioned [47], which is at least partially due the lack of consistent method comparison and also different concepts and expectations of antigenicity and immunogenicity. On the other hand, differences between predicted and naturally occurring immune responses are not necessarily identical to failure of prediction. Failure of a vaccine to create antibodies cross-reactive with the target (protein) is, however. So far no organism-specific epitope predictor is known to the author; hence the gap between antigenicity and immunogenicity in predictions persists. BEEPro, one of the newest although so far not publicly accessible methods, has been reported to very significantly improve prediction accuracy by a combination of SVM, multiple-alignment-based assessment of antigen variability, and careful selection of the training and validation data sets [48].

2.7 *Mimotopes*

Mimotopes are peptides mimicking discontinuous segments of a molecule's surface. So far no computational technique exists to predict mimotopes, although structural knowledge of a protein's surfaces, electron density and its structure dynamics should make such predictions possible. Particularly so with the use of nonnaturally occurring amino acids and the application of retro-inverso technology for peptide synthesis. The latter may help to better mimic structure of spatially closely associated amino acids.

3 T-Cell Epitope Prediction

T-cell epitope prediction has long been considered more straightforward than that of B-cell epitopes and has, in addition, significantly improved over the last decade [49]. The reason is that the biology of peptide/MHC interaction constrains ligands (and thus potential epitopes) more clearly and in discrete units (peptides). As cytotoxic T-cell epitopes are critical for clearing intracellular infections and helper T-cell epitopes are important for boosting immunity and developing memory cells they are critical for many

vaccine strategies. Their clear drawback in comparison to application of B-cell epitopes in vaccines is that one B-cell epitope can potentially cover an entire host species (or multiple species) to be vaccinated, while MHC restriction in ligand presentation makes approaches aiming for T-cell approaches very dependent on knowledge of MHC types present in a given host population and data or algorithms on their peptide ligand specificity.

While prediction of B-cell epitopes is often not position specific (not directly considering the precise position of amino acids within an epitope), T-cell epitopes are fairly stringently defined by binding pockets in the MHC molecule, thus making amino-acid position specific classification the natural (and effective) choice. Additional improvements for prediction of cytotoxic epitopes are achieved by adding assessment of proteasomal processing of proteins into peptides and transport of these peptides into the ER compartment via TAP transporters [50, 51]. For the prediction of peptide affinity to class I and class II MHC quantitative matrix [52] and neural network [53] based prediction methods have been proposed, where generally both approaches have been reported to work well. Due to the structurally fairly conserved nature of MHC molecules (primarily swapping certain positions affecting ligand specificity) a method has been developed and implemented in the netMHCpan [54] server to extrapolate binding specificity of ligands for MHCs for which so far no experimental data is available. While the precise quality of these methods is sometimes doubted, they currently represent the by far fastest approach to obtain predictions for MHC molecules of unknown ligand specificity and have been successfully applied for vaccine development in pigs [55]. As MHC molecules are highly diverse among humans and other vertebrates coverage of this diversity is a crucial obstacle for peptide vaccines aiming at broad population coverage. One way to circumvent this limitation is the mentioned supertype concept, aiming to identify peptides sharing specificities with multiple MHCs (HLAs) within a supertype [56–58], or optimally even among supertypes. *See* Fig. 2 for an example of a peptide ligand in complex with a human class I MHC (HLA) molecule.

T-cell epitopes can also be predicted using molecular dynamics, an interesting strategy but requiring significantly more effort in setting up a simulation and very substantial computing power in comparison to sequence based approaches. The true advantage may lie in the potential to simulate also other aspects of immunity along with peptide ligand binding [59].

A biological phenomenon which can be utilized in the selection of peptides for stimulating T-cell responses is the identification of epitope clusters. These clusters are comprised of potential, overlapping or also adjacent MHC ligand peptides within a protein sequence. These regions may provide substantially better coverage of both, pathogen variability (especially in conserved hydrophobic regions)

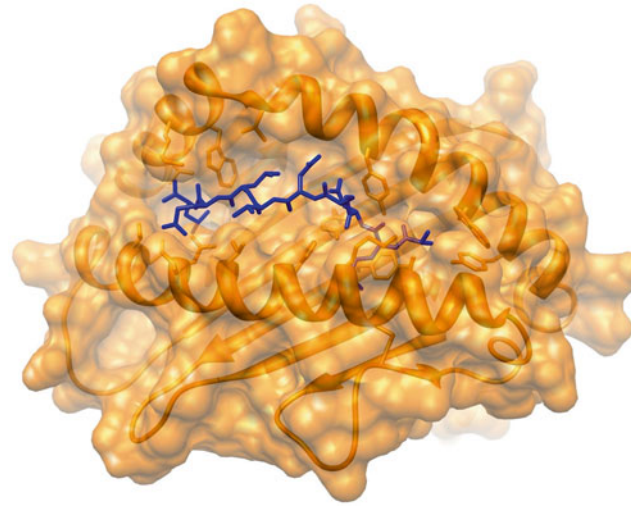


Fig. 2 Visualization of PDB entry 2X4S, a peptide derived from H5N1 (avian influenza) virus nucleoprotein in complex with a human MHC (HLA; human leukocyte antigen) molecule class I. The image demonstrates the interaction and embedding of a peptide of defined length in the surface groove of the MHC/HLA molecule and binding pockets formed by the MHC to allow affine interaction in this complex. Buried N- and C-termini of the peptide explain length restriction of MHC lass I presented peptides. The bound peptide/MHC complex is ready for recognition by a T-cell receptor. Image prepared with chimera

and MHC/supertype coverage [60]. A prerequisite for defining value as a vaccine antigen is suitable (possibly also alternative) proteolytic processing of the peptide into individual MHC ligands. Successful methods for prediction preferred proteasomal cut-sites exist [61], where for vaccine development it is important to choose a method which is based on the immunoproteasome [62] rather than the standard proteasome. Tools for combining these approaches namely supertype ligands, consideration of pathogen variability and epitope clusters, exist, for example Hotspot Hunter [63]. Recently numerous methods for proposing peptide sets for covering both population MHC population coverage and pathogen variability have been compared [64]. A simple algorithm named Conservancy Constrained T-cell Epitope Cluster (CCTEC) and some cluster visualisation are shown here [65]. Another option for limiting the number of peptides is to form synthetic concatenations of shorter peptides into longer peptides. The primary disadvantage of this approach (for both B- and T-cell targeting vaccines) is the formation of synthetic epitopes in the peptide junctions. For T-cell epitopes this can be minimized by arranging order of peptides so that proteasomal cleavage between them is optimized and minimized within the actually targeted epitopes. At the same time peptide order (possibly with constant spacers to optimize processing) is chosen to avoid novel T-cell epitopes in the junctions. For B-cell epitopes these

optimization steps are not as straightforward although at least in theory also possible by optimizing order to minimize antigenicity of the junctions. Excessive concatenation should probably be avoided unless junctional epitopes can be experimentally excluded.

4 Rational Vaccine Workflows

Rational vaccine workflows should result in the selection of vaccine antigens via integration of as much currently available knowledge (or at least data) as possible, filling up critical unknown aspects by predictions. Depending on the intended vaccine platform (recombinant protein, DNA, peptide formulations) a workflow will use different tools and result in different proposed antigens or constructs. Typically computational workflows focus on proteins, mostly because polysaccharides are bioinformatically less tractable. Reverse vaccinology [66] is the term coined by Rino Rappuoli for computationally filtering candidate proteins from the complete genome/proteome of a target pathogen. Reverse vaccinology thereby applies a number of rational steps to focus on characteristics of possible vaccine target proteins (such as trans-membrane proteins), resulting in a short or at least manageable list of candidates for experimental validation.

For peptide vaccinology these workflows are prerequisites and a sample workflow for Epstein Barr virus (EBV) comprising various methods as well as B- and T-cell targeting strategies has been published [65]. Based on suitable antigen peptides can be selected or engineered. Procedures involving selection of peptides from proteins (or engineering based on proteins) usually make the basic assumption that antibodies or T-cell responses (or both) against this particular antigen can lead to a protective immune response, possibly even sterilizing immunity capable of completely clearing the pathogen from the host. In-depth consideration of these workflows is unfortunately out of scope for this chapter as they touch much more than peptide associated technologies. However as they determine effectiveness of the vaccine we need to discuss some aspects here. It has to be particularly stressed that systems trained to predict relevant (peptide) epitopes may only be able to find best epitopes for a particular protein. If the entire protein does not have the potential for a neutralizing immune response all predictions will fail, hence the importance to see the entire workflow. The following questions should be considered and steps taken.

4.1 When Is the Antigen Expressed?

Possibly the most critical question is when an antigen is expressed. Many factors (including proteins) are only generated within certain periods of a pathogen's life cycle or in certain subpopulations of cancer cells or in specific patients (in the case of cancer). This is particularly true for chronically infecting organisms including

latent viruses, Mycoplasmas, Toxoplasma, Mycobacteria, or Plasmodia to name a few. For prevention of infection humoral immune responses are particularly suited, where targeting functional units such as adhesion molecules required for cell entry is particularly straightforward. For therapeutic infections other targets are equally or more relevant, mostly because the "early" gene products may not be produced anymore or spread works from cell to cell. For selection of antigens often literature mining (or manual curation/knowledge) are very useful, but again particularly functional considerations. In viruses typically "late" genes involved in replication and virion packaging or alternatively those for maintenance of latency can be defined. As such it is always important to define the scope of a vaccine: prevention versus therapy of disease. Integration of experimental data sources can provide the essential clues on expression status of specific antigens during the life cycle of a pathogen [67].

4.2 Which Type of Immunity Is Required for Vaccine Success?

First of all, rather B- or rather T-cell based? And particularly for B-cell antigens, functional regions (basically blocking interactions) or maximally immunogenic regions, where the latter is particularly suitable for antigens which are highly expressed and available at high density on a pathogen's surface, so to promote antibody mediated opsonization and generally complement-mediated immunity. B-cell epitopes will be the choice primarily for prevention and also for targeting secreted factors with known or hypothetical activity as pathogenicity factors (affecting disease phenotype and severity) or limiting pathogen spread. Class I T-cell epitopes (for stimulating cytotoxic T-cell responses) primarily for clearing existing infections on the cellular level. Both require integration of diverse data-sources, and optimally as much curated data and knowledge as possible, especially regarding the natural immune response against the pathogen (or cancer) and whether in which way it fails to resolve the infection. The latter especially interesting in the class of so-called difficult pathogens, those against which classical vaccine approaches have largely failed so far, including *M. tuberculosis*, HIV and Malaria parasites. Interesting tool for selecting the potential immunome include NERVE, Jenner-predict, Vaxign, and VaxiJen [68–72].

4.3 How Conserved Is the Antigen?

Aside of the degree of variability among known sequences, how many isolates of a particular pathogen do actually carry a homologue of the target? Commonly genes which can be lost easily by an organism may not be assumed to be highly conserved on a sequence level, but this is not an assumption to rely on. Loss of certain genes is perfectly possible and part of evolution, in some cases also backup systems exist. Experimentally this can be checked by PCR. Computationally the increasing wealth of completely sequenced genomes helps out, so presence of encoded proteins or

gene families can be checked by protein BLAST or comparative genomics. In some cases also paralogous families may exist and different isolates can feature different numbers of gene members, such as the VlhA adhesin cluster in Mycoplasma synoviae. If not many different variants of a particular pathogen have been sequenced, also related pathogens can be used. For example by means of a homology-based comparison of adhesion systems among related pathogens such as Brucella isolates which can be found in various host species. Also proteins shared among multiple species can be sufficiently similar on sequence level to allow selection of distinct peptides for vaccine development. However, similarity to other related (or unrelated) pathogens can also be detrimental because of reactivated memory cells producing (in this new context) low-affinity antibodies. Similarities among pathogens infecting the same host species should therefore be handled with care, as these similarities may also be an evasive strategy. The degree of similarity and existing functional constraints within the sequence may give further clues on the biological impact of sequence similarity among pathogens in organisms with diverse immunological histories.

4.4 Will There Be Self-Reactivity?

Selected pathogens antigens should not be similar to antigens of the host. This is conceptually difficult to determine for conformational B-cell epitopes, primarily because relatively few antigens of the pathogen may have been resolved on 3D structure level. As a first-line approach sequences (including variants) of the pathogen can be subjected to a BLAST [73] or other similarity search method against a comprehensive sequence database of the host organism on protein level. Potential similarities can be avoided on the peptide level, a specific advantage of peptide vaccines. While there are no strict rules, similarities larger than 6–8 amino acids should be avoided, where also similar amino acids may be counted into this stretch.

4.5 Is the Peptide Immunologically Available?

For B-cell epitopes the target peptide should be available to antibodies. This commonly means surface exposed (extracellular) domains of trans-membrane proteins or possibly soluble extracellular antigens if they comprise pathogen functionality to interfere with. Many methods for prediction of subcellular localization of proteins and trans-membrane domain structure exist. In certain cases also other proteins can be relevant targets. For example, Rotaviruses can be effectively targeted by secretory IgA antibodies because these are channeled through epithelial cells (including infected cells) in a process called transcytosis [74].

4.6 Minimize Number of Peptides

For practical reasons as well as regarding the approval process including a few peptides only is preferable. For T-cell epitopes supertype ligands and longer peptides including epitope clusters (class I and/or class II epitopes) can be helpful. Generally conserved and functionally constrained regions will provide peptide vaccines

with reduced chance for mutational escape of the pathogen. Another concept to optimize coverage is the selection of consensus sequences including multiple sequences covering subtypes. Optimal choice of variants has to the author's knowledge never been formalized so far. The aim thereby is to select distinct peptides (including consensus peptides or hypothetical peptides forming evolutionary intermediates between observed variants) which should cover functionally viable mutational intermediates with sufficiently intense stimulated immune reactions. While this is currently done by biological intuition and the use of classical substitution matrices such as the BLOSUM series, a defined metric for cross-reactivity (in B- and T-cells) would be highly appreciated.

4.7 Integrate as Much Knowledge as Possible

For rational selection of vaccine candidates a rational basis is required. This involves integration of literature with predictions and resources which already do that, like diverse NCBI and EBI resources, but also especially UNIPROT. Integration of curated data where possible is highly advisable as well as overlay of a number of different prediction methods, independent of what they predict. As strange as it may sound, immunoinformatics including peptide vaccinology is not about one perfect prediction system, but mostly about using whatever works to get an as good as possible understanding of host/pathogen (host/cancer) interaction to select the optimal points for disrupting this harmful interaction. Particularly for chronically infecting pathogens and there especially therapeutic (peptide) vaccines this understanding can make the critical difference between success or failure. Therapeutic vaccines are essentially ways to re-program the immune response into an effective mode, involving intelligent selection of antigens (including their peptides) and suitable adjuvanting to stimulate effective populations of immune cells. Systems vaccinology [75] is the exciting arm of system biology striving to obtain a systems view of immune reactions, including effects of adjuvants. Also host/pathogen interactions can be better understood by visualizing experimentally determined (or predicted) host/pathogen interactions. *See* Fig. 3 for an example of the Epstein–Barr virus/human interactome, derived from publicly available experimental as well as predicted protein–protein interaction data and differential expression of involved proteins. This approach can provide suitable insights into crucial contact points between host and pathogen at certain points of pathogenesis and propose suitable vaccine targets, and in some cases also drugs to support the vaccine and interfere with the host-pathogen interaction. Another way to make use of existing knowledge is to integrate immunogenicity information from related pathogens. For example, the group of Herpes viruses includes numerous pathogens of relevance among humans and animals, eight in humans alone. Herpes viruses can be distinguished in three major groups (alpha, beta, and gamma) where each is

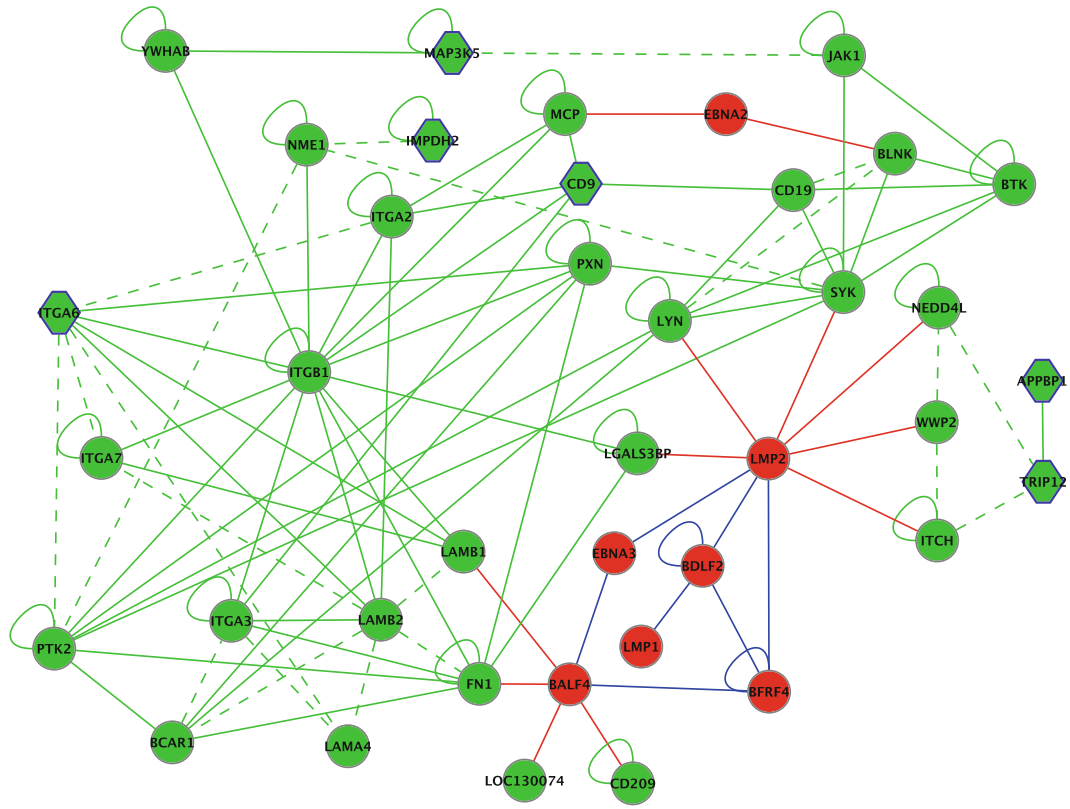


Fig. 3 Selection of vertices and edges from the EBV-human interaction graph centered around differentially regulated CD9. EBV proteins are shown in *red*, human proteins in *green*. *Solid lines* indicate physical interactions, *dashed lines* predicted connections. *Red*, *blue*, and *green* edges indicate EBV-EBV, EBV-human and human–human interactions, respectively. Human genes significantly differentially regulated upon infection/reactivation are shown as hexagon

characterized by specific proteins, tissue tropism (preferred target tissues/cells), and modes of pathogenesis. They also share numerous homologous proteins among groups. Some are too different on sequence level to be detected directly. However, when using each sequence of each Herpes virus as a seed for building a PSI-BLAST search matrix on an equilibrated sequence set such as UNIPROT90 (meaning, without certain sequences being over-represented) and then searching all other Herpes proteins one can establish a complete homology network. This procedure is also referred to as cascade PSI-BLAST [76]. The interesting aspect is that even with PSI-BLAST certain proteins cannot be linked directly, they are just too different. They can be linked via intermediate viruses however, which share sufficient similarity with both. Certain types of information can now be mapped through this homology network. Figure 4 shows peptide B-cell epitope (peptide) data mapped through this homology network to define potentially immunogenic regions on (based on sequence level)

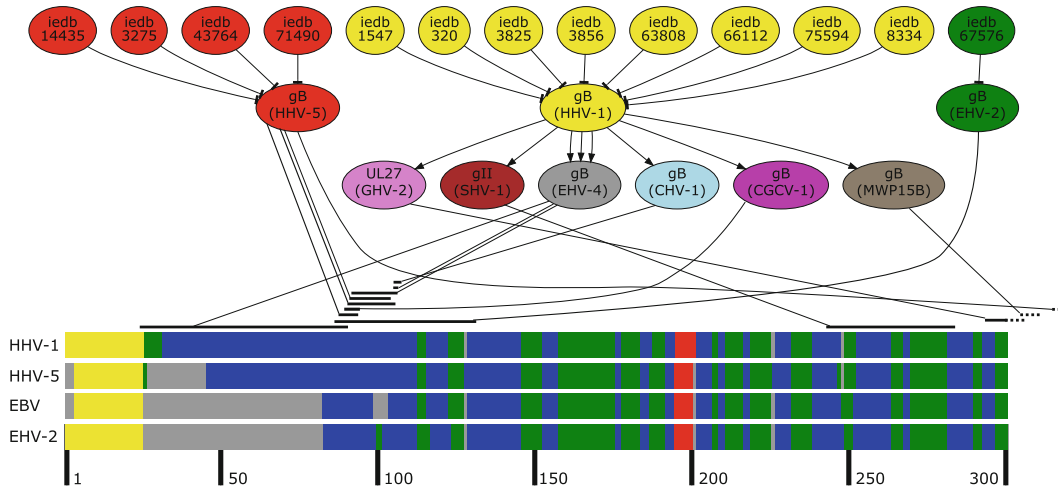


Fig. 4 The epitope mapping process. The *upper part* depicts a subgraph comprising shortest homology paths between proteins of diverse Herpes viruses and EBV gp110. Along these edges peptide positions are mapped utilizing PSI-BLAST based alignments. The *lower part* of the figure shows the first 300 columns of a multiple secondary structure alignment of homologous envelope glycoproteins of EBV, HHV-5, HHV-1, and EHV-2. HHV-1 epitopes are mapped to EBV via a multitude of intermediate viruses, other viral epitopes can be mapped directly. To improve readability secondary structures are color coded (helical areas in *red*, beta sheets in *green*, coils in *blue*, signal peptides in *yellow*, and gaps in *grey*). The *black lines* above the multiple alignment mark possible peptide/immunogenicity mapping positions. Figures 3 and 4 have been reproduced from ref. 65, originally published by BioMed Central

very different pathogens. The source peptide and the one mapped to are, in many cases, hardly related. However, from the original pathogen we know the region is immunogenic, and secondary structure comparison of proteins indicates these may also be structurally very similar. While there is no definite guarantee the overall 3D topology/accessibility of the domain may not have changed, the approach allows to enrich information from among fairly distinct organisms, providing indicators on peptide selection for vaccine applications. As typically certain organisms among a group are better analyzed and understood than others this type of immunogenicity mapping can provide very useful insights regarding location of potential peptide epitopes.

On a general basis, computationally designed peptide vaccines are knowledge based vaccines. Their primary prerequisite is therefore the integration of diverse knowledge sources regarding protein structure and function as well as application of predictors derived from biological data to select as few peptides as possible to get maximally robust and suitably intense immune reactions to prevent or clear disease. To be of practical use computational peptide vaccinology projects always have to interact with the targeted vaccine platform, its strengths and limitations, and regulatory requirements.

References

1. Gonzalez-Galarza FF, Christmas S, Middleton D et al (2011) Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. Nucleic Acids Res 39:D913–D919
2. Thomsen M, Lundegaard C, Buus S et al (2013) MHCcluster, a method for functional clustering of MHC molecules. Immunogenetics 65:655–665
3. Hunzeker JT, Elftman MD, Mellinger JC et al (2011) A marked reduction in priming of cytotoxic CD8+ T cells mediated by stress-induced glucocorticoids involves multiple deficiencies in cross-presentation by dendritic cells. J Immunol (Baltimore, MD : 1950) 186:183–194
4. Tomar N, De RK (2010) Immunoinformatics: an integrated scenario. Immunology 131:153–168
5. Dowd KA, Pierson TC (2011) Antibody-mediated neutralization of flaviviruses: a reductionist view. Virology 411:306–315
6. Meyer K, Banerjee A, Frey SE et al (2011) A weak neutralizing antibody response to hepatitis C virus envelope glycoprotein enhances virus infection. PLoS One 6:e23699
7. Schietinger A, Philip M, Schreiber H (2008) Specificity in cancer immunotherapy. Semin Immunol 20:276–285
8. Black M, Trent A, Tirrell M et al (2010) Advances in the design and delivery of peptide subunit vaccines with a focus on Toll-like receptor agonists. Expert Rev Vaccines 9:157–173
9. Caoili SEC (2012) On the meaning of affinity limits in B-cell epitope prediction for antipeptide antibody-mediated immunity. Adv Bioinformatics 2012:346765
10. Rose PW, Bi C, Bluhm WF et al (2013) The RCSB protein data bank: new resources for research and education. Nucleic Acids Res 41:D475–D482
11. Gutmanas A, Alhroub Y, Battle GM et al (2014) PDBe: protein data bank in Europe. Nucleic Acids Res 42:D285–D291
12. Haas J, Roth S, Arnold K et al (2013) The protein model portal: a comprehensive resource for protein structure and model information. Database 2013:bat031
13. Biasini M, Bienert S, Waterhouse A et al (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252–W258
14. Webb B, Sali A (2014) Protein structure modeling with MODELLER. Methods Mol Biol (Clifton, NJ) 1137:1–15
15. Kaufmann KW, Lemmon GH, Deluca SL et al (2010) Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry 49:2987–2998
16. Kabsch W, Sander C, Scharff M et al. DSSP. http://www.cmbi.ru.nl/hsspsoap/
17. Winn MD, Ballard CC, Cowtan KD et al (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr 67:235–242
18. Joo K, Lee SJ, Lee J (2012) Sann: solvent accessibility prediction of proteins by nearest neighbor method. Proteins 80:1791–1797
19. Petersen B, Petersen TN, Andersen P et al (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 9:51
20. Dor O, Zhou Y (2007) Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins 68:76–81
21. Yachdav G, Kloppmann E, Kajan L et al (2014) PredictProtein: an open resource for online prediction of protein structural and functional features. Nucleic Acids Res 42:W337–W343
22. Go EP, Irungu J, Zhang Y et al (2008) Glycosylation site-specific analysis of HIV envelope proteins (JR-FL and CON-S) reveals major differences in glycosylation site occupancy, glycoform profiles, and antigenic epitopes' accessibility. J Proteome Res 7:1660–1674
23. UniProt Consortium (2014) Activities at the universal protein resource (UniProt). Nucleic Acids Res 42:D191–D198
24. Xu Y, Wang X, Wang Y et al (2014) Prediction of posttranslational modification sites from amino acid sequences with kernel methods. J Theor Biol 344:78–87
25. Steentoft C, Vakhrushev SY, Joshi HJ et al (2013) Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J 32:1478–1488
26. Lam PVN, Goldman R, Karagiannis K et al (2013) Structure-based comparative analysis and prediction of N-linked glycosylation sites in evolutionarily distant eukaryotes. Genomics Proteomics Bioinformatics 11:96–104
27. Buchan DWA, Minneci F, Nugent TCO et al (2013) Scalable web services for the PSIPRED protein analysis workbench. Nucleic Acids Res 41:W349–W357
28. Pryor EE Jr, Wiener MC (2014) A critical evaluation of in silico methods for detection of membrane protein intrinsic disorder. Biophys J 106:1638–1649
29. Wilkins MR, Gasteiger E, Bairoch A et al (1999) Protein identification and analysis tools

in the ExPASy server. Methods Mol Biol (Clifton, NJ) 112:531–552
30. Kawashima S, Pokarowski P, Pokarowska M et al (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202–D205
31. Kolaskar AS, Tongaonkar PC (1990) A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett 276:172–174
32. Söllner J, Mayer B (2006) Machine learning approaches for prediction of linear B-cell epitopes on proteins. J Mol Recognit 19: 200–208
33. Ashkenazy H, Erez E, Martz E et al (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38:W529–W533
34. Gillet L, May JS, Colaco S et al (2007) The murine gammaherpesvirus-68 gp150 acts as an immunogenic decoy to limit virion neutralization. PLoS One 2:e705
35. Sollner J, Grohmann R, Rapberger R et al (2008) Analysis and prediction of protective continuous B-cell epitopes on pathogen proteins. Immunome Res 4:1
36. Ojosnegros S, Beerenwinkel N (2010) Models of RNA virus evolution and their roles in vaccine design. Immunome Res 6(Suppl 2):S5
37. Kim Y, Ponomarenko J, Zhu Z et al (2012) Immune epitope database analysis resource. Nucleic Acids Res 40:W525–W530
38. Haste Andersen P, Nielsen M, Lund O (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci 15:2558–2567
39. Ponomarenko J, Bui H-H, Li W et al (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics 9:514
40. Hopp TP (1993) Retrospective: 12 years of antigenic determinant predictions, and more. Pept Res 6:183–190
41. Yao B, Zhang L, Liang S et al (2012) SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. PLoS One 7:e45152
42. Davydov II, Tonevitskiĭ AG (2009) Linear B-cell epitope prediction. Mol Biol 43: 166–174
43. Gao J, Faraggi E, Zhou Y et al (2012) BEST: improved prediction of B-cell epitopes from antigen sequences. PLoS One 7:e40104
44. Costa JG, Faccendini PL, Sferco SJ et al (2013) Evaluation and comparison of the ability of online available prediction programs to predict true linear B-cell epitopes. Protein Pept Lett 20:724–730
45. El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting linear B-cell epitopes using string kernels. J Mol Recognit 21:243–255
46. Saha S, Raghava GPS (2007) Prediction methods for B-cell epitopes. Methods Mol Biol (Clifton, NJ) 409:387–394
47. Bergmann-Leitner ES, Chaudhury S, Steers NJ et al (2013) Computational and experimental validation of B and T-cell epitopes of the in vivo immune response to a novel malarial antigen. PLoS One 8:e71610
48. Lin SY-H, Cheng C-W, Su EC-Y (2013) Prediction of B-cell epitopes using evolutionary information and propensity scales. BMC Bioinformatics 14(Suppl 2):S10
49. Lundegaard C, Lund O, Buus S et al (2010) Major histocompatibility complex class I binding predictions as a tool in epitope discovery. Immunology 130:309–318
50. Doytchinova IA, Guan P, Flower DR (2006) EpiJen: a server for multistep T cell epitope prediction. BMC Bioinformatics 7:131
51. Larsen MV, Lundegaard C, Lamberth K et al (2005) An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. Eur J Immunol 35:2295–2303
52. Ivanov S, Dimitrov I, Doytchinova I (2013) Quantitative prediction of peptide binding to HLA-DP1 protein. IEEE/ACM Trans Comput Biol Bioinform 10:811–815
53. Lundegaard C, Lund O, Nielsen M (2011) Prediction of epitopes using neural network based methods. J Immunol Methods 374:26–34
54. Hoof I, Peters B, Sidney J et al (2009) NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics 61:1–13
55. Liao Y-C, Lin H-H, Lin C-H et al (2013) Identification of cytotoxic T lymphocyte epitopes on swine viruses: multi-epitope design for universal T cell vaccine. PLoS One 8:e84443
56. Larsen MV, Lundegaard C, Lamberth K et al (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinformatics 8:424
57. Zhang GL, Bozic I, Kwoh CK et al (2007) Prediction of supertype-specific HLA class I binding peptides using support vector machines. J Immunol Methods 320:143–154
58. Wang S, Bai Z, Han J et al (2014) Improving the prediction of HLA class I-binding peptides using a supertype-based method. J Immunol Methods 405:109–120
59. Flower DR, Phadwal K, Macdonald IK et al (2010) T-cell epitope prediction and immune complex simulation using molecular dynamics: state of the art and persisting challenges. Immunome Res 6(Suppl 2):S4

60. Chakraborty S, Rahman T, Chakravorty R et al (2013) HLA supertypes contribute in HIV type 1 cytotoxic T lymphocyte epitope clustering in Nef and Gag proteins. AIDS Res Hum Retroviruses 29:270–278
61. Xie J, Xu Z, Zhou S et al (2013) The VHSE-based prediction of proteasomal cleavage sites. PLoS One 8:e74506
62. Ferrington DA, Gregerson DS (2012) Immunoproteasomes: structure, function, and antigen presentation. Prog Mol Biol Transl Sci 109:75–112
63. Zhang GL, Khan AM, Srinivasan KN et al (2008) Hotspot Hunter: a computational system for large-scale screening and selection of candidate immunological hotspots in pathogen proteomes. BMC Bioinformatics 9(Suppl 1):S19
64. Schubert B, Lund O, Nielsen M (2013) Evaluation of peptide selection approaches for epitope-based vaccine design. Tissue Antigens 82:243–251
65. Söllner J, Heinzel A, Summer G et al (2010) Concept and application of a computational vaccinology workflow. Immunome Res 6(Suppl 2):S7
66. Donati C, Rappuoli R (2013) Reverse vaccinology in the 21st century: improvements over the original design. Ann N Y Acad Sci 1285: 115–132
67. Etz H, Minh DB, Henics T et al (2002) Identification of in vivo expressed vaccine candidate antigens from Staphylococcus aureus. Proc Natl Acad Sci U S A 99:6573–6578
68. Flower DR, Macdonald IK, Ramakrishnan K et al (2010) Computer aided selection of candidate vaccine antigens. Immunome Res 6(Suppl 2):S1
69. Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinformatics 8:4
70. Jaiswal V, Chanumolu SK, Gupta A et al (2013) Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. BMC Bioinformatics 14:211
71. He Y, Racz R, Sayers S et al (2014) Updates on the web-based VIOLIN vaccine database and analysis system. Nucleic Acids Res 42:D1124–D1132
72. Vivona S, Bernante F, Filippini F (2006) NERVE: new enhanced reverse vaccinology environment. BMC Biotechnol 6:35
73. Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
74. Aiyegbo MS, Sapparapu G, Spiller BW et al (2013) Human rotavirus VP6-specific antibodies mediate intracellular neutralization by binding to a quaternary structure in the transcriptional pore. PLoS One 8:e61101
75. Nakaya HI, Li S, Pulendran B (2012) Systems vaccinology: learning to compute the behavior of vaccine induced immunity. Wiley Interdiscip Rev Syst Biol Med 4:193–205
76. Kaushik S, Mutt E, Chellappan A et al (2013) Improved detection of remote homologues using cascade PSI-BLAST: influence of neighbouring protein families on sequence coverage. PLoS One 8:e56449

Chapter 14

Computational Modeling of Peptide–Aptamer Binding

Kristen L. Rhinehardt, Ram V. Mohan, and Goundla Srinivas

Abstract

Evolution is the progressive process that holds each living creature in its grasp. From strands of DNA evolution shapes life with response to our ever-changing environment and time. It is the continued study of this most primitive process that has led to the advancement of modern biology. The success and failure in the reading, processing, replication, and expression of genetic code and its resulting biomolecules keep the delicate balance of life. Investigations into these fundamental processes continue to make headlines as science continues to explore smaller scale interactions with increasing complexity. New applications and advanced understanding of DNA, RNA, peptides, and proteins are pushing technology and science forward and together. Today the addition of computers and advances in science has led to the fields of computational biology and chemistry. Through these computational advances it is now possible not only to quantify the end results but also visualize, analyze, and fully understand mechanisms by gaining deeper insights. The biomolecular motion that exists governing the physical and chemical phenomena can now be analyzed with the advent of computational modeling. Ever-increasing computational power combined with efficient algorithms and components are further expanding the fidelity and scope of such modeling and simulations. This chapter discusses computational methods that apply biological processes, in particular computational modeling of peptide–aptamer binding.

Key words Molecular dynamics, Aptamers, Peptides, Docking, Computational peptidology

1 Introduction

Short sequences of amino acids and peptides have many applications that include self-assembly, cell signaling, nutritional enhancement, and biomarker research. The application of peptides as biomarkers is of particular interest due to their use in disease detection. Biomarkers are molecules that correspond with biochemical changes in the body. Changes in concentration, physiology, and morphology are indicators that allow tracking of disease progression and drug effectiveness in the body [1]. For some diseases the biomarker is an individual peptide, but peptide sequences can be harvested from within protein biomarkers. As proteins are polypeptides, one can look at peptide segments of proteins that are important in the binding of the system. Experimentally targeting the peptides gives

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7_14,

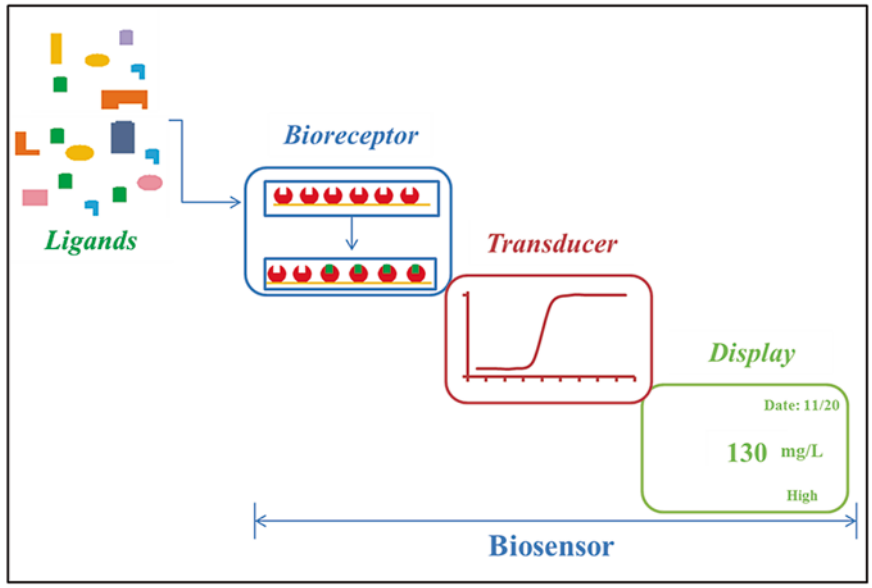


Fig. 1 Schematic representation of biosensor working principle

specific insights into the binding as well as provides a cost-effective alternative to using entire proteins. The concentration of these molecules coincides with the severity of the disease. This corollary has opened a window of opportunity in the diagnosis of many diseases. However, discovering a suitable biomarker alone is not sufficient to create a viable diagnostic platform. Selecting a bio-receptor for the biomarker is essential as it specifically recognizes the biomarker among millions of other molecules. Bio-receptor–biomarker combinations act as a lock and key mechanism to create a biological complex which can be interpreted by a biosensor. A biosensor is a receptor–transducer device that provides quantitative information using a bio-recognition element/bio-receptor and a transducer [2]. The transducer is generally based on electrochemical, mass, optical, or thermal properties while the bio-recognition element or bio-receptor action involves biochemical mechanism [2, 3]. When a biological sample is loaded into the sensor, the bio-recognition element/bio-receptor recognizes the target in the sample and binds to it. The transducer registers the change which is quantified and displayed on the device (*see* Fig. 1).

The potential of diagnostic devices for various diseases are hinged on the ability to gather the appropriate bio-receptors [2]. The most common method of making biosensors involves using antibodies as the bio-receptor [2]. Antibodies accompany the biological response to disease and injury which facilitates their use in biosensors. Glucose meters and pregnancy tests are well-known examples of biosensors because of their ease and immediacy of use.

A comprehensive sensor for many diseases would require a multi biomarker platform. Antibody based biosensors are common; however, there are distinct disadvantages of using antibodies in a multi biomarker biosensor. Antibodies are large molecules that are not readily synthesized and can be chemically unstable [4, 5]. Instability can cause errors and inaccuracies in readings of the biosensor. Their relatively larger size limits the number of antibodies that can be placed on the surface of the biosensor. Not only are antibodies large, but they are often good for a single usage in a biosensor [5]. These challenges motivated the investigation for better bio-recognition elements. Aptamers are one such bio-recognition element [6].

2 Aptamer Selection

Aptamers are broadly classified as either nucleic or peptide aptamers. Nucleic aptamers are synthetic oligonucleotides sequences made of single-stranded DNA or RNA [7]. Peptide aptamers are combinatorial protein molecules consisting of a variable peptide sequence inserted within a constant scaffold protein [8, 9]. Aptamers are advantageous as bio-receptors since they are relatively small, chemically stable and have a high binding affinity [4]. The aptamers have similar or better binding affinity compared to antibodies [7]. Such binding affinity is due to the aptamers' ability not only to bind to a specific structure but also to adapt conformation that favors the binding.

Nucleic and peptide aptamer types have distinct advantages as bio-recognition elements. Peptide aptamers have added chemical diversity compared to nucleic acid aptamers as binding does not occur on the sequence level [10]. Both RNA and DNA aptamers are reusable. However, RNA aptamers are susceptible to ribonuclease degradation, limiting their reusability [2, 4]. Due to small size, unlike antibodies, it is possible to affix large number of aptamers in a single location, creating a high-density receptor area. Aptamers can also be easily functionalized and immobilized to surfaces to create highly ordered receptor layers [2].

Over the years, a compilation of oligonucleotides and peptides have been made into aptamer libraries. A standard nucleic 25-mer library compilation currently stands at 10^{15} available aptamers [11]. In solution, these aptamers are quite flexible and adopts a tertiary conformation that complements the target molecule [7]. In 1990, a reasonable experimental solution for nucleic aptamer selection was provided by the Systematic Evolution of Ligands by Exponential Enrichment (SELEX) process, in which developed libraries undergo incubation with the desired target molecule [11]. Aptamers that do not bind to the target are removed and bound aptamers are separated from the target and amplified using

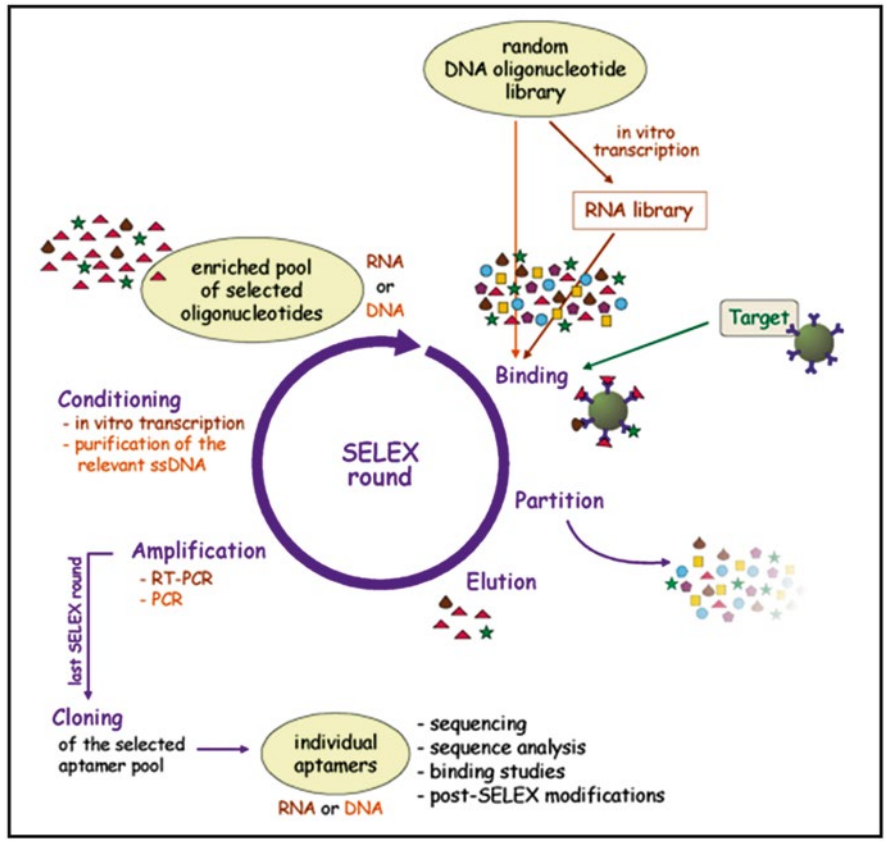


Fig. 2 Schematic diagram of the SELEX process (reproduced from ref. [11] with permission)

polymerase chain reactions (PCR) [11]. In the PCR process primers are added to the aptamers and are replicated making many double-stranded copies. These double strands are then separated, transcribed, and purified into single-stranded DNA (ssDNA) [11]. This pool of aptamers goes through several more rounds of SELEX until the pool is reduced to a handful of sequences (*see* Fig. 2). Target features, concentration, design of the initial library, experimental environment, and specificity of the binding are all determinants for the number of SELEX rounds that need to be done [11]. After the SELEX process the pool of aptamers must be sequenced for identification. The resulting SELEX aptamers should contain a select group that has the highest binding affinity for the target molecule.

Several methods have evolved for the selection of peptide aptamers. One method is using phage display. In this process peptide libraries constrained in loops of capsid protein are presented to a filamentous bacteriophage [8, 10, 12]. The gene needed for the capsid protein is contained within the phage effectively isolating

the target and its aptamer. A second approach, also named "two-hybrid" approach, is to select peptide aptamers that bind to their targets within the cytoplasm of living cells [10]. In this process, a protein is used as a scaffold for the display of a random library. A transcription factor is attached to the target protein within a cell containing a marker that is dependent on the expression of the transcription factor. Peptide aptamers are selected based on the inhibition of the selected protein. Alternative methods have been introduced using ribosomes and mRNA displays [10]. These methods still follow the same protocols but utilize DNA and mRNA libraries. Though all of these methods are constrained by the size of the library, they are still effective in generating peptide aptamers.

Once identified, these aptamers must go through optimization and validation experiments. Aptamer binding can be optimized by examining slight variations in the sequenced aptamer and varying the solvent environments. Variations in solvents and ion concentrations are known to influence the binding event [13]. Consideration must also be given to the target molecule source. For example, if the target is introduced from a blood or urine sample, one needs to make sure that the aptamer is also viable in the associated environment. Aptamer binding validation is generally done using ELISA, microarrays, and surface plasmon resonance imaging (SPRi) [14]. These methods allow one to find binding affinity, association and dissociation constants which determine the strength of the binding combination.

3 Experimental Analysis of Peptide–Aptamer Binding: Challenges and Limitations

Despite several experimental studies for aptamer selection and binding, there are still challenges and limitations. Since its introduction, the SELEX process has evolved; however, for this process to be successful, sequencing of the aptamers is still required. Due to the massive size of the aptamer library, SELEX must be done in small batches and there are risks of damaging the aptamers during the process. It is worth mentioning that during the PCR process of SELEX, the aptamers are amplified with the addition of primers and extension regions. Although primers are later removed, any residual nucleotides would alter the sequence. This addition could also cause a change in the binding characteristics or location. Though this process can select a group of aptamers over time, a major drawback is its inability to identify the specific binding site or natural progression of binding. Experimentally, Nuclear Magnetic Resonance (NMR) and X-Ray crystallography are considered to be the current best tools to obtain molecular structures. These techniques provide a snapshot of stable molecular structures in a solution, but are unable to provide insights into the natural progression of binding.

On the other hand, validation methods such as SPRi and microarrays have their own shortcomings. Both methods are efficient in determining binding, but they do not provide structural details of binding events. Many larger aptamers, peptides, and proteins will have multiple binding sites. SPRi and microarray results give authentication to the formation of a binding complex but do not reveal explicitly where the binding occurs. Similar shortcomings are associated with the methods used in binding of the aptamer targets to a surface. It is important to note that while dealing with small molecules like peptides and aptamers, even minute changes can impact the binding. For example, surface chemical methods applied to aptamers and peptides can cause changes in the molecular structure or interrupt possible binding sites.

The understanding of the natural progression of binding that is a currently a limitation in experimental studies can be further enhanced through computational modeling. A better understanding of biomarker or bio-receptor interactions can be obtained by developing computational models based upon their associated molecular systems. Computational modeling can facilitate the selection of target molecules for any biomarker. Such modeling can also aid the SELEX process as it enables one to analyze and understand the progression of the binding process, and not just the end outcome. Computational peptide–aptamer binding experiments can help identify binding sites and structural motifs obtained under various conditions.

4 Computational Modeling

The power of today's computational modeling can be an avenue to test, analyze, and visualize the peptide–aptamer binding that forms the basis of the aptamer selection process. The size of a nucleic acid aptamer library depends on the length of the variable region and can be approximated as

$$\text{library size} = 4^n \tag{1}$$

where n is the length of the variable region in the aptamer [2]. Going through each sequence of such a library is nearly impossible using the current wet lab procedures. As explained before, using aptamers as bio-recognition elements offers high selectivity and specificity. Reducing the multi-trillion aptamer pool through computational modeling and analysis can potentially cut down the time and resources needed for optimal binding aptamer selection. It is now known that the open regions of the aptamer's 3D structure provide the binding sites for peptides. There can be several of these sites, but it is unclear which site is used, and whether the site changes under varying conditions. Using aptamers in a biosensor

has the additional challenge of identifying the optimum orientation that favors binding. Generally, aptamers are bound to a surface within the device. Surface chemistry and target orientation influence the effectiveness of biosensors. Modeling and understanding how binding occurs will aid in the device development providing insights into binding specificities.

Computational modeling includes the effect of the associated processes and physical phenomena and can provide an emulation of the real behavior through relevant mathematical and computational formulations. In the case of molecular structures involving peptide–aptamer interactions, computational simulations can provide insight into the progression of the binding process. The efficacy, applicability, and insight from computational analysis of peptide–aptamer systems depend upon the fidelity of associated molecular models.

For the computational analysis of peptide and aptamer binding, docking and transient dynamic simulations of the relevant molecular systems are general approaches that can be employed. Simulation modeling and docking cannot be performed without a structure. Having a structure allows one to explore anomalies, mutations, and destruction of molecules and evaluate how these could lead to changes in molecular function and phenotype. These molecular structures are typically generated from NMR or X-Ray crystallographic solutions of the associated molecular systems. If the structures are not available, structure prediction analysis can help predict the structure [15]. Structure prediction methods start from the primary structures and compute secondary or tertiary structures based on other closely related known structures or de novo physics [15]. It is important to examine the molecular makeup prior to simulation analysis to identify key features and areas of interest. For peptides (short amino acid chains), one need to look at the sequence of residues as well as their length. Amino acids can be characterized into subgroups defined by their residues (Fig. 3). Each residue serves a specific function in peptide and protein structure. Such features are important as they determine their role in binding. For example, proline tends to make kinks in peptide chains due to its formation of a nitrogen ring on a peptide backbone [15]. Therefore, a proline heavy peptide tends to be more rigid. Identification of such peptide characteristics aids in the simulation analysis and design. For aptamers, one must consider the tertiary structure, as well as the open regions that may act as potential areas for ligand interactions.

The function of active biological molecules leads to the understanding of biological pathways and mechanisms [16]. The ability to readily change molecules and obtain a corresponding binding affinity of each combination can provide a preemptive look into the efficiency and efficacy of a binding combination. Testing different ligands with specific target molecules can be performed with

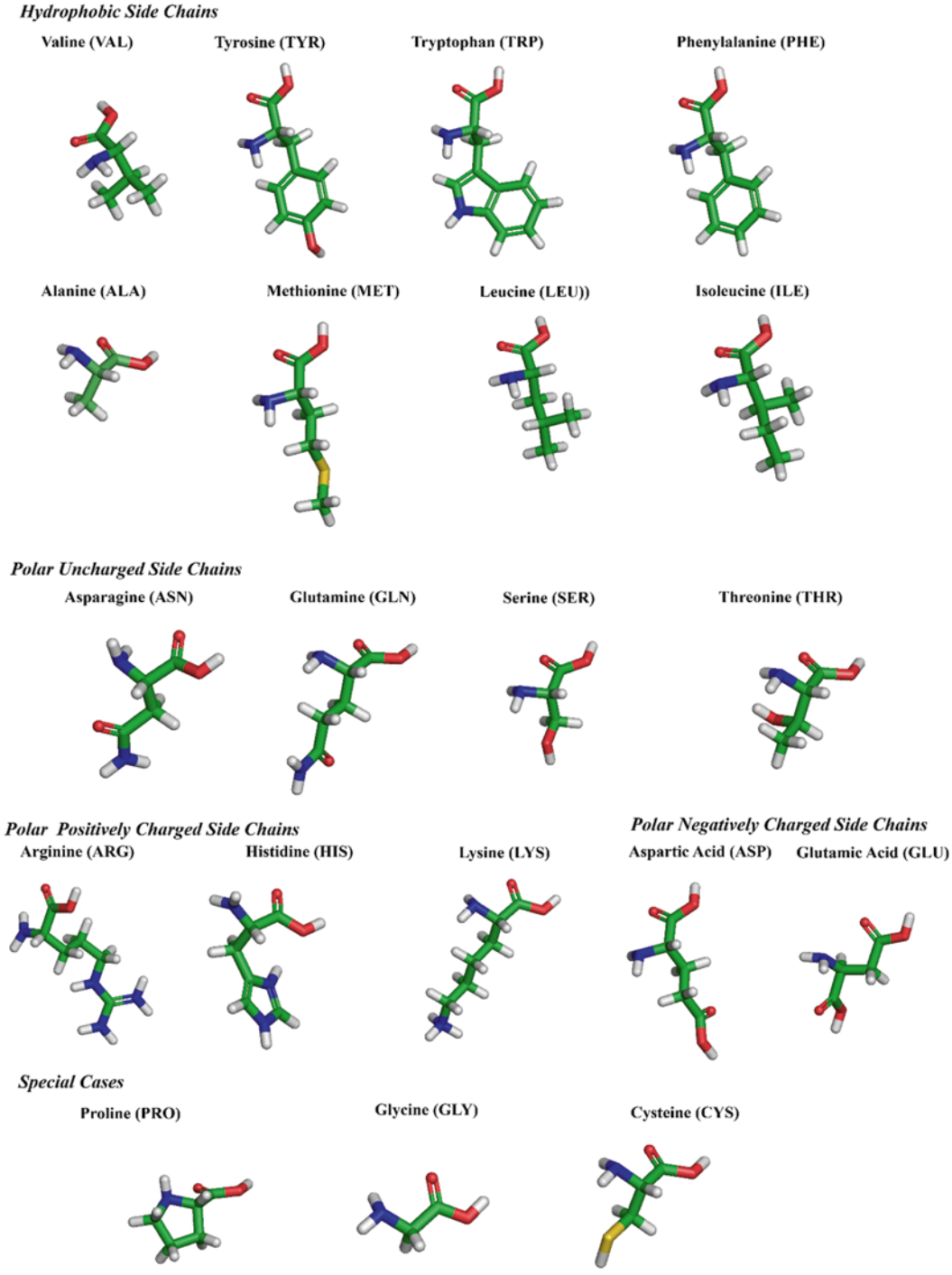


Fig. 3 Chemical structures of amino acids shown in the stick representation. Color code: Carbon—*green*, oxygen—*red*, nitrogen—*blue*, hydrogen—*white*, and sulfur—*yellow*

docking method. Docking is a method of bringing a target and ligand together to assess binding and its ability to form a stable structure [17]. This is done by bringing a biomolecule into a receptor binding site and moving the ligand to ascertain the location and conformation that is most advantageous for this binding to occur. Docking analysis can score each conformation to express the quality [17]. As aptamers and peptides can have multiple binding sites; using methods like docking is one avenue of evaluating the quality of peptide–aptamer binding. Docking method and its application to a specific case of peptide–aptamer binding are discussed in detail in a later section.

In addition to the docking, computational modeling based on transient dynamic analysis methods provide a detailed insight of the progression of peptide–aptamer binding process. Transient dynamic models are based on physics based mathematical formulations involving different length scales and features of interest. Based on the time and length scales involved, computational modeling and analysis approaches can be broadly classified into three main categories: quantum mechanics, atomistic modeling, and mesoscale dynamics [18].

Mesoscale Dynamics focuses on systems that involve billions of atoms and are generally based on larger geometrical sizes represented by appropriate physical laws [18]. The algorithms in this model are generally based on Newtonian Physics. In biological systems this type of modeling can be used for organ, large biomolecules, and their interface dynamics. However, modeling small atomic scale interactions using the mesoscale dynamics would cause inaccuracy with singular or small groups of atoms.

Atomistic Modeling is suitable for small systems where individual atoms and or small clusters of atoms are involved, and the phenomenon is influenced by the motion of individual atoms [18]. Molecular dynamics and Monte Carlo simulations are common examples of atomistic modeling [19]. These models can routinely explore time scales of picoseconds (10^{-12} s) to hundreds of nanoseconds. Although both approaches are based on interatomic potentials, they are inherently different. Monte Carlo modeling uses probabilistic approach to determine the lowest energy [20, 21]. On the other hand, the governing equations in molecular dynamics follow classical Newtonian mechanics [22]. This method is derived from Newton's equation of motion based on the selected force fields that defines the associated forces of the molecular interactions. This method is suitable to study dynamics and have been effectively used to model biological structures and interactions [19, 23, 24].

Quantum mechanics (QM) methods are highly suitable for simulating the electronic structure and properties of the system. Generally, chemical bond formation/breakage involves electron interactions between atoms [18]. Such bond formation and

breakage are accurately modeled in this approach. This method is the most accurate of the three methods for estimating the properties. However, this method is computationally expensive and is well suited only for extremely small systems. The high accuracy of this method is due to its ability to account for electron interactions through appropriate quantum mechanical equations. However, peptide–aptamer binding does not involve such electron interactions. Quantum mechanics methods are better suited for studying enzyme reactions, charge transfer, and analysis of chemically active regions in biomolecules [25, 26]. For this reason, atomistic based simulations are suitable for transient dynamics analysis of peptide–aptamer binding. In the following sections, we describe docking and transient dynamics approaches involving atomistic modeling of peptide–aptamer binding.

5 Docking Methods

Docking methods can be used in peptide–aptamer binding models for determining locations of binding, possible high-affinity sites and understanding structural isoforms. As described previously, binding is often depicted as "lock-and-key" mechanics where target molecules are considered the lock and its corresponding ligand the key [27]. Docking is a computational method of predicting the correct ligand as well as determining the structure and orientation of that ligand for a specific target molecule [28, 29]. The goal of docking is to optimize the binding event by considering the best fit between the ligand and target molecule. Two main approaches to docking are geometric and flexible. Geometric methods consider the structural geometry, sterics, shape of the ligand and its binding sites. This structure based method analyzes the binding site surface and its chemistry between the ligand and target molecule to determine the most complementary combination [30]. Investigating the binding site, one can define features that are distinct and necessary for docking before introducing a series of appropriate ligands. Though this method is sufficient in investigating the binding area, it considers molecules as rigid bodies. On the other hand, peptides and aptamers are flexible which means multiple conformations as well as binding areas are possible. There can be many ligands that fit into the binding area and vary in shape. Shape variations are not well accounted in geometric docking approach. Flexible docking is useful in investigating shape variations and customization beyond geometric constraints.

X-ray crystallographic inspection of proteins and ligands has shown that high-affinity ligands conform to the binding cavity to take advantage of the hydrogen bonding possibilities and hydrophobic interactions [27]. Flexible docking considers flexibility of

the molecules instead of treating them as completely rigid bodies [29]. Flexible docking analysis simulate the ligand near a target molecule active site and allow them to move based on energy minimization [31]. This allows for binding to occur in the most favorable conformation. Upon binding, the affinity of each conformation is scored and the confirmation with the best score is considered to be the optimized state of the biological complex. While this method looks at the most favorable conformation, there is also margin of error in the calculation when compared to wet lab experimentation [32]. The scoring and energy function calculation of a molecule or complex without solvent limits the method's accuracy in the interpretation of the experimental results.

6 Docking Studies of Peptide–Aptamer Binding

Docking methods have been applied to the discovery of peptide–aptamer binding of HIV Rev-RBE and BIV Tat-TAR complexes [33]. Bovine immunodeficiency virus (BIV) Tat peptides bound to the BIV TAR element (forming BIV tat peptide-TAR complexes) were solved using multidimensional NMR analysis. This combination was used as a control for the 17 amino acid peptide from the 34–50 residue of the human immunodeficiency virus (HIV) Rev protein that binds to a 30 nucleotide RNA aptamer. This aptamer, aptly named Rev binding element (RBE), was previously modeled with NMR constraints. Cedergren et al. provided a detailed strategy for the docking and modelling of the $Rev_{34–50}$ peptide–RBE aptamer complex interactions [33]. In their computational study, initial structures were treated as rigid bodies and individually minimized. Complexes were formed by combining the molecules into various binding conformations. Energetically most favorable conformations were determined by electrostatics and van der Waals interactions. The side chains and binding partners of those conformations were extensively analyzed. This process was applied to the known complex of BIV Tat-TAR to determine its efficacy by correctly identifying the orientation (5′–3′ relative to N and C termini) and register (juxtaposition) of two binding molecules. Validation was done by examining the binding free energy of the BIV Tat-TAR complex. Additional complexes were derived from the NMR solution of the BIV Tat-TAR complex by rotating the peptide and changing the register of the BIV Tat molecule in the binding site of the TAR element. It was found that the conformation identical to the NMR solution had the lowest binding energy. However, the calculations also showed that the docking method employed is insensitive to small changes in the peptide register. Nevertheless, the docking method can be used effectively in determining the global register and orientations of peptides docked with RNA aptamers.

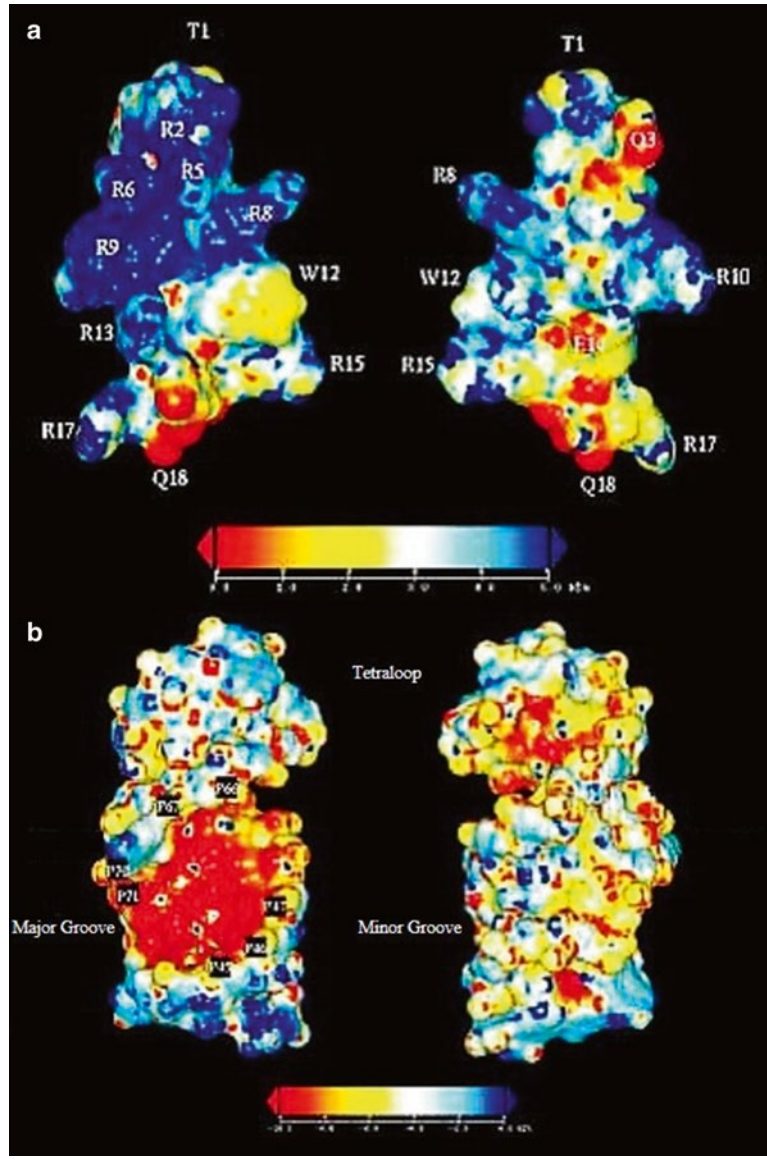


Fig. 4 Electrostatic images of the Rev Peptide and RBE. (**a**) Front and back views of the electrostatic surface of the Rev 34–50 peptide. Amino acid residues are indicated by single letter code and number. (**b**) Electrostatic surface of the RBE with groove and loop features indicated. Phosphate groups that are protected from chemical modification due to complex formation are indicated (reproduced from ref. [33] with permission)

After validating with the known BIV Tat-TAR complex the docking method was applied to the unknown structure of the $Rev_{34–50}$–RBE complex. Docking was guided by the electrostatic potentials and experimental data. Experimental work and the electrostatic potential surfaces indicated the major groove of the RBE aptamer to be the site of binding between key arginine residues (Arg2, Arg5, Arg6, Arg9, and Arg13) in the peptide (Fig. 4).

Fig. 5 Stereo view of the A-NL model of the $Rev_{34–50}$–RBE complex (reproduced from ref. [33] with permission)

Five initial models were generated from the *anti* (*A*) and *syn* (*S*) conformations of the peptide with the open end (NO) and tetraloop (NL) regions of the aptamer. The local minimum energy was calculated for each complex model in various conformations. The lowest energy complex was formed when the N terminus of the *anti* (A) or *syn* (S) peptide points towards the UUGG tetraloop region of the RBE. Further investigation of the binding energy showed that the RBE NL model with the *anti*-form of the peptide was the best model for the complex (Fig. 5). Based on the interaction of the peptide and RBE, the roles of the arginine residues and other side chains were also identified. The results indicated that the Arg2, Arg5, and Arg11 residues are important in binding but no single arginine side chain is singularly responsible.

The prediction of the optimal intermolecular geometry and interaction energy provides details of the binding area and the residues essential for the binding events. This work also showed that the major grove was the site of binding in the RNE RNA aptamer and predicted a possible structure based on the binding energy of peptides and aptamers [33].

Advances in docking studies in the mid 1990's through the 2000's led to the development of efficient docking analysis codes. More sophisticated analysis methods such as quantum mechanics

and molecular dynamics were available but restricted by computing power at that time [23]. The increase in computer power through parallel processing introduces allows for more detailed and accurate analysis of larger and complex problems. Docking can be improved with transient dynamics analysis, in particular molecular dynamics, which include fully solvated systems and more accurate models. As stated before, docking methods do not fully consider the flexibility of the molecules during binding. Movements such as the relaxation of active site around the ligand are still not considered in flexible docking. Further, such contributions make calculations of binding energy less reliable [34]. Such factors not considered in docking can be modeled in molecular dynamics and other transient dynamics methods. Peptide–aptamer modeling has now moved toward the more sophisticated analyses of molecular dynamics.

7 Transient Dynamic Analysis: Molecular Dynamics Methodology

One computational modeling technique applicable for the analysis of biomolecular motion and interactions is based on molecular dynamics (MD) modeling [35, 36]. MD methodology allows for the natural progression of the biomolecules in solution [37]. MD method has been applied to determine the chemical, physical, and mechanical properties of materials based on their molecular structures.

In the case of peptide–aptamer binding, the use of molecular dynamics (MD) analysis is most suitable for studying proteins and aptamers from a molecular level. The important question to be addressed is the behavior of the atoms within the macromolecules and how binding occurs under given conditions. Using the Newtonian equations of motion (Eq. 2), MD simulations can predict the movement of the atomic behavior of molecules as closely as possible to that of laboratory conditions [38].

$$M_{\mathrm{I}} \frac{\mathrm{d}^2 R}{\mathrm{dt}^2} = f_{\mathrm{I}}(R,t) \tag{2}$$

This equation relates the mass (M_{I}) and change in position (R) over time (t) to the force at each point in time (f_{I}). This calculation is done for every atom present in the simulation system at each time step. For accuracy of such simulations along with the structure, we must consider a suitable force field that is used for calculating the energy changes in the system.

For MD simulations, force fields are an important and influence the accuracy of the simulations [24, 39]. The energy summation (Eq. 3) is written as

$$E = E_{\mathrm{bonded}} + E_{\mathrm{nonbonded}} + E_{\mathrm{other}} \tag{3}$$

Bonded energies include bond stretching, bond angle, and torsional energy [40]. Non-bonded energy includes interactions such as van der Waals and electrostatic forces. Energetic contributions from other interactions are included in the E_{other} term. Over the years, several MD analyses codes and packages have been developed [41–43]. The choice of MD simulation analysis package depends upon the system studied and the availability of associated force field for the corresponding system of interest. Various forcefields such as CHARMM, AMBER, OPLS-AA, and GROMOS have been shown to be successful in simulating various physical, biological, and material systems [39, 44–46]. The AMBER force field in particular has often been used with peptides as well as RNA and DNA based molecules [47].

The molecular system configurations for the simulation studies are generally derived from a structure file in which the individual atoms, atom types, bonds, and positions are defined to form the initial molecular structure. In general, the starting structure files need to be converted into readable topology files in the respective data format for the chosen MD simulation analysis package. From the starting molecular conformations, the enclosed work space must be defined, solvated, minimized, and equilibrated via different established methods before simulation [48]. These preliminary steps are essential in establishing a stable initial system that best represents the real physical problem. The ability to follow a natural progression of a system is the advantage that MD gives over the docking method. Both quantitative parameters and visual analysis are used to assess and analyze the results from the dynamic simulation study. A variety of visualization data analysis software tools are available that are compatible with the output files of commonly used computational molecular modeling packages. Visualization analysis of the simulation trajectory provides guidance into the quantitative analysis.

One can also quantify structural changes that may occur in the simulation process. For example, RNA and DNA structures are held together through hydrogen bonds and it is possible to track these bonds throughout the simulation. Solvent interaction can also be investigated and quantified. The distribution of the solvent surrounding the molecules can also be considered as a method to understand the influence of the molecules on the environment.

This well-defined and constantly growing method of modeling biological molecules has become more prevalent over the years. In 1977, modeling globular protein dynamics in vacuum for short moments (10 ps) in time was a huge leap in computational applications [49]. The advances in parallel processing and dynamic algorithms in the late 1990s and 2000s allowed moderate scale molecular dynamic simulations of nucleic acids and small proteins in solution to be explored [50–53]. Today super, high end computing, greater dynamic algorithms and software have pushed the effectiveness of MD simulations to better understand, visualize, and analyze bimolecular events.

8 Molecular Dynamics of Peptide: Aptamer Binding

One recent study focused on molecular dynamics approach to study peptide–aptamer binding [54]. The initial work was done using the breast cancer aptamer–peptide binding combination of S2.2 Anti-Mucin 1 (MUC1) aptamer and a 8 amino acid Mucin 1 peptide. This combination was simulated using the GROMACS molecular dynamics analysis package. To parallel wet lab experiments, a 9 amino acid peptide–aptamer binding was simulated in 0.15 M NaCl solution at standard temperature and pressure. Visual analysis of this work showed the transient progression of peptide–aptamer binding as well as the identification of conformational changes (see Fig. 6). Simulation and analysis of the MUC1-G peptide and Anti-MUC1 aptamer show that binding occurs in the open loop region of the aptamer after 51 ns of simulation. In the loop region the thymine residue locks onto the arginine residue. As the simulation continues the thymine residue rotates and interacts with peptide backbone which helps the peptide and aptamer stay bound.

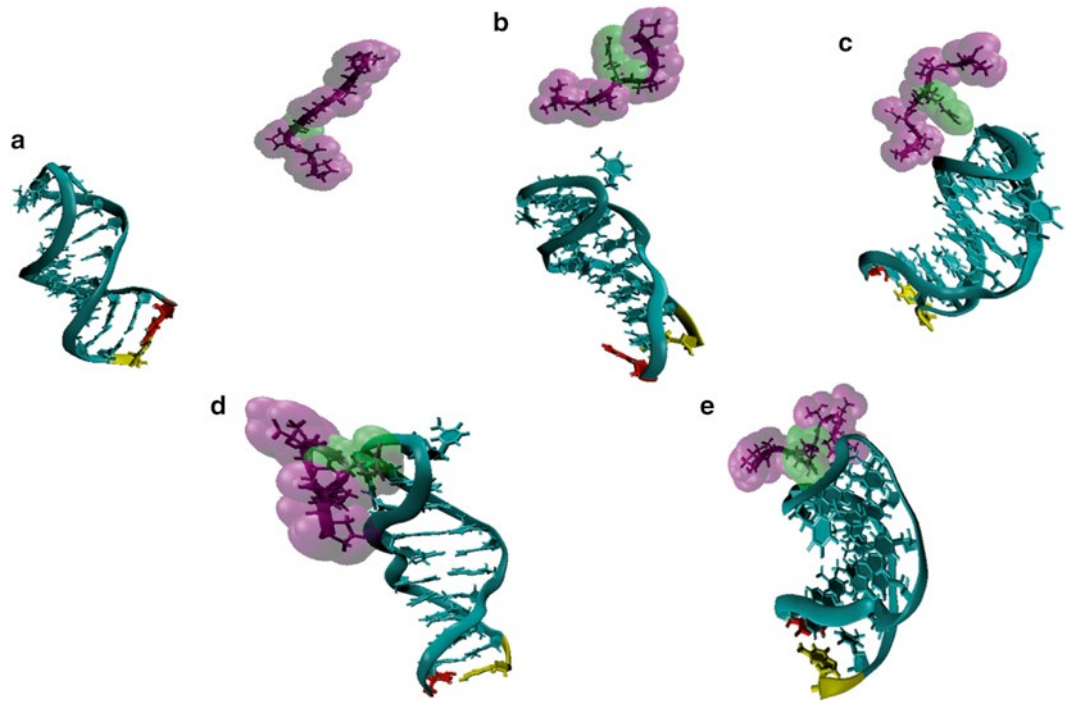


Fig. 6 Visualization of Anti-MUC1 aptamer (*blue*) and MUC1-G peptide (*purple*) binding 300 ns Simulation. (**a**) Starting peptide–aptamer configuration. (**b**) Peptide–aptamer configuration after 27 ns. (**c**) Interaction of the 11th thymine residue of the aptamer and the peptide backbone after 51 ns of simulation. (**d**) Continued peptide–aptamer interaction at the open loop region of the aptamer and the arginine residue after 127 ns. (**e**) Magnified image of the peptide–aptamer interaction at the end of simulation

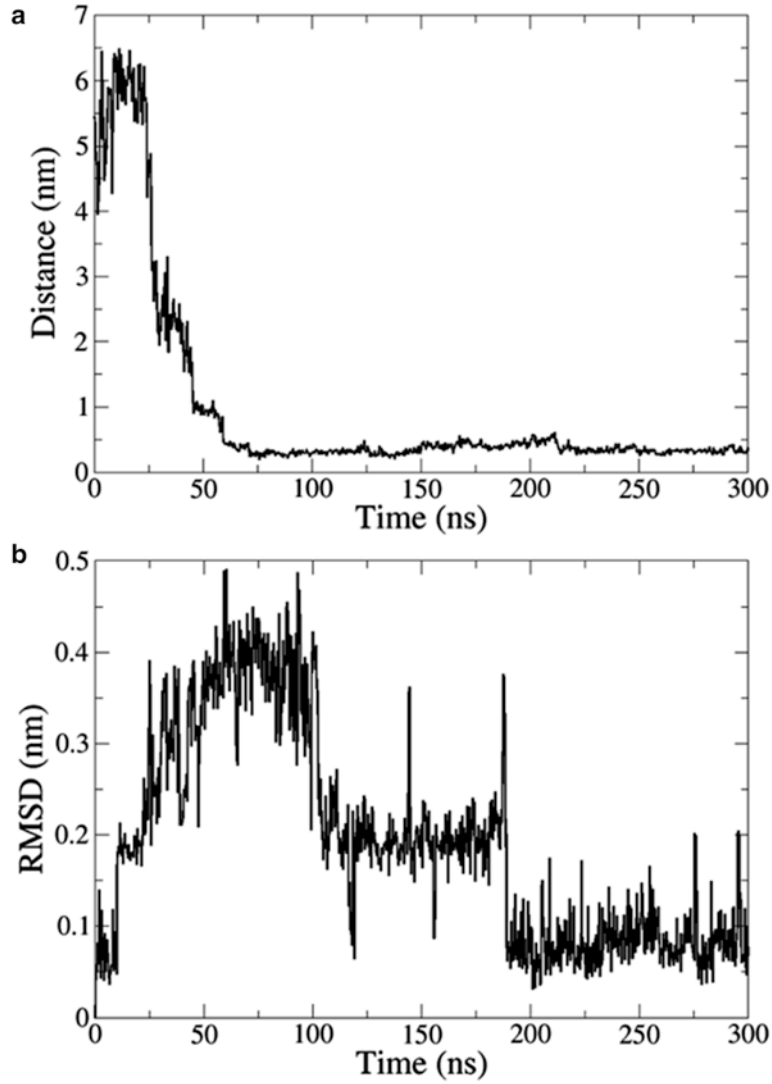


Fig. 7 Quantitative analysis of MUC1-G peptide and aptamer for 300 ns. (**a**) Atom Distance between the aptamer and peptide in the binding region (**b**) RMSD of the aptamer 5′–3′ ends during simulation

Repeated simulation of this peptide–aptamer combination with different initial configurations each converged and bound with the peptide interacting with the thymine loop region of the aptamer.

Quantitative analysis of this combination further reiterated the visual analysis results. As biomolecules bind using non-covalent interactions, we must consider electrostatics interactions, van der Waals forces and hydrophobic interactions. At this distance the atoms in the binding region should be less than 4.5 Å to indicate that the aptamer and peptide are noncovalently bound [55]. The distance between the aptamer and peptide in the binding site averaged 3.5 Å indicating binding has occurred (Fig. 7a).

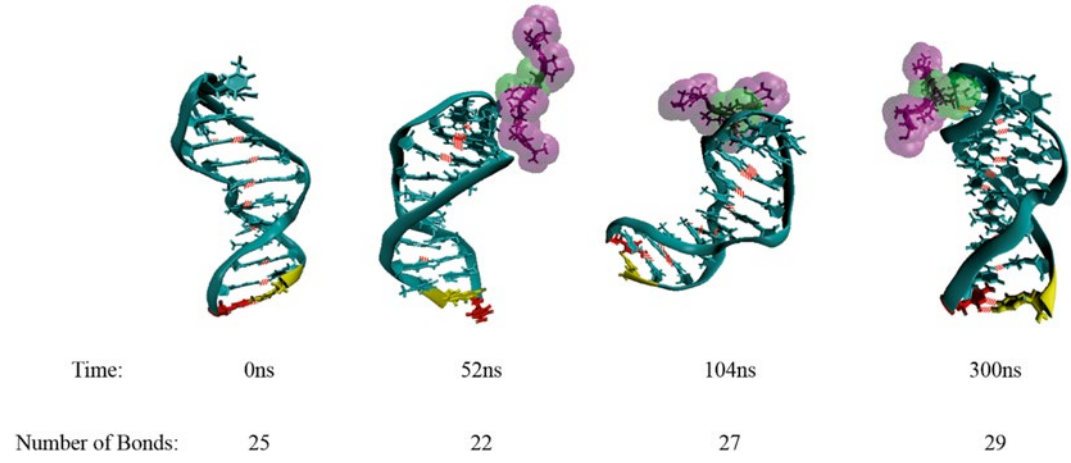


Fig. 8 Number of hydrogen bonds in the Anti-MUC1 Aptamer (*blue*) during simulation with the MUC1-G peptide (*purple*). Corresponding simulation snapshots at 25, 52, 104, and 300 ns of the anti-MUC1 Aptamer with the peptide and hydrogen bonds (*red dash lines*) are shown

The root mean squared deviation (RMSD) at the open 5′ and 3′ ends showed significant conformational changes in the aptamer that corresponded to binding events (Fig. 7b). Interactions of the peptide and aptamer residues during binding induce conformational changes along the backbone of the aptamer molecule. Though binding happens at the loop and open ends of the aptamer, changes in the aptamer appear to manifest in the 5′ and 3′ ends of the aptamer as it is the most loosely associated area of the aptamer backbone. The increase in the RMSD of the aptamer in this open region is due to the initial interaction between the MUC1 peptide and the aptamer. As the simulation continues and the peptide begins to interact more with the aptamer there are distinctive bouts of stability before the aptamer open ends settle. As the peptide forms a butterfly-like motif with the loop region of the aptamer, the open ends begin to settle to an RMSD value closer to that seen at the start of the simulation.

RMSD along with the visual analysis of the aptamer structure indicate changes in conformation that correspond to peptide–aptamer binding. Figure 8 shows the number of hydrogen bonds in one case of the aptamer with the MUC1 peptide as a function of simulation time studied. As the aptamer structure is held together by the formation of hydrogen bonds, disruptions in the structure would be evident in number of hydrogen bonds within the aptamer. In the beginning several hydrogen bonds were found to hold the structure together except in the loop and helical regions of the aptamer. The decrease in the number of hydrogen bonds shown at 52 ns (1 ns after the interaction) is due to the arginine residue of the peptide disrupting the bonds at the open ends. Extended time after the binding event indicates that the number of hydrogen bonds increases as the hydrogen bonds in the open ends reform.

The simulation results were quantitatively analyzed for the conformational changes, and the overall behavior of aptamer and peptide system was observed from a molecular view point, which is not always possible in the wet lab experiments. The observed changes in structure are reflected in the atomic distance, RMSD values, and hydrogen bonds. The changes in hydrogen bonds show direct correspondence to structural changes resulting from binding events. The present MD simulations, analyses, and discussions clearly show that the aptamer and peptide binding could be efficiently simulated and analyzed using computational methods.

9 Summary

Computational modeling provides effective means to understand peptide–aptamer bindings. This method can offer additional insights into aptamer selection and binding processes by providing a visual and quantitative scope through biomolecular modeling. Computational docking was used to study peptide–aptamer combination. Its application provides a fundamental view into the binding site and ligand identification and interaction. However, the natural movement and behavior of the aptamer and peptide cannot fully be investigated with docking alone. To interpret the experimental peptide–aptamer complexes and gain a greater insight into their binding interactions one needs to look into the transient dynamic analysis. Molecular dynamics provides more in-depth look into the natural progression as demonstrated by an example case of peptide–aptamer binding discussed in this chapter.

Acknowledgments

This work was supported in part by the U. S. Army Research Office via award/contract no. W911NF-11-1-0168. We thank Dr. M. Sandros for scientific discussions during the course of this work.

References

1. Jain KK (2010) The handbook of biomarkers. Springer, New York, NY
2. Strehlitz B, Nikolaus N, Stoltenburg R (2008) Protein detection with aptamer biosensors. Sensors 8:4296–4307
3. Erickson D, Mandal S, Yang A, Cordovez B (2008) Nanobiosensors: optofluidic, electrical and mechanical approaches to biomolecular detection at the nanoscale. Microfluid Nanofluid 4:33–52
4. Song S, Wang L, Li J, Fan C, Zhao J (2008) Aptamer-based biosensors. Trends Anal Chem 27:108–117
5. Wang J (2000) From DNA biosensors to gene chips. Nucleic Acids Res 28:3011–3016
6. McCauley TG, Hamaguchi N, Stanton M (2003) Aptamer-based biosensor arrays for detection and quantification of biological macromolecules. Anal Biochem 319:244–250
7. Clark SL, Remcho VT (2002) Aptamers as analytical reagents. Electrophoresis 23:1335–1340
8. Mascini M, Palchetti I, Tombelli S (2012) Nucleic acid and peptide aptamers: fundamentals and bioanalytical aspects. Angew Chem Int Ed 51:1316–1332

9. Colas P, Cohen B, Jessen T, Grishina I, McCoy J, Brent R (1996) Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. Nature 380:548–550
10. James W (2001) Nucleic acid and polypeptide aptamers: a powerful approach to ligand discovery. Curr Opin Pharmacol 1:540–546
11. Stoltenburg R, Reinemann C, Strehlitz B (2007) SELEX – a (r)evolutionary method to generate high-affinity nucleic acid ligands. Biomol Eng 24:381–403
12. Baines IC, Colas P (2006) Peptide aptamers as guides for small-molecule drug discovery. Drug Discov Today 11:334–341
13. Ferreira CS, Matthews CS, Missailidis S (2006) DNA aptamers that bind to MUC1 tumour marker: design and characterization of MUC1-binding single-stranded DNA aptamers. Tumour Biol 27:289–301
14. Ferreira C, Papamichael K, Guilbault G, Schwarzacher T, Gariepy J, Missailidis S (2008) DNA aptamers against the MUC1 tumour marker: design of aptamer–antibody sandwich ELISA for the early diagnosis of epithelial tumours. Anal Bioanal Chem 390: 1039–1050
15. Tramontano A (2006) Protein structure prediction: concepts and applications. Wiley-VCH, Weinheim
16. Bader DA (2004) Computational biology and high-performance computing. Commun ACM 47:34–41
17. Schneider G, Baringhaus K-H (2008) Molecular design: concepts and applications. John Wiley & Sons, New York, NY
18. Gomperts R, Renner E, Mehta M (2005) Enabling technologies for innovative new materials. Am Lab 37:12–14
19. Sim AYL, Minary P, Levitt M (2012) Modeling nucleic acids. Curr Opin Struct Biol 22:273–278
20. Berg BA (2004) Markov chain Monte Carlo simulations and their statistical analysis: with web-based fortran code. World Scientific, Hackensack, NJ
21. Scherer POJ (2010) Computational physics: simulation of classical and quantum systems. Springer, New York, NY
22. Rapaport DC (2004) The art of molecular dynamics simulation. Cambridge University Press, Cambridge
23. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 9:646–652
24. Karplus M, Petsko GA (1990) Molecular dynamics simulations in biology. Nature 347:631–639
25. Senn HM, Thiel W (2009) QM/MM methods for biomolecular systems. Angew Chem Int Ed Engl 48:1198–1229
26. Náray-Szabó G, Oláh J, Krámos B (2013) Quantum mechanical modeling: a tool for the understanding of enzyme reactions. Biomolecules 3:662–702
27. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
28. Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 32:335
29. Knegtel RMA, Kuntz ID, Oshiro CM (1997) Molecular docking to ensembles of protein structures. J Mol Biol 266:424–440
30. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288
31. Österberg F, Åqvist J (2005) Exploring blocker binding to a homology model of the open hERG K+ channel using docking and molecular dynamics methods. FEBS Lett 579:2939–2944
32. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ 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
33. Srinivasan J, Leclerc F, Xu W, Ellington AD, Cedergren R (1996) A docking and modelling strategy for peptide–RNA complexes: applications to BIV Tat–TAR and HIV Rev–RBE. Fold Des 1:463–472
34. Okimoto N, Futatsugi N, Fuji H, Suenaga A, Morimoto G et al (2009) High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations. PLoS Comput Biol 5:e1000528
35. Auffinger P, Westhof E (1998) Simulations of the molecular dynamics of nucleic acids. Curr Opin Struct Biol 8:227–236
36. Jayapal P, Mayer G, Heckel A, Wennmohs F (2009) Structure–activity relationships of a caged thrombin binding DNA aptamer: insight gained from molecular dynamics simulation studies. J Struct Biol 166:241–250
37. Hansson T, Oostenbrink C, van Gunsteren W (2002) Molecular dynamics simulations. Curr Opin Struct Biol 12:190–196
38. Stavrakoudis A, Tsoulos I, Uray K, Hudecz F, Apostolopoulos V (2011) Homology modeling and molecular dynamics simulations of MUC1-9/H-2K(b) complex suggest novel binding interactions. J Mol Model 17:1817–1829
39. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM 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
40. Guvench O, MacKerell AD (2008) Comparison of protein force fields for molecular dynamics

simulations. Methods Mol Biol 443:63–88, Molecular modeling of proteins. A Kukol (ed.), Humana Press, pp. 63–88
41. Plimpton S (1995) Fast parallel algorithms for short-range molecular dynamics. J Comput Phys 117:1–19, http://lammps.sandia.gov
42. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. Comput Phys Commun 91:43–56
43. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E et al (2005) Scalable molecular dynamics with NAMD. J Comput Chem 26:1781–1802
44. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD et al (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616
45. Oostenbrink C, Villa A, Mark AE, van Gunsteren WF (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25:1656–1676
46. 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
47. Guvench O, MacKerell A Jr (2008) Comparison of protein force fields for molecular dynamics simulations. In: Kukol A (ed) Molecular modeling of proteins, vol 443. Humana Press, Totowa, NJ, pp 63–88
48. Hünenberger P (2005) Thermostat algorithms for molecular dynamics simulations. In: Holm C, Kremer K (eds) Advanced computer simulation, vol 173. Springer, Berlin, pp 105–149
49. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267: 585–590
50. Freddolino PL, Liu F, Gruebele M, Schulten K (2008) Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. Biophys J 94:L75–L77
51. Schaeffer RD, Fersht A, Daggett V (2008) Combining experiment and simulation in protein folding: closing the gap for small model systems. Curr Opin Struct Biol 18:4–9
52. Duan Y, Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 282:740–744
53. Pérez A, Luque FJ, Orozco M (2007) Dynamics of B-DNA on the microsecond time scale. J Am Chem Soc 129:14739–14745
54. Rhinehardt K, Mohan R, Srinivas G, Kelkar A (2013) Computational modeling of peptide - aptamer binding in biosensor applications. Int J Biosci Biochem Bioinform 3: 639–642
55. Schalley CA (2012) Analytical methods in supramolecular chemistry, vol 1. Wiley-VCH Verlag GmbH & Company KGaA, Weinheim

Index

Peng Zhou and Jian Huang (eds.), *Computational Peptidology*, Methods in Molecular Biology, vol. 1268,
DOI 10.1007/978-1-4939-2285-7, © Springer Science+Business Media New York 2015

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Z