NATO Science for Peace and Security Series - B:
Physics and Biophysics

Biophysics and Structure to Counter Threats and Challenges

Edited by
Joseph D. Puglisi
Manolia V. Margaris

Springer

NATO OTAN

This publication is supported by:

The NATO Science for Peace and Security Programme

Biophysics and Structure to Counter Threats and Challenges

NATO Science for Peace and Security Series

This Series presents the results of scientific meetings supported under the NATO Programme: Science for Peace and Security (SPS).

The NATO SPS Programme supports meetings in the following Key Priority areas: (1) Defence Against Terrorism; (2) Countering other Threats to Security and (3) NATO, Partner and Mediterranean Dialogue Country Priorities. The types of meeting supported are generally "Advanced Study Institutes" and "Advanced Research Workshops". The NATO SPS Series collects together the results of these meetings. The meetings are coorganized by scientists from NATO countries and scientists from NATO's "Partner" or "Mediterranean Dialogue" countries. The observations and recommendations made at the meetings, as well as the contents of the volumes in the Series, reflect those of participants and contributors only; they should not necessarily be regarded as reflecting NATO views or policy.

Advanced Study Institutes (ASI) are high-level tutorial courses intended to convey the latest developments in a subject to an advanced-level audience

Advanced Research Workshops (ARW) are expert meetings where an intense but informal exchange of views at the frontiers of a subject aims at identifying directions for future action

Following a transformation of the programme in 2006 the Series has been re-named and re-organised. Recent volumes on topics not related to security, which result from meetings supported under the programme earlier, may be found in the NATO Science Series.

The Series is published by IOS Press, Amsterdam, and Springer, Dordrecht, in conjunction with the NATO Public Diplomacy Division.

Sub-Series

A.	Chemistry and Biology	Springer
B.	Physics and Biophysics	Springer
C.	Environmental Security	Springer
D.	Information and Communication Security	IOS Press
E.	Human and Societal Dynamics	IOS Press

http://www.nato.int/science
http://www.springer.com
http://www.iospress.nl

Series B: Physics and Biophysics

Biophysics and Structure to Counter Threats and Challenges

edited by

Joseph D. Puglisi
Stanford University, Stanford
CA, USA

and

Manolia V. Margaris
Stanford University, Stanford
CA, USA

Published in cooperation with NATO Emerging Security Challenges Division

Proceedings of the NATO Advanced Study Institute on
Biophysics and Structure to Counter Threats and Challenges
Erice, Sicily, Italy
22 June–2 July 2010

Library of Congress Control Number: 2012953288

ISBN 978-94-007-4925-2 (PB)
ISBN 978-94-007-4922-1 (HB)
ISBN 978-94-007-4923-8 (e-book)
DOI 10.1007/978-94-007-4923-8

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

Preface

This volume is a collection of articles from the Proceedings of the International School of Biological Magnetic Resonance 10th Course: Biophysics and Structure to Counter Threats and Challenges. This NATO Advance Study Institute (ASI) was held in Erice, Sicily (Italy) at the Ettore Majorana Foundation and Centre for Scientific Culture on 22 Jun through 2 July 2010. This ASI brought together a diverse group of experts who span virology, biology, biophysics, chemistry, physics and engineering. Prominent lecturers representing world renowned scientists from nine (9) different countries, and students from around the world representing eighteen (18) countries, participated in the ASI organized by Profs. Joseph Puglisi (Stanford University, USA) and Alexander Arseniev (Moscow, RU).

The ASI focussed on these interdisciplinary approaches to treatment, detection and understanding of emerging infectious diseases. Biology has been revolutionized by the application of genomics, proteomics and sequencing. Virologists have delineated novel pathways of both viral and bacterial infection, using new physical approaches, which highlight the complex interplay of host and pathogen. The goal of biophysics is to bridge these biological discoveries to the molecular scale, yielding new opportunities for therapeutic development. The opportunity for students to interact intimately with great scientists during the long duration of the ASI is perhaps the most thrilling part of the meeting. The theme of the ASI embraced three topics: (1) the pathway of infection by viruses and bacteria; (2) the role of biophysics in designing novel therapeutics; and (3) new technologies to improve detection of pathogens and provide for a cleaner environment. The didactic focus of the ASI was the techniques underlying these areas, and included NMR and fluorescence spectroscopy, x-ray crystallography, electron and optical microscopy, single-molecule spectroscopy and computational methods.

The central hypothesis underlying this ASI was that interdisciplinary research, merging principles of physics, chemistry and biology, can drive new discovery in detecting and fighting chemical and bioterrorism agents, lead to cleaner environments and improved energy sources and help propel development in NATO partner countries. At the end of the ASI, students had an appreciation of how to apply each technique to their own particular research problem and to demonstrate that

multifaceted approaches and new technologies are needed to solve the biological challenges of our time. The course succeeded in training a new generation of biologists and chemists who will probe the molecular basis for life and disease. We thank NATO, and in particular the Science for Peace and Security Division, for supporting the course and this book of proceedings.

Stanford University
Stanford, California

Joseph D. Puglisi
Manolia V. Margaris

Contents

Contributors

Elisabeth D. Balitskaya Laboratory of Biomolecular Modeling, Russian Academy of Sciences, M. M. Shemyakin & Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia

Department of Bioengineering, M.V. Lomonosov Moscow State University, Biological faculty, Moscow, Russia

Mark Bowen Department of Physiology and Biophysics, Stony Brook University Medical Center, Stony Brook, NY, USA

Axel T. Brunger The Howard Hughes Medical Institute and Department of Molecular and Cellular Physiology, Stanford University, Stanford, CA, USA

Steven Chu Departments of Physics and Molecular and Cell Biology, University of California, Berkeley, CA, USA

Lawrence Berkeley National Laboratory, Berkeley, CA, USA

United States Department of Energy, Washington, DC, USA

Roman G. Efremov Laboratory of Biomolecular Modeling, Russian Academy of Sciences, M.M. Shemyakin & Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia

Måns Ehrenberg Department of Cell and Molecular Biology, Biomedical Center, Uppsala University, Uppsala, Sweden

Angela M. Gronenborn Department of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

Leonardus M.I. Koharudin Department of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

David M.J. Lilley College of Life Sciences, Cancer Research UK Nucleic Acid Structure Research Group, The University of Dundee, Dundee, UK

Alexander McPherson Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA, USA

Harriet Mellenius Department of Cell and Molecular Biology, Biomedical Center, Uppsala University, Uppsala, Sweden

Ivan V. Ozerov Laboratory of Biomolecular Modeling, Russian Academy of Sciences, M.M. Shemyakin & Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia

Department of Bioengineering, M.V. Lomonosov Moscow State University, Moscow, Russia

Sandra E. Pineda-Sanabria Department of Biochemistry, School of Molecular and Systems Medicine, University of Alberta, Edmonton, AB, Canada

Elisabetta Viani Puglisi Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA

Joseph D. Puglisi Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA

Ian M. Robertson Department of Biochemistry, School of Molecular and Systems Medicine, University of Alberta, Edmonton, AB, Canada

Pavel Strop The Howard Hughes Medical Institute and Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences, Structural Biology, and Photon Science, Stanford University, Stanford, CA, USA

Brian D. Sykes Department of Biochemistry, School of Molecular and Systems Medicine, University of Alberta, Edmonton, AB, Canada

Marija Vrljic The Howard Hughes Medical Institute and Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences, Structural Biology, and Photon Science, Stanford University, Stanford, CA, USA

Keith R. Weninger Department of Physics, North Carolina State University, Raleigh, NC, USA

Ada Yonath Department of Structural Biology, The Weizmann Institute of Science, Rehovot, Israel

Chapter 1
Macromolecular Models by Single Molecule FRET

Axel T. Brunger, Pavel Strop, Marija Vrljic, Mark Bowen, Steven Chu, and Keith R. Weninger

Abstract Single molecule fluorescence energy transfer (FRET) experiments enable investigations of macromolecular conformation and folding by the introduction of fluorescent dyes at specific sites in the macromolecule. Multiple such experiments can be performed with different labeling site combinations in order to map complex conformational changes or interactions between multiple molecules. Distances that are derived from such experiments can be used for determination of the fluorophore positions by triangulation. When combined with a known structure of the macromolecule(s) to which the fluorophores are attached, a three-dimensional model of the system can be determined by docking calculations. Here we discuss recent applications of single molecule FRET to obtain a model of the synaptotagmin-1:SNARE complex and to study the conformation of PSD-95.

A.T. Brunger (✉) • P. Strop • M. Vrljic
The Howard Hughes Medical Institute and Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences, Structural Biology, and Photon Science, Stanford University, James H. Clark Center, Room E300-C, Stanford, CA 94305, USA
e-mail: brunger@stanford.edu

M. Bowen
Department of Physiology and Biophysics, Stony Brook University Medical Center, Stony Brook, NY 11794-8661, USA

S. Chu
Formerly Lawrence Berkeley National Laboratory and Physics and Cell Biology Departments, University of California at Berkeley, Berkeley, CA 94720, USA

K.R. Weninger (✉)
Department of Physics, North Carolina State University, Raleigh, NC 27695, USA
e-mail: keith_weninger@ncsu.edu

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_1,

1.1 Introduction

Detailed structural studies extending to the atomic level are effective approaches to reveal the molecular mechanisms underlying function of biological molecules. High-resolution methods such as X-ray diffraction crystallography and nuclear magnetic resonance have led the way in providing the highest spatial information, but a host of other methods, such as small angle X-ray scattering, cryo-electron microscopy, hydrodynamic assays (gel filtration chromatography, light scattering, and analytical ultra centrifugation), and spectroscopic measures of circular dichroism and fluorescence also can provide a wealth of knowledge about molecular structure. Among these approaches, fluorescence resonance energy transfer (FRET) studies have flourished in recent years as a result of the capability to detect the signal reporting intra and intermolecular distances from samples as small as single molecules [17, 57].

The benefits of studying single molecules have opened new avenues of investigation in biomolecular science. By recording properties and dynamics of single molecules one at a time, the effects of averaging that are inherent in ensemble studies are absent, allowing discovery of phenomena not otherwise observable. The single molecule approach is particularly effective at revealing heterogeneous behaviors across different individual molecules within a sample and also reporting dynamic trajectories of molecules without the need for synchronous behavior across a population. Single molecule FRET (smFRET) is well suited for structural studies because it provides a unique tool with the applicability to transient and dynamic molecular conformations and can reveal weak interactions that often are not resolvable when averaging over a larger sample.

Motivated by the dramatic successes of smFRET in the past decade, there has been a rapid development of instrumentation, sample preparation, labeling, and data analysis [14, 41, 50, 59]. By combining multiple smFRET experiments involving different labeling site combinations one obtains a network of FRET-derived distances between the labeling sites. If the distance network is augmented by structural information about the molecules to which the fluorophores are attached, powerful computational approaches can be used to obtain three-dimensional models of the entire system. As a particular example for multi-molecule docking we discuss the determination of the model of the synaptotagmin-1:SNARE complex derived from smFRET-derived distances [11]. We also discuss a recent smFRET study of the conformation of PSD-95 [42].

1.2 Förster Theory

FRET occurs between two fluorescent dyes when the emission spectrum of an excited donor fluorophore overlaps the absorption spectrum of a nearby acceptor fluorophore [23]. For applications with biomolecules, site specific chemistry is

generally used to link the fluorescent dyes at known locations on the molecule so that FRET between the fluorophores is interpretable at the distance between those points [62]. The FRET efficiency E_{FRET} depends strongly on the distance R between fluorophores:

$$E_{FRET} = \frac{1}{1 + (R/R_0)^6} \tag{1.1}$$

where the Förster radius R_0 is a donor/acceptor-pair specific constant (the distance at which the FRET efficiency is 50%). When the distance between the donor and acceptor fluorphores is ½ R_0, the FRET efficiency is nearly maximal and, so, any further decrease in distance is difficult to measure. Conversely, if the distance increases beyond 2 R_0 then, the distance dependence is also very shallow. Thus, the most sensitive range for a typical FRET experiment is in the distance range ½ to 2 R_0.

The Förster radius R_0 depends on the spectroscopic properties of the FRET fluorophore pair and the surrounding medium of the fluorophores [32]. The FRET efficiency can be obtained by measuring either the donor and acceptor fluorescence intensities, or the donor lifetimes in the presence and absence of an acceptor [55]

$$E_{FRET} = \left[1 + \gamma \left(\frac{I_D}{I_A}\right)\right]^{-1} \tag{1.2}$$

where I_{D} and I_{A} are donor (D) and acceptor (A) fluorescence intensities. The factor γ combines the probabilities of the donor and acceptor fluorophores to relax to the ground state from the fluorescent excited state by emitting a photon and the likelihood of experimentally detecting emitted photons.

In summary, the Förster theory that relates measured donor and acceptor intensities (I_{D} and I_{A}) to the interfluorophore distance R is dependent on two factors involving fluorophore and instrument properties: R_o (the Förster Radius) (Eq. 1.1) and γ (Eq. 1.2). R_o is generally different for each specific pair of fluorescent dyes. The γ factor depends on a combination of fluorophore and instrument properties.

1.3 Corrections to Raw Single Molecule Fluorescence Intensity Data

The raw fluorescence intensity data (I_{A} and I_{D} in Eq. 1.2) must be corrected by background scattering, leakage of fluorescence intensity of donor and acceptor fluorophores into each other's detection channels, and by the γ factor. Leakage of one fluorophore's emission into the detection channel of the other fluorophore is typically characterized using measurements of samples prepared with only one of the fluorophores. The fraction of the emission for each fluorophore that appears

in the unintended channel is a fixed function of the specific configuration of the detection channel as long as the emission spectral density is constant during the experiment (see ref. [12] for a notable exception where emission spectral density changes). This fraction is subtracted from measured intensities when analyzing the FRET efficiency data. There are several approaches to background subtraction for experiments using immobilized molecules. They all are generally based around the fact that if the surface density of immobilized molecules is sufficiently dilute then background contributions to the emission intensity can be estimated from locations near observed fluorescent spots that are free of other fluorescent molecules.

1.4 Empirical Determination of γ

The γ factor accounts for differences between the donor and acceptor fluorophores in the probability that emitted photons will be detected (detector efficiency) and the probability of photon emission upon excitation (quantum yield). These properties can be determined experimentally by measuring the detected intensities of acceptor and donor fluorophores separately as a function of illumination power, and by measuring the relative quantum yields of the fluorophores attached to the particular biomolecule under study. Quantum yields are commonly measured by comparing ensemble fluorescence measurements of samples with known concentrations to the emissions of quantum yield standards (for example, rhodamine 101 in ethanol). For single molecule studies of freely diffusing molecules, alternative approaches to determine the γ factor that do not require independent determination of detection efficiencies and quantum yields are available. One such method exploits the linear relationship between the apparent FRET ratio and the stoichiometry by using an alternative laser excitation scheme and anti-correlated photobleaching events [35]. Another approach relies on measurement of the lifetime of the fluorescent dye excited states [55].

A systematic study of γ normalization methods using DNA found that determination of γ from the magnitude of anti-correlated photobleaching events was the most effective at achieving convergence of measured FRET values and high resolution structures of duplex DNA [41]. Variability in γ raises questions as to how normalization should be applied to recover FRET efficiency effectively. It is possible to measure γ once for an instrument and apply this as a "universal" normalization to all measurements using the same dyes and optical path. This method does not account for the observed sample-to-sample variation in γ. One could normalize an individual data set using the data set specific mean or "global" γ factor. This does not account for variation within a data set or the outliers with γ values significantly different from the mean. To account for these outliers, one would have to normalize each molecule with an "individual" γ factor. These approaches have all been previously reported [13, 27, 28, 56, 67].

A systematic comparison of these approaches [41] found that the global γ factor was sufficient to correct the mean FRET efficiency. A universal γ factor was less effective because it fails to account for actual variability in γ between samples. However, only normalization with individual γ factors for each molecule resulted a narrowing of width of smFRET distributions. Both systematic factors, including instrumental or photophysical effects as well as dynamic molecular motion, can contribute to broadening the widths of FRET efficiency distribution peaks in smFRET experiments [8, 43, 44]. This study showed that γ variability also contributes to broadening FRET histogram widths. Molecules with FRET values near the edges of the FRET distribution also commonly had outlying γ values and per molecule γ normalization brings these values closer to the mean. The γ outliers may be due to differences in detection efficiency introduced during image recording or processing. In agreement with this notion, it has been noted that one can change the value of γ by misaligning the detectors when measuring diffusing single molecules [35, 44]. Empirical measurements of the terms composing γ cannot account for aberration this of kind. As a consequence of the fact that γ outliers are not representative of the population, applying a γ cutoff as a means of selecting accepted molecules could further affect peak width and shape.

1.5 Empirical Determination of R_0

In principle, R_0 can be calculated from the spectral overlap of the fluorophore pairs, the donor quantum yield, and the orientation factor [32]. Some of these parameters can be obtained experimentally, but the orientation factor requires some knowledge of the fluorophores's dynamics with respect to the attached molecule. For the synaptotagmin-1:SNARE complex [11], an entirely empirical approach was used to calibrate R_o inspired by previous work [1, 2]. For a fluorophore pair (Alexa555:Alexa647) attached to one of the two rigid C2 domains of synaptotagmin-1 FRET efficiency was measured and compared with the calculated value using the fluorophore separation from a crystal structure of Syt1 [24]. The resulting FRET efficiency distributions produced a single Gaussian peak with a width consistent with shot noise [11]. The FRET efficiency and fluorophore separation yielded an empirical value of $R_o = 5.55$ nm (using $\gamma = 1$, Eqs. 1.1 and 1.2), compared to the theoretical R_o for the Alexa555:Alexa647 fluorophore pair of 5.1 nm [30]. R_o is expected to deviate from the theoretical value due to changes in the fluorophore microenvironment when conjugated to the molecule of interest. In the work with synaptotagmin-1, the spread of the R_o values derived from three label-site combinations (using the same fluorescent dyes but different amino acid attachment sites) was 0.23 nm, which is smaller than the error bounds used in the docking calculation [11].

1.6 Simulation of Fluorophore Center Positions

As a pre-requisite for docking and fitting calculations, the average fluorophore center positions have to be computed relative to the position of the molecule or domain that the fluorophore is attached to. Depending on the particular fluorophore and linker, the fluorophore center position is separated from the coordinates of the covalently attached residue (often generated by site-specific mutagenesis) by ~ 1 nm. In earlier work, the fluorophore center position was simply taken at the protein residue site or displaced by a certain amount away from the molecule's center [49]. To obtain a more precise estimate, a molecular dynamics simulation can be used to obtain the average position of the fluorophore center using an atomic model of the fluorophore linked to a protein at the residue position used for labeling [11, 66] (Fig. 1.1).

For these particular simulations, the protein atoms of the molecule to which the fluorophore is attached were kept fixed while the linker and the fluorophore

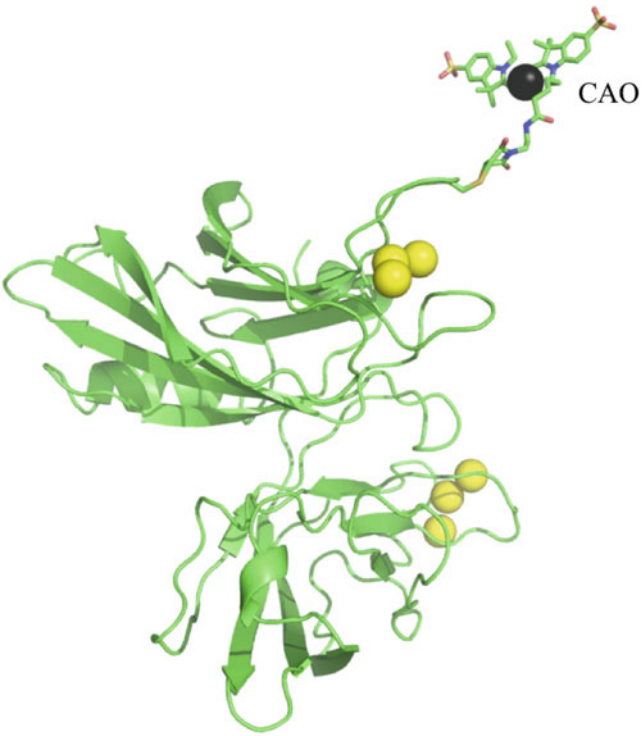


Fig. 1.1 Cy3 fluorescent dye (*sticks*) attached to the C2B domain of synaptotagmin-3 (*green cartoon*) [66]. *Yellow spheres* indicate the Ca^{2+} positions of the crystal structure of synaptotagmin-3. *Black sphere* indicates the position of atom CAO of the dye that is used to define the center of the fluorescent dye (Color figure online)

atoms were allowed to move. Chemical structures are available for some of the commonly used fluorescent fluorophores, such as Cy3, Cy5, and Alexa 647 [66]. A large number of simulations were performed starting from different initial velocities in order to obtain good conformational sampling.

1.7 Docking Calculations

Determination of a three-dimensional model from smFRET derived distances is reminiscent of rigid body docking approaches using nuclear magnetic resonance (NMR)-derived distances between protons [19, 61]. We discuss here the method that was used to determine a model of the synaptotagmin-1:SNARE complex [11]. In the case of smFRET data, the distances refer to distances between fluorophore centers. The fluorophore center position is treated as a "pseudoatom" that is rigidly associated with the molecule to which the fluorophore is covalently attached; the position of the pseudoatom is taken as the average position of the fluorophore center relative to the molecule, as obtained from a molecular dynamics simulation (see above) [11]. The term pseudoatom is used since these points have no chemical energy terms associated with them during the docking calculation, but are rather restricting the possible conformations of the rigid molecule or domain that the pseudoatom is attached to. Extensive torsion angle/rigid body molecular dynamics simulations [52] were performed using a simulated annealing slow-cooling protocol. The total energy function consists of a repulsive term for the nonbonded interactions (i.e., excluding electrostatic and attractive van der Waals terms) [20] and the distance restraints term. This type of energy function is widely used for three-dimensional structure determination based NMR data [6]. A harmonic square-well potential was used to restrain the distances between fluorophore center pseudoatom positions [4]. The smFRET efficiencies were converted to distances (Eqs. 1.1 and 1.2). FRET efficiency becomes less sensitive to fluorophore separation at distances much less and much greater than R_o (see Eq. 1.1). Therefore, variable bounds were used for the square well potential depending how close the expected distance is to R_o [11].

Many trials (typically $\sim$ 1,000) with different randomly assigned orientations of domains and molecules, different relative conformations of flexible domains, and initial velocities were performed for each set of calculations (Schema 1.1). Each resulting model was then characterized by the root mean square (rms) deviation between the distances predicted by the model and the distance ranges used as the square well potentials derived from the FRET efficiency measurements. The solutions of the docking simulations were sorted by rms deviation satisfaction in increasing order. The solutions were clustered according to the rms deviation using an algorithm implemented in the program HADDOCK [15]. For each cluster, the structure with the best distance satisfaction was used for subsequent analysis. The derived models could also be further refined using local perturbation and refinement in docking programs to optimize local interactions. This step would allow for side chain refinement and inclusion of electrostatics and van der Waals energy terms.

- **SNARE complex and C2 domains kept rigid**
- **C2A-C2B linker flexible**
- **distance bounds +/- 2.5 Å , +/- 5 Å**
- **1000 simulated annealing docking calculations starting from random placements of the domains**
- **Cluster analysis of the top solutions**

Schema 1.1 Key points of the docking calculations for the synaptotagmin-1:SNARE complex. Further details can be found in the single molecule FRET tutorial of CNS, version 1.3, section on "Docking calculations with single-molecule FRET derived distances", http://cns-online.org [60]

1.8 smFRET Derived Model of the Synaptotagmin-1:SNARE Complex

Ca^{2+}-induced membrane fusion of synaptic vesicles at synapses is the central phenomenon that results in triggered inter-neuron signaling. The membrane protein synaptotagmin-1 is the Ca^{2+} sensor for synchronous neurotransmitter release [22, 51]. Highly coordinated interactions among synaptotagmin, SNARE (soluble N-ethylmaleimide-sensitive factor attachment protein receptor) proteins and other neuronal factors are required to create robust and adaptive neural circuits [5, 54]. Synaptotagmin-1 is primarily located on synaptic vesicles and contains two independent C2-type Ca^{2+}-sensing domains [9] (termed C2A and C2B, respectively) that are connected by a linker (the fragment containing both domains is designated C2AB). Synaptotagmin 1 interacts with both anionic membranes and SNARE complexes, and both interactions are physiologically relevant [22, 45].

A general model has emerged where inhibitory and activating interactions among synaptotagmin, complexin, and the SNARE complex (that juxtaposes synaptic vesicles and the plasma membrane) yield a stall in membrane fusion that is released by Ca^{2+} influx following an action potential [25, 40]. Understanding the molecular mechanism underlying the release of the stall requires knowledge of the structures and dynamics of the complexes formed by these proteins. However, the structure of the complex between synaptotagmin and SNAREs has remained elusive. Extensive smFRET measurements were performed between a set of 34 fluorophore pairs located at various amino acid positions in the C2AB fragment of synaptotagmin-1 and the cis (post-fusion) state of the neuronal SNARE complex [11] (Fig. 1.2).

The SNARE complex, and the two C2 domains of synaptotagmin-1 were treated as independent rigid bodies (residues 140:262 and 273:418, respectively), while the torsion angles of the linker connecting the two domains (residues 263–272) were simulated in torsion angle space, i.e., with bond lengths and bond angles fixed. The coordinates for the SNARE complex were obtained from the crystal structure of the neuronal SNARE complex (PDB ID 1SFC) [63] and those of the C2 domains of synaptotagmin-1 from the Ca^{2+}-free crystal structure of synaptotagmin-1 (PDB ID 2R83) [24].

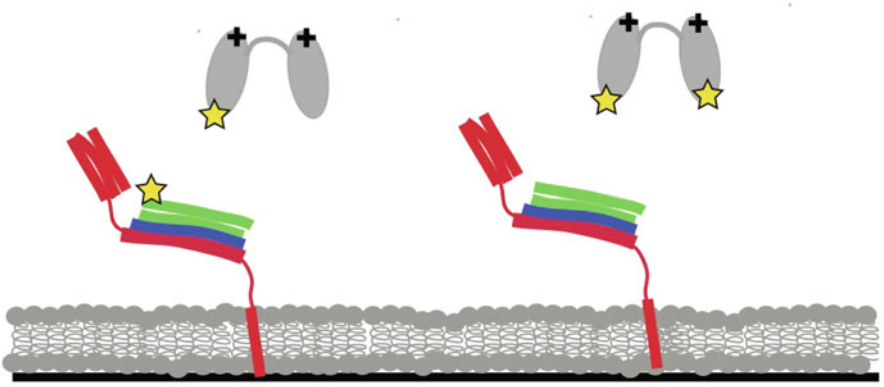


Fig. 1.2 SNARE complex is anchored to a supported bilayer through the transmembrane domain of syntaxin (*red*). The cytoplasmic domain of synaptobrevin (*blue*) and soluble SNAP-25 (*green*) were added and the complex extensively purified prior to membrane reconstitution to obtain mostly parallel complex [68]. After reconstitution into liposomes, a deposited bilayer was formed by liposome condensation. Soluble C2A-C2B fragment (*grey*) of synaptotagmin-1 was added. The "+" signs indicate the calcium binding regions. FRET label pairs were placed in several positions in the C2A, C2B domains and the SNARE complex [11] (*yellow stars*)

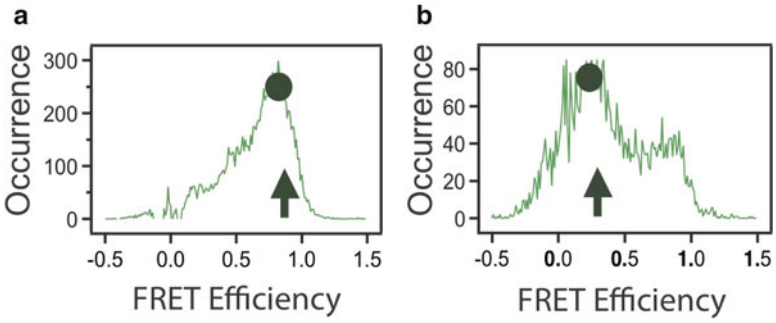


Fig. 1.3 smFRET efficiency histograms of complex between synaptotagmin-1 and SNARE complex with labels placed at position 350 of synaptotagmin-1 (Cy3) and position 61 of synaptobrevin (Cy5) [11]. *Black circles* indicate the FRET efficiency value used to derive the distance restraint for the docking calculations. *Arrows* indicate the FRET efficiency value calculated from best model of the synatotagmin-1:SNARE complex. (**a**) Example of a unimodal FRET efficiency distribution. (**b**) Example of a biomodal FRET efficiency distribution

The FRET measurements used for the synaptotagmin-1:SNARE complex study [11] were acquired with a 0.1 s integration period (referred to as "time bin"). The histograms of FRET efficiency for the different sets of label attachment sites were characterized by either one or two well defined Gaussian peaks whose width was near the shot noise expected width or up to a factor of two wider (Fig. 1.3). The center of a Gaussian function fitted to the histograms was used to represent the

FRET efficiency value for a defined state of the complex. The fraction of area under that Gaussian compared to the area under the complete histogram indicated the fractional population within that FRET efficiency state. In this system, 26 out of 34 label site combinations had dominant states with at least 70% fractional population. These states were used to generate distance restraints for molecular modeling. The remaining label site combinations indicated a mixture of states with none reaching 70% occupancy. For these cases the state was selected using an iterative modeling approach to make use of these mixed-state FRET efficiency distributions. An intermediate model was generated using only the restraints derived from FRET states with >70% in one Gaussian and the mixed states were compared to this preliminary model. The FRET value in the mixed states that was more consistent with the intermediate model was selected and then the entire set of restraints was used for a final round of modeling.

The presence of distinct FRET states for some label pairs along with the fact that the widths of some FRET efficiency peaks were wider than expected from shot noise alone indicate some degree of heterogeneity within the synaptotagmin-1:SNARE complex. For most fluorophore attachment site combinations a dominant configuration could be identified, highlighting one of the advantages of the single molecule approach. On the other hand, this example serves as a warning that in future applications possible heterogeneity within a configurational ensemble may prevent convergence of the modeling calculations to a single solution. Advances in single-molecule imaging technology that allow improved temporal resolution in single molecule FRET studies will allow multiple interconverting states to be better resolved and will extend the applicability of the modeling approaches discussed here.

For 26 of the 34 measured FRET pairs, the major fitted peaks capture 70% or more of the total non-zero smFRET distribution; ten label pairs have distributions with a major peak comprising 90% of the total distribution (Fig. 1.3a), seven label pairs between 90% and 80%, nine between 80% and 70%. The assignments of the dominant FRET states at the 70% level were robust against run-to-run variation. Repeating measurements of single label pair combinations generated the same central FRET efficiency values within experimental error for all label pairs and for most label pairs the dominant population was consistently above the 70% value. For all pairs the dominant population was observed at levels above 70% in at least 66% of the repeated experiments. Many were confirmed greater than 70% dominant in all repeats. Some of the 34 distance restraints involved multiple distinct FRET populations (Fig. 1.3b). Therefore, the docking calculations proceeded in two steps.

For final docking calculations all of the repeated experiments for each label pair were pooled into a single histogram to address the most probable configuration observed across multiple repeated experiments (at least three repeats for every label site pair). For label pairs with FRET efficiency distributions with a dominant peak of >70% of the total area, the FRET measurements were converted to distances and used as restraints for a first round of docking calculations (26 out of 34 pairs were included at this first step). The FRET histograms for the remaining eight label combinations required sums of two Gaussians to fit where neither comprised more

than 70% of the total population. These measured distances were compared to the best model from the first simulation using the first 26 distances and the measured distance that was closer to the model distance was selected. Then the simulation was repeated using all 34 restraints. The resulting models did not change significantly upon inclusion of the eight additional restraints.

A docking simulation using the minor populations from each of the FRET efficiency distributions was also performed (if only one FRET population was present, it was used for both the major and minor population simulation) but convergence was much poorer than for the calculations with the major population. Thus, the conformations arising from docking using the major FRET population restraints are much more likely to occur than those derived using the minor FRET populations.

Uniformly increasing or decreasing all major FRET derived distance restraints by 1 nm led to non-physical results where the proteins were far away from each other or overlapped in space respectively. If the intra-C2 domain restraints were released then the models converged to the same C2B docking state, but the location of the C2A domain became variable.

In order to cross-validate the model, the docking calculations were repeated omitting all of the restraints that involved one particular fluorophore position (residue 383) on synaptotagmin-1. This site showed some of the highest FRET efficiency values when combined with acceptor fluorophores on the SNARE complex. Remarkably, the resulting top model was very similar to the docking calculation using all distances. Because the number of restraints exceeds the number of degrees of freedom for the docking calculations, it is reasonable that omitting a few restraints does not lead to drastically different solutions. The similarity of the top models with and without distances involving the 383 site illustrates the robustness of the top solution with respect to such cross-validation.

The best model from the docking calculations is shown in Fig. 1.4. It showed up consistently in all docking calculations as the top solution. One helix of synaptotagmin-1 (residues 385–395) is directly positioned at the interface with the SNARE complex (Fig. 1.4). The central region of the SNARE complex that mediates synaptotagmin-1 binding (as predicted by the smFRET-derived model) is essential for function. Mutations of glutamates near this area of SNAP-25 (Glu51, Glu52, Glu55) to lysines eliminated in vitro binding of synaptotagmin to the SNARE complex and greatly reduced Ca^{2+} stimulated release in PC12 cells [53]. These same SNAP-25 mutations as well as additional SNAP-25 mutations directly adjacent to this region (Leu50 and Ile171) are critical in the context of docking vesicles in adrenal chromaffin cells [16]. Additionally, in the smFRET-derived model the conserved arginine residues at the bottom of C2B [69] are close to the interface with the SNARE complex but also are sufficiently exposed to allow potential interactions with membranes. Mutation of these residues results in decreased synchronous neurotransmitter release in hippocampal glutamatergic neurons [69]. These independent functional assessments of the synaptotagmin-1 - SNARE interactions lend further credence to the smFRET-derived model.

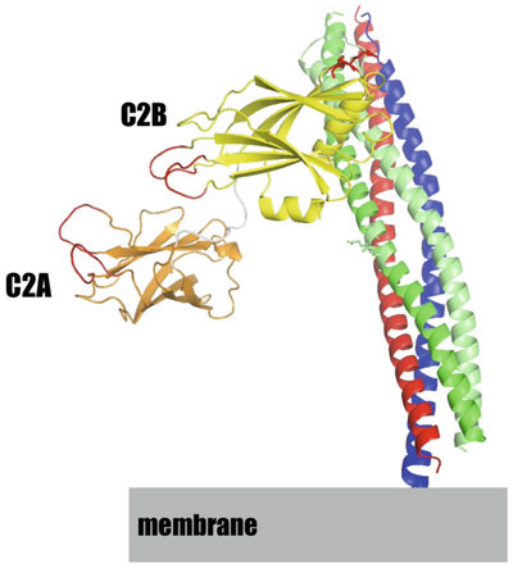


Fig. 1.4 Model of the synaptotagmin-1:SNARE complex obtained from docking calculations using 34 smFRET derived distances [11]. Show are cartoon representations of synaptobrevin (*blue*), syntaxin (*red*), SNAP-25 (*green*), C2A domain of synaptotagmin-1 (*orange*), C2B domain of synaptotagmin-1 (*yellow*) (Color figure online)

The synaptotagmin-1:SNARE complex is not a rigid structure since occasional transitions between different FRET efficiency states were observed, and some of the smFRET efficiency distributions have a multimodal appearance (Fig. 1.3b). This intrinsic flexibility of the synaptotagmin-1:SNARE complex may allow the complex to adapt to the particular geometry of the interacting membranes in the pre-fusion state.

1.9 Application to PSD-95

Multidomain scaffold proteins are critical organizers of signal transduction and junctional communication [46]. Often these proteins contain a series of archetypal protein interaction domains connected by flexible linkers into a larger protein. It is convenient to try and understand scaffold proteins by studying each individual domain as an isolated unit. However, dimerization and incorporation into larger proteins can alter the structure as well as the binding specificity relative to the isolated domain [33, 47, 65, 70]. Thus, studies of isolated domains can be only partially successful at explaining function in their biological context.

PDZ (PSD-95/Dlg/ZO-1) domains are the most common protein-interaction domains in the human genome [3, 29]. PDZ domains typically form part of a larger multidomain protein and often appear in tandem with instances of up to 13 PDZ domains in a single protein [31]. PSD-95 was one of the first PDZ-containing proteins identified [10]. PSD-95 contains three tandem PDZ domains followed by an SH3 domain and an enzymatically inactive guanylate kinase-like domain. The first

two tandem PDZ domains are connected by a five residue linker. Such tandem domain arrays are often conceptualized as folded beads on a disordered string. However, recent structural studies have found that tandem PDZ domains can adopt a fixed interdomain orientation [21, 26, 34, 37–39].

Most studies of PDZ tandems to date have examined portions of much larger proteins so the degree to which the structure of the tandem PDZ supramodules depends on the remainder of the protein is therefore unknown. The crystallization of flexible, multidomain scaffold proteins is challenging. The PDZ tandem from PSD-95 "yielded poorly diffracting crystals" so a self-interacting PDZ ligand sequence appended to the C-terminus that induced non-native protein interactions which lead to diffracting crystals [58]. The resulting crystals contained two different conformations with slightly different orientations of the PDZ domains. The large molecular weight of scaffold proteins often necessitates the use of truncated constructs. The limited interdomain contact provides few unambiguous NMR restraints. In contrast, FRET experiments allow for the accurate measurement of long intramolecular distances (3–8 nm) so it is particularly useful for studying the orientation of domains that are not in intimate contact. In addition, fluorescence measurements are not subject to molecular weight restrictions, which makes measurements in full-length scaffold proteins possible.

Single molecule FRET observations and other data were used to characterize the interdomain orientation of the first two PDZ domains in full-length PSD-95 [42]. Using the high-resolution structures of the first two PDZ domains [37, 48, 64] four surface-exposed sites in PDZ1 and five in PDZ2 were selected for fluorescent labeling and mutated to cysteine (Fig. 1.5a). Accepted labeling sites showed maximal solvent exposure and minimal near neighbor interactions, which should minimize any positional dependence of the photophysical properties. This expectation was confirmed by measuring the anisotropy and quantum yield of the attached dyes, which showed minimal environmental impacts on fluorescence emission [42].

The donor anisotropy was similar for all samples and was minimally higher than that measured for unconjugated Alexa 555. Using the 9 single labeling sites, 11 pairwise combinations were generated in full-length PSD-95 to create 11 double-cysteine mutants for smFRET measurements [42] (Fig. 1.5a). This ensemble of labeling combinations was chosen to sample any possible orientation between these two domains.

All 11 labeling combinations produced smFRET histograms containing a single peak of narrow width that was well described by a single Gaussian curve (Fig. 1.5b). The narrow widths are similar to those observed in duplex DNA and could either arise from a fixed interdomain orientation or time averaging of rapid motions. Paradoxically, these two extremes are indistinguishable in this assay. The different labeling site combinations showed a wide dispersion in their mean smFRET efficiencies (from 0.37 to 0.94), which would not be expected if the two domains were undergoing isotropic motion. To see if domain orientation is dependent on interdomain contacts within the full-length protein, smFRET efficiencies were measured for all 11 labeling site combinations in truncated PSD-95 constructs

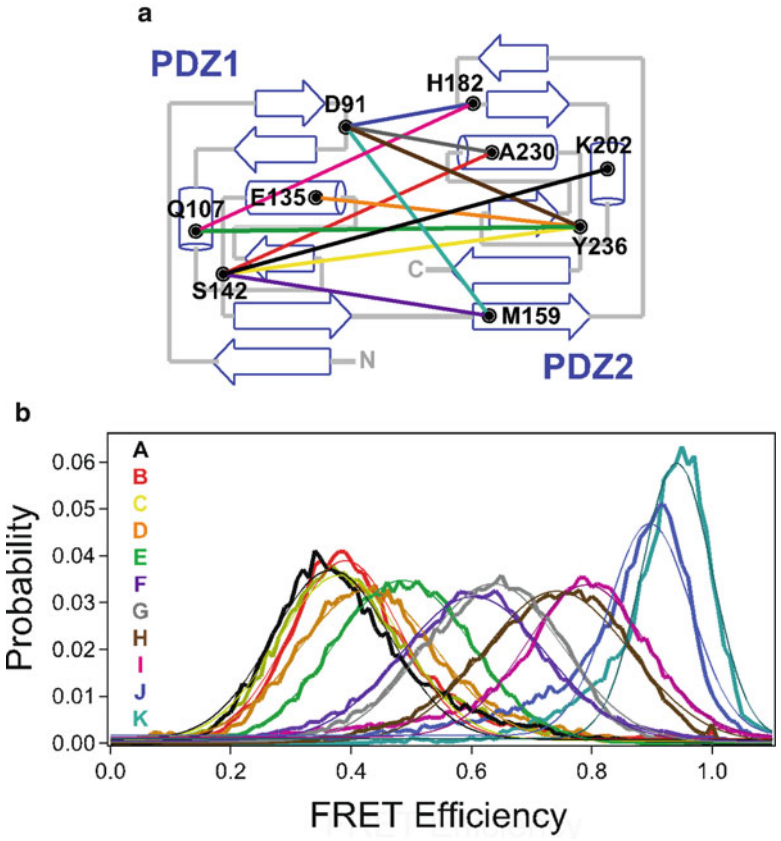


Fig. 1.5 Single molecule FRET measurements between PDZ1 and PDZ2 in full length PSD-95. (**a**) Topology diagram showing the position of the cysteine labeling sites in each domain and the 11 combinations of labeling sites used for fluorescence studies [42]. (**b**) smFRET histograms for all 11 FRET pairs made in full-length PSD-95. Letter codes (indicated within the panel) were assigned to each mutant in order of increasing FRET efficiency. The lettering of the FRET distributions corresponds to the *lettering* denoting labeling combinations in panel A. *Thin lines* indicate the fit to a single Gaussian function (Color figure online)

lacking the 3 additional domains. None of the labeling site combinations showed significant changes in mean FRET efficiency or the width of the FRET distribution relative to full length PSD-95. Thus, domain orientation in the PDZ tandem is not affected by intramolecular interactions with other domains present in full length PSD-95.

Molecular dynamics simulations were performed to calculate the mean dye position relative to the protein backbone for each of the nine labeling sites similar to the studies of the synaptotagmins [11, 66]. The mean FRET efficiencies were converted to distances using a calibrated Förster's radius. The γ factor was determined from photobleaching and applied as individual γ correction at the

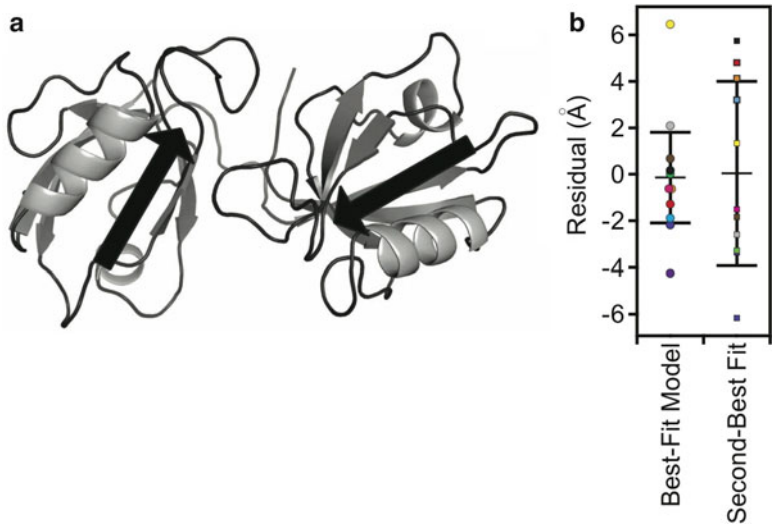


Fig. 1.6 (**a**) Cartoon representation of the best fit model to the 11 FRET distance restraints [42]. The model shows a relatively compact orientation without interdomain contacts. The position and direction of canonical PDZ ligand binding sites are represented as *arrows*. (**b**) Residuals (r.m.s. distance differences) for the first and second best-fit models based on the smFRET distance restraints. Residuals are colored for each labeling site according to Fig. 1.5

single molecule level [41]. Extensive torsion angle/rigid body molecular dynamics simulations were performed with atoms in PDZ1 and PDZ2 fixed while atoms in the interdomain linker were unrestrained. FRET-derived distance restraints were applied to guide domain docking [42].

From 500 trials with randomly assigned starting orientations, cluster analysis identified a single model was found that best satisfied the FRET distance restraints (Fig. 1.6a) [42]. The best-fit model for the PDZ tandem showed a relatively closed configuration but with the distance between domains greater than a single bond. The orientation in the model shows an antiparallel alignment of the two ligand binding pockets (Fig. 1.6a). The root mean square error (E_{RMS}) for the best-fit model was 2.54 Å, while in the second model E_{RMS} was 5.21 Å (Fig. 1.6b). Outlying data points were within loops that were held rigid during the simulation but are expected to fluctuate in the native protein. The second-best fit model was more compact, with the centers of mass for the two domains positioned 7.7 Å closer. The second model had the relative orientation flipped with PDZ2 positioned on the opposite face of PDZ1. To insure that FRET data were not misfit of overfit, docking simulations were repeated with each of the FRET restraints omitted from the refinement. The final results from all 11 simulations were highly similar structures, indicating that FRET restraints oversampled the domain orientation in the PDZ tandem. This represents the first structural model for the PDZ tandem of PSD-95 in the context of the full-length protein.

1.10 Conclusions

Many important biological structures have resisted analysis by high-resolution structural methods. The causes of these difficulties are often multiple and wide-ranging. Some proteins are not stable at the high concentrations required; some molecular systems do not reside in a single stable configuration; other interesting complexes are present only rarely within an ensemble. Single molecule FRET has proven to be an effective tool that can provide a window into these difficult systems. Although the distance resolution provided in any single FRET pair measurement does not approach the atomic dimensions, detailed structural information can be gleaned by oversampling in distance space with combinations of multiple FRET measurements across different locations. Easier access to such oversampling will be provided by the recent development of four-color single molecule FRET measurements of individual complexes [18, 36]. If four distinct fluorescent dyes can be attached at specific locations to the domains of larger multi-molecular complexes, then the FRET coupling between the dyes can simultaneously report 6 interdye distances [36]. Such advances will allow more confident and efficient application of the FRET-restrained docking approaches we have described.

These studies of the synaptotagmin-1:SNARE complex and the conformation of PDZ domains within full-length PSD-95 we have described are examples of particularly effective combinations of single molecule FRET with available high-resolution structures of individual domains. The domains are either connected by flexible linkers or are bound to other stable domains in the final complex. Single molecule FRET measurements restrained docking calculations of the known high-resolution structures of the individual domains and allowed three-dimensional models ranked by distance satisfaction to be obtained. This hybrid strategy will likely be useful to determine the configuration of other biological complexes.

Acknowledgments We thank the National Institutes of Health for support (to A.T.B., RO1-MH63105), and a Career Award at the Scientific Interface from the Burroughs Wellcome Fund (to K.W.). Part of this material has been published in modified form in the Journal of Structural Biology [7].

References

1. Amir D, Haas E (1987) Estimation of intramolecular distance distributions in bovine pancreatic trypsin inhibitor by site-specific labeling and nonradiative excitation energy-transfer measurements. Biochemistry 26:2162–2175
2. Amir D, Haas E (1988) Reduced bovine pancreatic trypsin inhibitor has a compact structure. Biochemistry 27:8889–8893
3. Bhattacharyya RP, Remenyi A, Yeh BJ, Lim WA (2006) Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. Annu Rev Biochem 75:655–680
4. Brunger AT (1992) X-PLOR, version 3.1. A system X-ray crystallograpjy and NMR. Yale University Press, New Haven

5. Brunger AT (2005) Structure and function of SNARE and SNARE-interacting proteins. Q Rev Biophys 38:1–47
6. Brunger AT, Nilges M (1993) Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. Q Rev Biophys 26:49–125
7. Brunger AT, Strop P, Vrljic M, Chu S, Weninger KR (2011) Three-dimensional molecular modeling with single molecule FRET. J Struct Biol 173:497–505
8. Cherny DI, Eperon IC, Bagshaw CR (2009) Probing complexes with single fluorophores: factors contributing to dispersion of FRET in DNA/RNA duplexes. Eur Biophys J 38:395–405
9. Cho W, Stahelin RV (2006) Membrane binding and subcellular targeting of C2 domains. Biochim Biophys Acta 1761:838–849
10. Cho KO, Hunt CA, Kennedy MB (1992) The rat brain postsynaptic density fraction contains a homolog of the drosophila discs-large tumor suppressor protein. Neuron 9:929–942
11. Choi UB, Strop P, Vrljic M, Chu S, Brunger AT, Weninger KR (2010) Single-molecule FRET-derived model of the synaptotagmin 1-SNARE fusion complex. Nat Struct Mol Biol 17:318–324
12. Chung HS, Louis JM, Eaton WA (2009) Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. Proc Natl Acad Sci U S A 106:11837–11844
13. Dahan M, Deniz AA, Ha T, Chemla DS, Schultz PG, Weiss S (1999) Ratiometric measurement and identification of single diffusing molecules. Chem Phys 247:85–106
14. Dave R, Terry DS, Munro JB, Blanchard SC (2009) Mitigating unwanted photophysical processes for improved single-molecule fluorescence imaging. Biophys J 96:2371–2381
15. de Vries SJ, van Dijk AD, Krzeminski M, van Dijk M, Thureau A, Hsu V, Wassenaar T, Bonvin AM (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins 69:726–733
16. de Wit H, Walter AM, Milosevic I, Gulyas-Kovacs A, Riedel D, Sorensen JB, Verhage M (2009) Synaptotagmin-1 docks secretory vesicles to syntaxin-1/SNAP-25 acceptor complexes. Cell 138:935–946
17. Deniz AA, Mukhopadhyay S, Lemke EA (2008) Single-molecule biophysics: at the interface of biology, physics and chemistry. J R Soc Interface 5:15–45
18. DeRocco V, Anderson T, Piehler J, Erie DA, Weninger K (2010) Four-color single-molecule fluorescence with noncovalent dye labeling to monitor dynamic multimolecular complexes. Biotechniques 49:807–816
19. 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
20. Engh RA, Huber R (1991) Accurate bond and angle parameters for X-Ray protein- structure refinement. Acta Crystallogr A 47:392–400
21. Feng W, Shi Y, Li M, Zhang M (2003) Tandem PDZ repeats in glutamate receptor-interacting proteins have a novel mode of PDZ domain-mediated target binding. Nat Struct Biol 10: 972–978
22. Fernandez-Chacon R, Konigstorfer A, Gerber SH, Garcia J, Matos MF, Stevens CF, Brose N, Rizo J, Rosenmund C, Südhof TC (2001) Synaptotagmin I functions as a calcium regulator of release probability. Nature 410:41–49
23. Förster T (1948) Zwischenmolekulare Energiewanderung und Fluoreszenz. Ann Phys 2:55–57
24. Fuson KL, Montes M, Robert JJ, Sutton RB (2007) Structure of human synaptotagmin 1 C2AB in the absence of Ca2+ reveals a novel domain association. Biochemistry 46:13041–13048
25. Giraudo CG, Garcia-Diaz A, Eng WS, Chen Y, Hendrickson WA, Melia TJ, Rothman JE (2009) Alternative zippering as an on-off switch for SNARE-mediated fusion. Science 323:512–516
26. Goult BT, Rapley JD, Dart C, Kitmitto A, Grossmann JG, Leyland ML, Lian LY (2007) Small-angle X-ray scattering and NMR studies of the conformation of the PDZ region of SAP97 and its interactions with Kir2.1. Biochemistry 46:14117–14128
27. Ha T, Ting AY, Liang J, Caldwell WB, Deniz AA, Chemla DS, Schultz PG, Weiss S (1999) Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. Proc Natl Acad Sci USA 96:893–898

28. Ha T, Ting AY, Liang J, Deniz AA, Chemla DS, Schultz PG, Weiss S (1999) Temporal fluctuations of fluorescence resonance energy transfer between two dyes conjugated to a single protein. Chem Phys 247:107–118
29. Harris BZ, Lim WA (2001) Mechanism and role of PDZ domains in signaling complex assembly. J Cell Sci 114:3219–3231
30. Haugland RP (2005) The handbook: A guide to fluorescent probes and labeling technologies. Molecular Probes, Carlsbad
31. Hung AY, Sheng M (2002) PDZ domains: structural modules for protein complex assembly. J Biol Chem 277:5699–5702
32. Iqbal A, Arslan S, Okumus B, Wilson TJ, Giraud G, Norman DG, Ha T, Lilley DM (2008) Orientation dependence in fluorescent energy transfer between Cy3 and Cy5 terminally attached to double-stranded nucleic acids. Proc Natl Acad Sci U S A 105:11176–11181
33. Irie M, Hata Y, Takeuchi M, Ichtchenko K, Toyoda A, Hirao K, Takai Y, Rosahl TW, Sudhof TC (1997) Binding of neuroligins to PSD-95. Science 277:1511–1515
34. Kang BS, Cooper DR, Jelen F, Devedjiev Y, Derewenda U, Dauter Z, Otlewski J, Derewenda ZS (2003) PDZ tandem of human syntenin: crystal structure and functional properties. Structure 11:459–468
35. Lee NK, Kapanidis AN, Wang Y, Michalet X, Mukhopadhyay J, Ebright RH, Weiss S (2005) Accurate FRET measurements within single diffusing biomolecules using alternating-laser excitation. Biophys J 88:2939–2953
36. Lee J, Lee S, Ragunathan K, Joo C, Ha T, Hohng S (2010) Single-molecule four-color FRET. Angew Chem Int Ed Engl 49:9922–9925
37. Long JF, Tochio H, Wang P, Fan JS, Sala C, Niethammer M, Sheng M, Zhang M (2003) Supramodular structure and synergistic target binding of the N-terminal tandem PDZ domains of PSD-95. J Mol Biol 327:203–214
38. Long JF, Feng W, Wang R, Chan LN, Ip FC, Xia J, Ip NY, Zhang M (2005) Autoinhibition of X11/Mint scaffold proteins revealed by the closed conformation of the PDZ tandem. Nat Struct Mol Biol 12:722–728
39. Long J, Wei Z, Feng W, Yu C, Zhao YX, Zhang M (2008) Supramodular nature of GRIP1 revealed by the structure of its PDZ12 tandem in complex with the carboxyl tail of Fras1. J Mol Biol 375:1457–1468
40. Maximov A, Tang J, Yang X, Pang ZP, Sudhof TC (2009) Complexin controls the force transfer from SNARE complexes to membranes in fusion. Science 323:516–521
41. McCann J, Choi UB, Zheng L, Weninger K, Bowen ME (2010) Optimizing methods to recover absolute FRET efficiency from immobilized single molecules. Biophys J 99:961–970
42. McCann J, Zheng L, Chiantia S, Bowen ME (2011) Domain orientation in the tandem PDZ supramodule from PSD-95 is maintained in the full-length protein. Structure 19:810–820
43. Merchant KA, Best RB, Louis JM, Gopich IV, Eaton WA (2007) Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. Proc Natl Acad Sci 104:1528–1533
44. Nir E, Michalet X, Hamadani KM, Laurence TA, Neuhauser D, Kovchegov Y, Weiss S (2006) Shot-noise limited single-molecule FRET histograms: comparison between theory and experiments. J Phys Chem B 110:22103–22124
45. Pang ZP, Shin OH, Meyer AC, Rosenmund C, Sudhof TC (2006) A gain-of-function mutation in synaptotagmin-1 reveals a critical role of Ca2 + −dependent soluble N-ethylmaleimide-sensitive factor attachment protein receptor complex binding in synaptic exocytosis. J Neurosci 26:12556–12565
46. Pawson T, Scott JD (1997) Signaling through scaffold, anchoring, and adaptor proteins. Science 278:2075–2080
47. Peterson FC, Penkert RR, Volkman BF, Prehoda KE (2004) Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition. Mol Cell 13:665–676
48. Piserchio A, Pellegrini M, Mehta S, Blackman SM, Garcia EP, Marshall J, Mierke DF (2002) The PDZ1 domain of SAP90. Characterization of structure and binding. J Biol Chem 277:6967–6973

49. Rasnik I, Myong S, Cheng W, Lohman TM, Ha T (2004) DNA-binding orientation and domain conformation of the E. coli rep helicase monomer bound to a partial duplex junction: single-molecule studies of fluorescently labeled enzymes. J Mol Biol 336:395–408
50. Rasnik I, McKinney SA, Ha T (2006) Nonblinking and long-lasting single-molecule fluorescence imaging. Nat Methods 3:891–893
51. Rhee JS, Li LY, Shin OH, Rah JC, Rizo J, Sudhof TC, Rosenmund C (2005) Augmenting neurotransmitter release by enhancing the apparent Ca2+ affinity of synaptotagmin 1. Proc Natl Acad Sci U S A 102:18664–18669
52. Rice LM, Brunger AT (1994) Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. Proteins 19:277–290
53. Rickman C, Jimenez JL, Graham ME, Archer DA, Soloviev M, Burgoyne RD, Davletov B (2006) Conserved prefusion protein assembly in regulated exocytosis. Mol Biol Cell 17:283–294
54. Rizo J, Rosenmund C (2008) Synaptic vesicle fusion. Nat Struct Mol Biol 15:665–674
55. Rothwell PJ, Berger S, Kensch O, Felekyan S, Antonik M, Wöhrl BM, Restle T, Goody RS, Seidel CAM (2003) Multiparameter single-molecule fluorescence spectroscopy reveals heterogeneity of HIV-1 reverse transcriptase:primer/template complexes. Proc Natl Acad Sci U S A 100:1655–1660
56. Roy R, Kozlov AG, Lohman TM, Ha T (2007) Dynamic structural rearrangements between DNA binding modes of E. coli SSB protein. J Mol Biol 369:1244–1257
57. Roy R, Hohng S, Ha T (2008) A practical guide to single-molecule FRET. Nat Methods 5:507–516
58. Sainlos M, Tigaret C, Poujol C, Olivier NB, Bard L, Breillat C, Thiolon K, Choquet D, Imperiali B (2010) Biomimetic divalent ligands for the acute disruption of synaptic AMPAR stabilization. Nat Chem Biol 7:81–91
59. Sakon JJ, Weninger KR (2010) Detecting the conformation of individual proteins in live cells. Nat Methods 7:203–205
60. Schroder GF, Levitt M, Brunger AT (2010) Super-resolution biomolecular crystallography with low-resolution data. Nature 464:1218–1222
61. Schwieters CD, Clore GM (2001) Internal coordinates for molecular dynamics and minimization in structure determination and refinement. J Magn Reson 152:288–302
62. Stryer L, Haugland RP (1967) Energy transfer: a spectroscopic ruler. Proc Natl Acad Sci USA 58:719–726
63. Sutton RB, Fasshauer D, Jahn R, Brunger AT (1998) Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. Nature 395:347–353
64. Tochio H, Hung F, Li M, Bredt DS, Zhang M (2000) Solution structure and backbone dynamics of the second PDZ domain of postsynaptic density-95. J Mol Biol 295:225–237
65. van den Berk LC, Landi E, Walma T, Vuister GW, Dente L, Hendriks WJ (2007) An allosteric intramolecular PDZ-PDZ interaction modulates PTP-BL PDZ2 binding specificity. Biochemistry 46:13629–13637
66. Vrljic M, Strop P, Ernst JA, Sutton RB, Chu S, Brunger AT (2010) Molecular mechanism of the synaptotagmin-SNARE interaction in Ca2+-triggered vesicle fusion. Nat Struct Mol Biol 17:325–331
67. Watkins LP, Chang H, Yang H (2006) Quantitative single-molecule conformational distributions: a case study with poly-(L-proline). J Phys Chem A 110:5191–5203
68. Weninger K, Bowen ME, Chu S, Brunger AT (2003) Single-molecule studies of SNARE complex assembly reveal parallel and antiparallel configurations. Proc Natl Acad Sci USA 100:14800–14805
69. Xue M, Ma C, Craig TK, Rosenmund C, Rizo J (2008) The Janus-faced nature of the C(2)B domain is fundamental for synaptotagmin-1 function. Nat Struct Mol Biol 15:1160–1168
70. Zhang Q, Fan JS, Zhang M (2001) Interdomain chaperoning between PSD-95, Dlg, and Zo-1 (PDZ) domains of glutamate receptor-interacting proteins. J Biol Chem 276:43216–43220

Chapter 2
System-Specific Scoring Functions: Application to Guanine-Containing Ligands and Thrombin

Ivan V. Ozerov, Elisabeth D. Balitskaya, and Roman G. Efremov

Abstract Molecular docking is one of the most common and popular computational methods in structural biology. It is widely used for investigations of molecular details of protein functioning and in drug design. Nevertheless, modern docking algorithms are still far from perfection. Development of scoring functions aimed at prediction of spatial structure and free energy of binding for molecular complexes remains a challenging task. With increasing amount of structural data, creation of precise system-specific scoring functions becomes possible. This article describes the physical phenomena underlying efficiency of such scoring functions and demonstrates the related quantitative approaches by the examples of guanine-containing ligands and thrombin.

2.1 Introduction

Molecular docking is a method of molecular modeling aimed at prediction of the spatial structure of a protein-ligand complex and the free energy of ligand binding. Docking procedure can be divided into two separate but interconnected parts: conformational search and scoring function.

The conformational search algorithm starts from a random orientation of two molecules and leads to final 3D structure of the biomolecular complex through a

I.V. Ozerov (✉) • E.D. Balitskaya • R.G. Efremov
Laboratory of Biomolecular Modeling, Russian Academy of Sciences, M.M. Shemyakin & Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Ul. Miklukho-Maklaya, 16/10, 117997 GSP, Moscow V-437, Russia

I.V. Ozerov • E.D. Balitskaya
Department of Bioengineering, M.V. Lomonosov Moscow State University, Biological faculty, Leninskie Gori, 1/73, 119991 GSP-1, Moscow, Russia
e-mail: varnivey@nmr.ru

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_2,

sequence of stepwise conformational changes. Generally, these are internal rotations around chemical bonds in molecules and translation/rotation of a rather small molecule (ligand) around another larger one (receptor) as a rigid body. In some cases, receptor flexibility is very important and so there are several methods to take it into account. Detailed information about the conformational protein flexibility problem one can find in reviews [1, 2].

Scoring function is used to score (or rank) various putative ligand conformations placed in the receptor binding site at each step of the conformational search procedure according to some predefined criteria. Sometimes, scoring functions are used to predict the free energy of binding. There are three types of scoring functions: forcefield based, stochastic and empirical [3]. Empirical functions are the most commonly used. They represent a natural choice when a proper training set could be constructed. Such scoring functions are often defined as a linear combination of terms with weighting coefficients, which can be derived using multiple linear regression methods during validation on the training set. Usually, the terms represent different types of intermolecular interactions. Correct description and/or weighting of such interactions are among the most important problems related to selection of acceptable criteria to investigate specific protein-ligand complexes.

The major part of molecular docking programs includes different universal scoring functions, which were validated for a wide range of molecular targets. Nevertheless, it is reasonably to assume that the more efficient way is to use functions validated for a given class of systems if the structural data available for them is enough to construct a representative training set. In this case, the information about specific interactions typical for these particular protein-ligand complexes can be considered. The first variant is to use experimental data on binding of different ligands to a certain target protein. The target-specific score is applied to search for other ligands from the libraries of low-molecular-weight compounds to find those that are capable of binding the target with greater affinity. Another approach is to design some ligand-specific criteria. This procedure is less wide-spread since it requires a considerable amount of experimental data on interactions of a particular class of ligands with proteins. However, there have been presented successful studies aimed at development of such scores for peptides [4], saccharides [5], ATP [6], and so forth.

Often, the same universal function is used for ranking of the docking poses and prediction of the free energy of binding. This approach does not seem reasonable when applied to system-specific criteria. The functions based on a set of correct structures could give insufficiently overrated estimation of the free energy for incorrect structure. Besides, certain interactions crucial for determination of the spatial structure of a complex could be useless for estimation of the binding energy. For example, the information about a pattern of hydrogen bonds which is found in all training structures is extremely important for ranking of docking solutions. At the same time, it is not helpful to construct an empirical function based on analysis of the structural differences between complexes in the training set and aimed at prediction of the free energy of ligand binding.

2.2 Validation of Empirical Scoring Functions

There are several ways to evaluate the accuracy of a given predictive model and docking solutions retrieved by it. The most common way is to calculate RMSD (root-mean-square deviation) of a particular docking pose from known crystal structure and to plot enrichment curves for every model. As a rule, the docking solution is considered correct if its RMSD from the crystal structure is less than 2.5 Å. Enrichment factor of the scoring function is the fraction of correct docking poses among several best ranked solutions. The enrichment curve graphically demonstrates percentage of the correct poses found in a top-ranked subset of docking solutions. These quantities were described in more details elsewhere (see, e.g. [3]).

In this work, we focus on the case when the amount of structural data is not enough to construct a full-blown test set. In such a situation, the "leave-one-out" cross validation is normally used to validate a scoring function. Members of the training set are consecutively excluded, and on each step, the excluded member is used to assess the predictive power of the model. Leave-one-out cross validation leads to a cross-validated correlation coefficient q^2 and indicator of predictive ability of the model (s_{press}):

$$q^2 = \frac{\sum (y_{pred} - y_{obs})^2}{\sum (y_{obs} - y_{mean})^2} \tag{2.1}$$

$$s_{press} = \sqrt{\sum (y_{pred} - y_{obs})^2 / (N - k - 1)}, \tag{2.2}$$

where y_{pred} is the predicted value, y_{obs} is the experimentally observed value, y_{mean} is the mean value for all complexes, N is the number of complexes, and k is the number of terms. Sometimes, more rigorous test like "leave-five-out" cross validation can be used. If the value of q^2 is higher than 0.3, then an approximate correlation is present.

2.3 Scoring Functions

Here, based on the analysis of thrombin complexes with various ligands and nucleotide-protein complexes, two corresponding system-specific scoring functions were designed to predict the spatial structure and the free energy of binding for these systems. Both functions represent a linear combination of the terms describing contributions of various ligand-receptor interactions. In a general form, they can be written as follows:

$$\Delta G_{bind} = K + \alpha(HB) + \beta(LIPO) + \gamma(BP) + \delta(ROT), \tag{2.3}$$

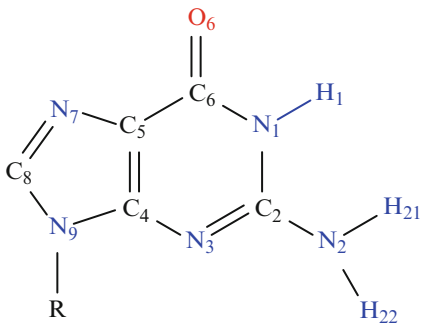


Fig. 2.1 Chemical formula of guanine. Standard atom numbering is used

where ΔG_{bind} is the free energy of binding in thrombin-ligand complexes; K is the constant; *HB* is the term which takes into account the energy of hydrogen bonds; *LIPO* is the term considering the energy of hydrophobic ligand – protein contact; *BP* is the term describing the energy of hydrophobic ligand – hydrophilic protein and hydrophilic ligand – hydrophobic protein contacts; *ROT* is the rotational term considering the entropy loss that occurs when single, acyclic bonds in the ligand become non-rotatable upon binding. Greek symbols indicate the weighting coefficients.

$$\begin{aligned} Thrombinscore = {} & K + \alpha_1(HB\_S1) + \alpha_2(LIPO\_S1) + \alpha_3(BP\_S1) + \beta(HB\_S2) \\ & + \gamma_1(HB\_S3) + \gamma_2(LIPO\_S3) + \delta(BP\_P4), \end{aligned} \tag{2.4}$$

where *thrombinscore* is the estimation of thrombin-ligand complex spatial structure; K is the constant; *HB_S1, HB_S2, HB_S3* are the terms related to hydrogen bonds in the pockets S1, S2, S3 of thrombin, respectively; *LIPO_S1, LIPO_S3* are the terms describing hydrophobic ligand–protein contacts in the pockets S1, S3 respectively; *BP_S1*, *BP_P4* are the terms of hydrophobic/hydrophilic ligand–protein (and *vice versa*) contacts in the pockets S1, P4, respectively. Greek symbols are the weighting coefficients.

$$\begin{aligned} Guascore = {} & K + \alpha_1(HB\_N2) + \alpha_2(HB\_O6) + \alpha_3(HB\_PO4) \\ & + \beta(STACKAR) + \gamma(LIPO), \end{aligned} \tag{2.5}$$

where *guascore* is the estimation (score) of a guanine-containing ligand-protein spatial structure; K is the constant; *HB_N2, HB_O6* are the terms which take into account hydrogen bonds between guanine N2, O6 atoms and protein (Fig. 2.1); *HB_PO4* is the term related to phosphate-protein hydrogen bonds; *STACKAR* characterizes stacking of two aromatic rings; *LIPO* is the term describing hydrophobic ligand–protein contact. Greek symbols are the weighting coefficients.

As follows from the Eqs. 2.1, 2.2, and 2.3, different intermolecular interactions can be used to formulate acceptable criteria for investigation of particular protein-ligand complexes [7]. Here, special attention is given to some of these interactions.

2.4 Hydrogen Bonds

Hydogen bonds (h-bonds) can be identified and included into the empirical scoring function using simple geometrical criteria – the distance between donor and acceptor r_{AD} and the angle between them and the hydrogen atom α_{ADH} (Fig. 2.2).

These parameters are included in the formula for the h-bonding term:

$$B = B_r(r_{AD}) \times B_\alpha(\alpha_{ADH}), \tag{2.6}$$

$$B_r(r_{AD}) = \begin{cases} 1.0; & r_{AD} \le r_1 \\ \dfrac{r_0 - r_{AD}}{r_0 - r_1}; & r_1 < r_{AD} < r_0, \\ 0.0; & r_0 \le r_{AD} \end{cases} \tag{2.7}$$

where r_1 и r_0 are the boundary values of r_{AD} for the existence ($B_r = 1$) and the absence ($B_r = 0$) of hydrogen bonds (Fig. 2.3).

$$B_\alpha(\alpha_{ADH}) = \begin{cases} 1.0; & \alpha \le \alpha_1 \\ \dfrac{\alpha_0 - \alpha}{\alpha_0 - \alpha_1}; & \alpha_1 < \alpha < \alpha_0, \\ 0.0; & \alpha_0 \le \alpha \end{cases} \tag{2.8}$$

where α_1 и α_0 are the boundary values of α_{AD} for the existence ($B_\alpha = 1$) and the absence ($B_\alpha = 0$) of hydrogen bonds (Fig. 2.3).

Different values of r_1, r_0, α_1, α_0 in the above criteria can be selected for a specific protein-ligand complex. For example, in thrombin-containing complexes, such parameters were taken rather close to those observed in the corresponding crystal structures. On the contrary, detailed inspection of the nucleotide-containing

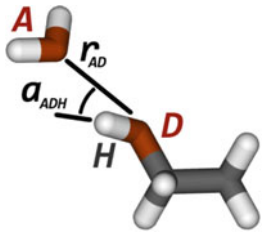


Fig. 2.2 Geometrical parameters of a hydrogen bond that are used in scoring functions: the distance between the donor and the acceptor r_{AD}; and the angle between them and the hydrogen atom α_{ADH}

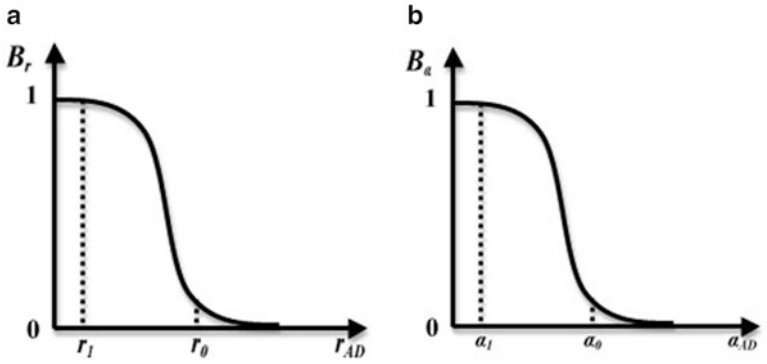


Fig. 2.3 The plots illustrating dependence of the values B_r and B_α on the parameters r_{AD} (**a**) and α_{AD} (**b**), respectively; r_1 and r_0, α_1 and α_0 are the corresponding boundary values of r_{AD} and α_{AD} determining the presence of hydrogen bonds

Table 2.1 Hydrogen bond parameters in thrombin-ligand and guanine-protein complexes

Parameter used in criteria	Mean value in thrombin-ligand complexes	Mean value in guanine-protein complexes
r_1, Å	3.2	3.4
r_0, Å	3.4	3.6
α_1	20°	40°
α_0	40°	60°

$\boldsymbol{r_1}$,$\boldsymbol{r_0}$ are the minimal and maximal values of the distance r_{AD};
$\boldsymbol{\alpha_1}$,$\boldsymbol{\alpha_0}$ are the minimal and maximal values of the angle α_{ADH}

complexes revealed that the best predictive power of the model (scoring function) is achieved when these parameters are distributed in a wider interval as comparing with the experimental data (Table 2.1).

Binding of nucleotide-containing ligands, which include a nitrogen base (ATP, ADP, GTP, NAD, FAD, and others) can be characterized by several patterns of hydrogen bond – different combinations of h-bonds between nucleotide atoms and surrounding protein groups. They resemble h-bonding network in nucleotide pairs in DNA duplex (Fig. 2.4).

In scoring function, this term represents a discrete function, which takes on a value 1 if the pattern exists and 0 otherwise.

2.5 Stacking

Aromatic stacking is another type of intermolecular contacts which plays a distinctive role in biological recognition. Since most of drug-like compounds contain aromatic fragments, stacking is often observed in receptor-ligand complexes. Nevertheless, it is not always paid proper attention in scoring functions.

Fig. 2.4 Examples of h-bonding patterns in complexes of GTP with molybdopterin-guanine dinucleotide biosynthesis protein (PDB code 1FRW), adenylosuccinate synthetase (PDB code 1LOO) and xanthine-guanine phosphoribosyltransferase (PDB code 1A97)

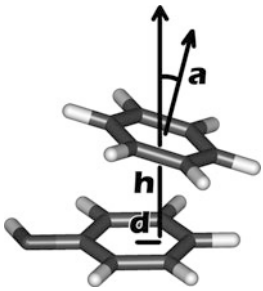


Fig. 2.5 Scheme of geometrical parameters used to describe a stacking contact between two aromatic rings. Displacement (d) and height (h) are calculated for the center of one aromatic ring relative to another ring's plane. α is the angle between the normal vectors of both rings

Stacking interaction occurs between two π-electron systems and results in their specific arrangement in space. The most typical example is DNA, where nitrogen bases form parallel stacking contacts [1, 8, 9]. Other variants are also possible – the so-called T-shaped arrangement is observed for such compounds as benzene [10]. Furthermore, aromatic substances tend to form cation-pi contacts where, positively charged groups interact with aromatic electron clouds [11–13].

Like hydrogen bonds, aromatic contacts can be described by geometrical criteria. To identify a stacking contact, we analyzed the mutual orientation of two aromatic fragments in terms of the height h and displacement d of one cycle with respect to the other, and the angle α between them (Fig. 2.5).

To describe the interaction of two aromatic fragments (including flat guanidine group), we proposed to use the following criterion:

$$S = S_d(d) \times S_h(h) \times S_\alpha(\alpha), \tag{2.9}$$

where d and h are, respectively, displacement and height of the center of one aromatic ring relative to the other, and α is the angle between their planes (Fig. 2.5); the weighting functions in the product are of the following form:

$$S_\alpha(\alpha) = \max(\cos^2\alpha; \sin^2\alpha), \tag{2.10}$$

where the particular form of $S_\alpha(\alpha)$ defines whether parallel ($\cos^2\alpha \geq 0.5$) or "T-shaped" edge-to-face ($\sin^2\alpha > 0.5$) stacking is observed;

$$S_h(h) = \begin{cases} 1.0; & h \leq h_1 \\ \dfrac{h_0 - h}{h_0 - h_1}; & h_1 < h < h_0, \\ 0.0; & h_0 \leq h \end{cases} \tag{2.11}$$

where $h_1 = 4.0$ Å and $h_0 = 5.0$ Å for parallel stacking; $h_1 = 5.0$ Å and $h_0 = 6.0$ Å for edge-to-face stacking;

$$S_d(d) = \begin{cases} 1.0; & d \leq d_1 \\ \dfrac{d_0 - d}{d_0 - d_1}; & d_1 < d < d_0, \\ 0.0; & d_0 \leq d \end{cases} \tag{2.12}$$

where $d_1 = 2.0$ Å and $d_0 = 3.0$ Å for any type of stacking arrangement. Aforementioned stacking parameters were derived from the statistical analysis of contacts of nucleobases with aromatic amino acids in proteins [6].

2.6 Hydrophobic Interactions

Hydrophobic interactions play an important role in protein-ligand recognition. In many cases, complementarity of hydrophobic properties is crucial for ligand binding [14–16].

Quanitative estimation of hydrophobic interactions was made in the framework of the Molecular Hydrophobicity Potential (MHP) formalism. The hydrophobic properties were calculated on molecular surfaces or in any arbitrary points of space using atomic hydrophobicity constants and a distance-dependent decay function (Fig. 2.6):

$$MHP_j = \sum_i f_i g(r_{ij}), \tag{2.13}$$

where MHP_j is the value of MHP in the point of space j; f_i is the hydrophobicity constant of atom i; $g(r_{ij})$ is the distance-decay function. Since there is no exact expression for the "hydrophobic field", all such functions are of intuitive and empirical nature. Often, the exponential function $g(r) = e^{-r}$ [17] is used. Calculations of hydrophobic surface areas were performed with the PLATINUM (Protein Ligand ATtractions Investigation NUMerically) software [18] available on the website http://model.nmr.ru/platinum.

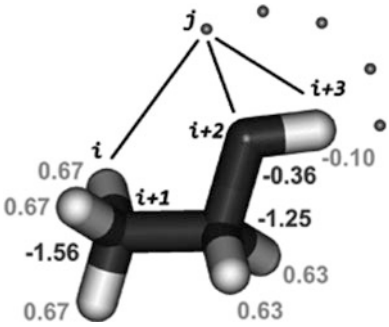


Fig. 2.6 The concept of MHP (Molecular Hydrophobicity Potential) is based on hydrophobicity constants (shown with numbers) of atoms *i*, which contribute to the MHP value in an arbitrary point of space *j*

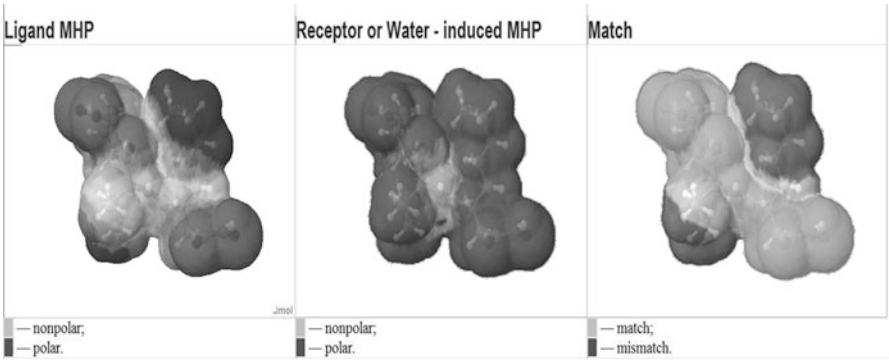


Fig. 2.7 Complementarity of hydrophobic properties of ligand and receptor. MHP induced by atoms of the ligand is calculated on the ligand surface (Ligand MHP). MHP induced by the receptor and water environment is calculated on the same surface (Receptor or water-induced MHP). Difference between the corresponding MHP values calculated in each surface point gives the hydrophobic match/mismatch property (Match)

MHP values on the surfaces of ligand and receptor were calculated separately. The areas of these surfaces, where MHP match or mismatch was observed, served as a criterion of hydrophobic interactions between ligand and receptor (Fig. 2.7). Usually, there are no coordinates of water molecules in molecule structure files. On the other hand, solvation in water may seriously affect intermolecular interactions. To take this phenomenon into account, the implicit water grid with a step of 2 Å and nodes hydrophobicity of −0,38 was suggested to mimic water environment.

Different expressions of the term describing hydrophobic interactions can include absolute or relative surface areas. For example, the following expression of the term includes relative areas.

$$MHP_{match} = \frac{2 \times LL}{LL + LH + HL}, \tag{2.14}$$

where LL, LH and HL are the areas of hydrophobic/hydrophobic, hydrophobic/hydrophilic and hydrophilic/hydrophobic contact of ligand and receptor, respectively.

In the case of thrombin, ligands vary in size significantly. In order to avoid artifacts, the areas of hydrophobic match and mismatch were considered as two independent terms. Thereby, the final expressions for these terms were:

$$LIPO = LL$$
$$BP = LH + HL, \tag{2.15}$$

Here the term LIPO describes lipophilic contacts when the term BP (buried polar) penalizes undesirable contacts between lipophilic and polar atoms of ligand and receptor. It is worth mentioning that these terms implicitly take into account van der Waals and electrostatic interactions, along with the entropic effects.

Besides, we have found that a moderate shift to the more hydrophobic range (+0.04 atomic *logP* units) improves the distribution of MHP-properties for low-molecular-weight compounds, particularly nucleosides and nucleotides. Meanwhile, such a moderate shift does not alter significantly the distribution of protein MHP.

2.7 Entropy Loss Upon Binding

Modern scoring functions (especially those designed for the free energy calculations) include a rotational term describing freezing of ligand rotational bonds caused by binding to the receptor. Here, we examined different types of this term. The simplest way to consider rotatable bonds freezing is to count up the number of such bonds (N_{rot}) in a ligand [19].

$$ROT = N_{rot} \tag{2.16}$$

Hereinafter, rotatable bond is referred to as a bond between sp3-sp3 or sp3-sp2 atoms, excluding bonds in cycles and bonds with terminal groups.

More precise approach is to take into account the freezing of each rotatable bond separately. In this case, one can mark out the bonds which were frozen during ligand binding. The bond is marked as frozen if the distance between atoms forming this bond and receptor atoms is less than a predefined cutoff value. The corresponding term can be defined as follows:

$$ROT = \left(\frac{N_{rot} - N_{froz}}{2}\right) + N_{froz}, \tag{2.17}$$

In the present thrombin-specific function, the expression proposed by Eldridge [20] was used:

$$ROT = 1 + (1 - 1/N_{rot}) \sum_{r} (P_p(r) + P_p^{'}(r))/2, \tag{2.18}$$

where the summation is carried out over frozen bonds only, $P_p(r)$ and $P'_p(r)$ – are the fractions of non-lipophilic heavy atoms on the each side of a particular frozen bond (ranging from 0 to 1). This kind of term is essential to penalize less the large hydrophobic areas of ligands, which are also almost frozen in unbound state in water. The cutoff radius was set to 3.5 Å.

2.8 Ligand Embedding

If the active site is located on the protein surface, it is useful to estimate docking poses by the value of embedding the ligand in the protein. In PLATINUM, it is represented by the sum of all kind of hydrophobic/hydrophilic interactions between the ligand and the protein. For example, it was found that if the docking pose of a guanine-containing ligand is embedded into the protein by less than 70 $Å^2$ of the surface area, this pose reliably was considered wrong and the value of the score was zero.

2.9 Pocket-Targeted Terms

Sometimes, active site of the receptor can be subdivided into several parts. Not all possible interactions may be important for binding to each part of the site. For example, there are three well-delimited pockets in the active site of thrombin. The first pocket (S1) contains an aspartic acid residue forming electrostatic contacts with positively charged ligand groups. In the second pocket (S2), ligands form antiparallel β-sheet-like structures. So, hydrogen bonding is the most important interaction type in this pocket. The third pocket (S3) is formed by hydrophobic residues. It specifically adopts apolar groups of the ligands (Fig. 2.8).

Thereby, sometimes it is better to sum up particular interactions by specific pockets. Such terms are called "pocket-targeted". Their usage can significantly improve the predictive power of the scoring function. However, this approach leads to increasing of the number of terms. So, it is not applicable in the case of very small training set.

2.10 Scoring Function for Guanine-Protein Complexes

Development of the scoring function *guascore* was based on the regression analysis (that fits the weighting coefficients so that the final equation reproduces experimental data such as the measured affinity or geometry of the receptor-ligand complex) by using of 130 guanine-protein complexes. They were divided into

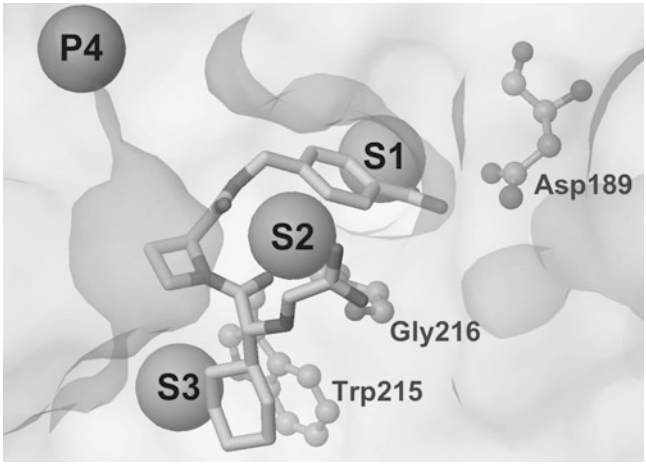


Fig. 2.8 Thrombin active site with inhibitor melagatran (PDB code 1K22). Pocket centers are shown as *spheres*

training (97 complexes) and test (33 complexes) sets. Adjustment of the weighting coefficients was based on the training set and the final scoring function was approved *via* calculations with the test set.

The *guascore* function has the following form (the exact meaning of the terms was described in the "Scoring functions" section):

$$\begin{aligned} Guascore = &-(0.28 \pm 0.06) + (0.12 \pm 0.03) \times (HB\_N2) + (0.15 \pm 0.03) \\ &\times (HB\_O6) + (0.019 \pm 0.007) \times (HB\_PO4) + (0.09 \pm 0.04) \\ &\times (STACKAR) + (0.0028 \pm 0.0009) \times (LIPO) \end{aligned} \tag{2.19}$$

In addition, two filters were proposed to estimate whether the docking pose is correct or not. The first one is h-bond pattern filter. If the docking pose included the patterns formed by three and more guanine atoms, it was considered absolutely correct. The second filter takes into account embedding of the ligand to the protein. If the contact receptor/ligand surface area was less than 70 $Å^2$, this pose was considered as incorrect.

Efficiency of the scoring function was estimated using three criteria: enrichment factor, average best rank of the correct solution and the number of complexes where the correct pose is on the top. The average best rank of the correct solution represents a ratio between the sum of the minimal rank values of correct poses and the total number of complexes.

Guascore was compared with the following three scoring functions – goldscore [21], ASP score [22], and chemscore (Table 2.2) [20]. Furthermore, the average best rank of the correct solution obtained with *guascore* was also compared with the average best rank obtained from a random distribution used instead of a score.

Table 2.2 The enrichment factor (for details see the section "Validation of empirical scoring functions" section), the average best rank of the correct solution, and the number of complexes where the correct pose was on the top for different scoring functions and for a random distribution used instead of a scoring function.

	Training set (97 complexes)			Test set (30 complexes)		
Scoring function	Enrichment factor	Average best rank of the correct solution	Number of complexes where the right pose is on the top	Enrichment factor	Average best rank of the correct solution	Number of complexes where the right pose is on the top
Guascore	0.92	3.4	82 (84%)	0.96	1.2	27 (90%)
Aspscore	0.89	4.3	80 (82%)	0.87	6.0	23 (77%)
Goldscore	0.84	6.1	73 (75%)	0.78	13.8	19 (63%)
Chemscore	0.68	11.6	40 (41%)	0.65	17.9	15 (50%)
Random distribution	0.50	3.7	48 (49%)	0.50	8.4	16 (53%)

The percent of correctly ranked docking poses is shown in brackets

As seen in Table 2.2, guascore is apparently the best scoring function according to all three criteria considering the training and the test sets.

2.11 Thrombin Specific Scoring Functions

Two different thrombin specific scoring functions (TSSF-1 and TSSF-2) were used for direct thrombin inhibitors screening. Twenty eight structures of thrombin-ligand complexes with known affinities were taken from the PDB Bind database [23].

List of the docking poses obtained with five well-known docking algorithms [20, 24, 25] was re-ranked with the TSSF-1 function. The function consisted of seven pocket-targeted terms (Table 2.3). P4 pocket is the place where binding almost doesn't occur. So, the last term BP_P4 doesn't really explain any kind of interactions, though it penalizes any ligand pose which comes into the wrong place. The function ranks docking poses significantly better in comparison with other popular functions examined (Fig. 2.9).

The TSSF-2 function was used to calculate free energies of binding for the best ranked poses obtained with TSSF-1. The mathematical expression for this function is:

$$\Delta G_{bind} = -12.9 - 2.1 \times HB - 25.9 \times LIPO - 24.0 \times BP - 9.5 \times ROT \quad (2.20)$$

After the "leave-one-out" cross validation performed for 28 complex structures, the following values of cross-validated correlation and standard error of prediction were obtained:

$$q^2 = 0.88$$
$$s_{press} = 4.32 \text{kJ/mol}$$

Table 2.3 Terms in the resulting thrombin specific scoring function designed for prediction of the spatial structure of thrombin-ligand complexes (TSSF-1)

Term	Description	Weighting coefficient
HB_S1	Hydrogen bonds in S1 pocket	0.129
LIPO_S1	MHP match in S1 pocket	−2.907
BP_S1	MHP mismatch in S1 pocket	1.334
HB_S2	Hydrogen bonds in S2 pocket	0.068
HB_S3	Hydrogen bonds in S3 pocket	−0.111
LIPO_S3	MHP match in S3 pocket	0.810
BP_P4	MHP mismatch in P4 pocket	−1.592
K	Constant	0.011

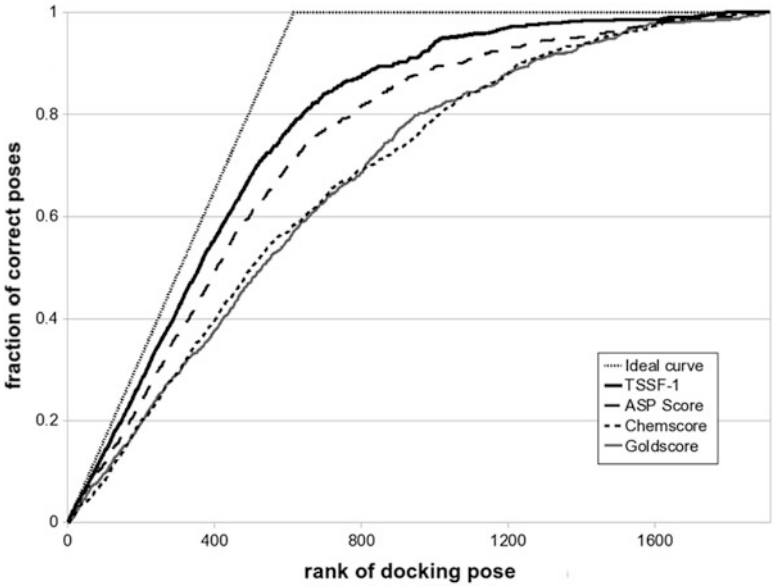


Fig. 2.9 Enrichment *curves* for popular scoring functions and TSSF-1. 1913 docking poses of thrombin inhibitors were re-ranked using different scoring functions. For each function, points of the chart show the part of correct docking poses (RMSD from the crystal structure is less than 2.5 Å) among the particular number of the best ranked solutions

On Fig. 2.10, the predicted free energy is plotted *versus* the experimental data. At the same time, none of the universal functions doesn't show any visible correlation. It is remarkable that the highest R^2 value obtained using the later ones is less than 0.1.

The results obtained clearly show that combination of two thrombin-specific scoring functions is a powerful tool. It allows correct assessment of the spatial structure and the free energies of binding. Therefore, this approach can be further used for rational design of new thrombin inhibitors.

2.12 Conclusions

The results demonstrated by the two system-specific scoring functions (one – for target, one – for ligand) look quite promising from the point of view of their future pharmacological applications. In both cases, the predictive power was better as compared with standard general-purpose evaluation models. At the same time, we should outline that the success rate of such a methodology may strongly depend on the two major factors. Firstly, it is determined by the specificity of protein-ligand interactions in a given type of systems. Thus, it may happen that for some classes of ligands and/or receptors such interactions are not well-defined, and binding

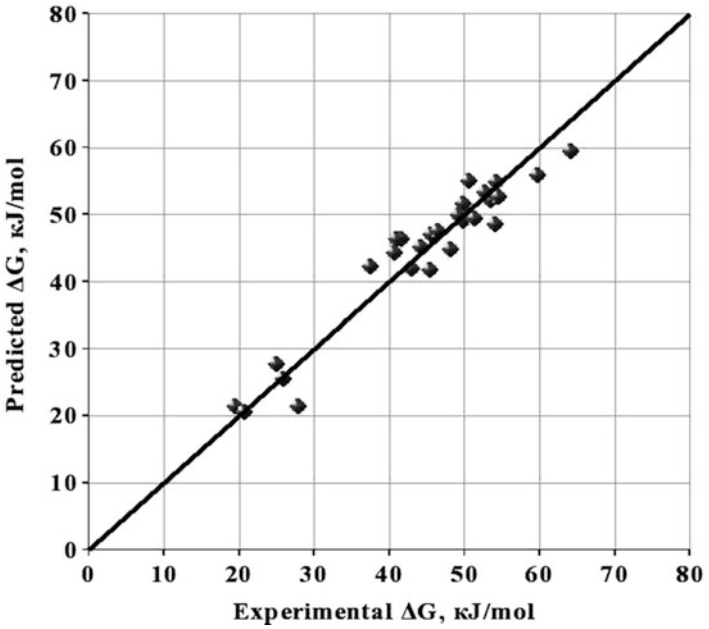


Fig. 2.10 Predicted versus experimental values of the free energy of binding for thrombin-ligand complexes

is determined by a subtle balance of different terms. In such a situation, minor conformational, solvation, *etc*. changes can dramatically alter the results. Most probably, the approach described here will be not useful in these cases. Secondly, development of each new system-specific model requires individual consideration and a representative set of experimental data for training and validation.

Acknowledgments This work was supported by the Russian Foundation for Basic Research and by the RAS Programmes (MCB and "Basic fundamental research for nanotechnologies and nanomaterials"). Access to computational facilities of the Joint Supercomputer Center RAS (Moscow) and Computer Center of M.V. Lomonosov Moscow State University is gratefully acknowledged.

References

1. Zhong H, Tran LM, Stang JL (2009) Induced-fit docking studies of the active and inactive states of protein tyrosine kinases. J Mol Graph Model 28:558–575
2. Kokh DB, Wenzel W (2008) Flexible side chain models improve enrichment rates in in silico screening. J Med Chem 51:5919–5931
3. Pyrkov TV, Chugunov AO, Krylov NA, Nolde DE, Efremov RG (2009) Complementarity of hydrophobic/hydrophilic properties in protein-ligand complexes. In: Joseph D (ed) A new tool to improve docking results, Biophysics and the challenges of emerging threats, Puglisi, pp 21–41

4. Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V (1999) Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. J Med Chem 42:4650–4658
5. Laederach A, Reilly PJ (2005) Modeling protein recognition of carbohydrates. Proteins 60: 591–597
6. Pyrkov TV, Kosinsky YA, Arseniev AS, Priestle JP, Jacoby E, Efremov RG (2007) Complementarity of hydrophobic properties in ATP-protein binding: a new criterion to rank docking solutions. Proteins 66:388–398
7. Pyrkov TV, Ozerov IV, Blitskaia ED, Efremov RG (2010) Molecular docking: role of intermolecular contacts in formation of complexes of proteins with nucleotides and peptides. Bioorg Khim 36:482–492
8. Gotoh O (1983) Prediction of melting profiles and local helix stability for sequenced DNA. Adv Biophys 16:1–52
9. Sponer J, Leszczynski J, Hobza P (1996) Hydrogen bonding and stacking of DNA bases: a review of quantum-chemical ab initio studies. J Biomol Struct Dynam 14:117–135
10. Jorgensen WL, Severance DL (1990) Aromatic-aromatic interactions: free energy profiles for the benzene dimer in water, chloroform, and liquid benzene. J Am Chem Soc 112:4768–4774
11. Waters ML (2002) Aromatic interactions in model systems. Curr Opin Chem Biol 6:736–741
12. Meyer EA, Castellano RK, Diederich F (2003) Interactions with aromatic rings in chemical and biological recognition. Angew Chem Int Ed 42:1210–1250
13. Tewari AK, Dubey R (2008) Emerging trends in molecular recognition: utility of weak aromatic interactions. Bioorg Med Chem 16:126–143
14. Efremov RG, Chugunov AO, Pyrkov TV, Priestle JP, Arseniev AS, Jacoby E (2007) Molecular lipophilicity in protein modeling and drug design. Curr Med Chem 14:393–415
15. Pyrkov TV, Priestle JP, Jacoby E, Efremov RG (2008) Ligand-specific scoring functions: improved ranking of docking solutions. SAR QSAR Environ Res 19:91–99
16. Novoseletsky VN, Pyrkov TV, Efremov RG (2010) Analysis of hydrophobic interactions of antagonists with the beta2-adrenergic receptor. SAR QSAR Environ Res 21:37–55
17. Fauchere JL, Quarendon P, Kaetterer L (1988) Estimating and representing hydrophobicity potential. J Mol Graph Model 6:203–206
18. Pyrkov TV, Chugunov AO, Krylov NA, Nolde DE, Efremov RG (2009b) PLATINUM: a web tool for analysis of hydrophobic/hydrophilic organization of biomolecular complexes. Bioinformatics 25:1201–1202
19. Bohm HJ (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput Aided Mol Des 8:243–256
20. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 11:425–445
21. 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
22. Mooij WTM, Verdonk LM (2005) General and targeted statistical potentials for protein-ligand interactions. Proteins Struct Funct Bioinf 61:272–287
23. Wang R, Fang X, Lu Y, Wang S (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem 47: 2977–2980
24. Banks JL, Beard HS, Cao Y, Cho AE, Damm W, Farid R, Felts AK, Halgren TA, Mainz DT, Maple JR, Murphy R, Philipp DM, Repasky MP, Zhang LY, Berne BJ, Friesner RA,Gallicchio E, Levy RM (2005) Integrated Modeling Program, Applied Chemical Theory (IMPACT). J Comput Chem 26:1752–1780
25. Stroganov OV, Novikov FN, Stroylov VS, Kulkov V, Chilov GG (2008) Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening. J Chem Inf Model 48:2371–2385

Chapter 3
Large DNA Template Dependent Error Variation During Transcription

Harriet Mellenius and Måns Ehrenberg

Abstract The accuracy of an enzymatic reaction system is the propensity to process the correct, or cognate, substrate over similar, non-cognate substrates. This is of particular importance to polymerization reactions with a template sequence, like transcription, translation and replication. A theoretical framework for the analysis of accuracy control is presented, including initial substrate selection and kinetic proofreading. This framework allows for analysis not only of the efficiency of accuracy control, but also its source in standard free energy differences and equilibrium constants and its relation to the rate of product formation. A key feature is the separation of the selection in a context dependent discard parameter and a context independent discrimination parameter. When the theory is applied to the example of prokaryote transcription, it is shown that the discard parameter, composed by experimentally well-defined values, induces a large template sequence dependent error rate variation.

3.1 Background and Chapter Layout

In an enzyme catalyzed reaction, it is possible that a substrate other than the intended diffuses into the active site. If two substrates, one cognate (S^c) and the other non-cognate (S^{nc}) to the enzyme, have similar structures, the enzyme may in addition to its catalysis of cognate product (P^c) erroneously catalyze formation of non-cognate product (P^{nc}) from the non-cognate substrate. Template directed enzymatic reactions in which substrate selection is essential are DNA replication to DNA by DNA polymerase [19], transcription of DNA to RNA by RNA polymerase [28]

H. Mellenius (✉) • M. Ehrenberg (✉)
Department of Cell and Molecular Biology, Biomedical Center, Uppsala University, Box 596, SE-751 24 Uppsala, Sweden
e-mail: harriet.mellenius@icm.uu.se; ehrenberg@xray.bmc.uu.se

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_3,

and translation of messenger RNA to protein by the ribosome [16]. Another set of enzymes for which high substrate selectivity is substantial are the aminoacyl-tRNA synthetases, which link each of the 20 canonical amino acids to its cognate set of isoaccepting transfer RNAs [13]. In this chapter we will first discuss principal issues of accuracy in simple, single step enzymatic selection systems. Then we will introduce the notion of free energy driven proofreading selection [5, 9, 11, 15, 20] that has been identified for DNA replication and damage repair [3, 21], aminoacylation of tRNA [12, 22, 32], translation of mRNA [24, 30]. There is, in addition, strong experimental evidence for proofreading in transcription of DNA [18], but such a mechanism seems to be lacking for the highly error prone reverse transcriptase [23]. After discussing principal issues regarding enzymatic selection systems, we will explain the standard free energy landscape of the transcription complex as the RNA polymerase moves along its DNA template during operon transcription. Then, we show how this standard free energy landscape affects basal accuracy parameters in both the initial and proofreading accuracy of the template dependent selection of the four RNA polymerase substrate bases A, U, G and C. Finally, we will discuss putative consequences of the large and template dependent variation of transcription errors that by our theoretical approach is predicted to exist in all living systems.

3.2 Single Step Substrate Selection by Enzymes

Consider an enzyme, E, which in the same living cell or in the same test tube catalyzes formation of a cognate product, P^c, from a cognate substrate, S^c, and also a non-cognate substrate, S^{nc}, to a non-cognate product, P^{nc}. As long as the enzyme follows Michaelis-Menten kinetics, the steady state flows j^c and j^{nc} into cognate and non-cognate product formation, respectively, are given by

$$\begin{aligned} j^c &= E\,[S^c]\,\frac{k^c_{cat}}{K^c_m}, \\ j^{nc} &= E\,[S^{nc}]\,\frac{k^{nc}_{cat}}{K^{nc}_m} \end{aligned} \tag{3.1}$$

Each flow is determined by the amount, E, of enzyme in the system, the substrate concentration $[S^c]$ or $[S^{nc}]$ and the Michaelis-Menten parameter $k^c_{cat}\,/\,K^c_m$ or $k^{nc}_{cat}\,/\,K^{nc}_m$. We define the current, normalized, accuracy, A, as the ratio between the cognate and the non-cognate $k_{cat}\,/\,K_m$ -value:

$$A = \left(k^c_{cat}\,/\,K^c_m\right)/\left(k^{nc}_{cat}\,/\,K^{nc}_m\right) \tag{3.2}$$

With this definition, A equals j^c/j^{nc} when $[S^c] = [S^{nc}]$. In general, the Michaelis-Menten parameter $k_{cat}\,/\,K_m$ is equal to the rate constant of association

of substrate to enzyme, k_a, multiplied by the probability, P_{prod}, that the substrate is transformed to and released as product rather than released as substrate [8]. Accordingly, A can be written

$$A = \left(k_a^c P_{prod}^c\right) \Big/ \left(k_a^{nc} P_{prod}^{nc}\right) \tag{3.3}$$

It is useful to relate the general relations in Eqs. (3.1, 3.2, and, 3.3) to the simple scheme:

$$\begin{aligned} E + S^c &\underset{k_d^c}{\overset{k_a^c}{\rightleftharpoons}} ES^c \longrightarrow k_c^c E + P^c, \\ E + S^{nc} &\underset{k_d^{nc}}{\overset{k_a^{nc}}{\rightleftharpoons}} ES^{nc} \xrightarrow{k_c^{nc}} E + P^{nc} \end{aligned} \tag{3.4}$$

Here, $P_{prod}^c = k_c^c/(k_d^c + k_c^c)$ and $P_{prod}^{nc} = k_c^{nc}/(k_d^{nc} + k_c^{nc})$ so that

$$A = \frac{k_{cat}^c/K_m^c}{k_{cat}^{nc}/K_m^{nc}} = \frac{k_a^c k_c^c/(k_d^c + k_c^c)}{k_a^{nc} k_c^{nc}/(k_d^{nc} + k_c^{nc})} = \frac{d_a + ad}{1 + a}, \tag{3.5}$$

where

$$\begin{aligned} a &= k_d^c/k_c^c \\ d_a &= k_a^c/k_a^{nc} \\ d &= d_a \left(k_d^{nc}/k_c^{nc}\right) / \left(k_d^c/k_c^c\right) = \left(K_D^{nc}/K_D^c\right)\left(k_c^c/k_c^{nc}\right) \end{aligned} \tag{3.6}$$

The equilibrium dissociation constants K_D^c and K_D^{nc} are defined as k_d^c/k_a^c and k_d^{nc}/k_a^{nc}, respectively. The discard parameter a [17] is the *common* part of cognate and non-cognate substrate kinetics, while d_a and d quantify their *difference*. We note that as the common factor a varies from zero to infinity, the accuracy A varies from d_a to d:

$$d_a < A < d \tag{3.7}$$

It follows that the parameter a determines for each position where in the range of the total accuracy that the system operates, while the parameter d determines the limits of accuracy control. The discard parameter is related to the cognate kinetic efficiency through

$$\left(\frac{k_{cat}}{K_m}\right)^c = \frac{k_a^c}{1 + a} \tag{3.8}$$

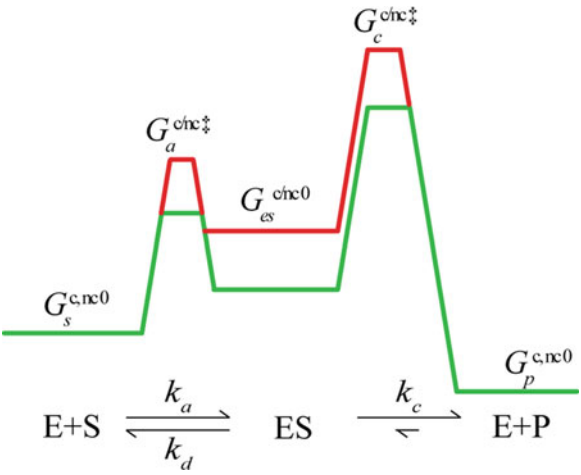


Fig. 3.1 Standard free energies for stationary and transition states of an example reaction system, showing energy level differences between cognate (*green*) and non-cognate substrates (when deviating, *red*, above green line) (Color figure online)

It is informative to relate the equilibrium and rate constants in scheme (3.4) along with the *d*-values in Eq. (3.6) to standard free energies that generate them. A standard free energy diagram relevant to the cognate and non-cognate reactions in scheme (3.4) is shown in Fig. 3.1.

According to transition state theory, the cognate rate constants are:

$$\begin{aligned} k_a^c &= k_{pre} \cdot e^{-\left(G_a^{c\ddagger} - G_s^{c0}\right)/k_B T} = k_{pre} \cdot e^{-\Delta G_a^{c\ddagger}/k_B T}, \\ k_c^c &= k_{pre} \cdot e^{-\left(G_c^{c\ddagger} - G_{es}^{c0}\right)/k_B T} = k_{pre} \cdot e^{-\Delta G_c^{c\ddagger}/k_B T}, \\ k_d^c &= k_{pre} \cdot e^{-\left(G_d^{c\ddagger} - G_{es}^{c0}\right)/k_B T} = k_{pre} \cdot e^{-\Delta G_d^{c\ddagger}/k_B T}. \end{aligned} \tag{3.9a}$$

Similarly, the non-cognate rate constants are:

$$\begin{aligned} k_a^{nc} &= k_{pre} \cdot e^{-\left(G_a^{nc\ddagger} - G_s^{nc0}\right)/k_B T} = k_{pre} \cdot e^{-\Delta G_a^{nc\ddagger}/k_B T}, \\ k_c^{nc} &= k_{pre} \cdot e^{-\left(G_c^{nc\ddagger} - G_{es}^{nc0}\right)/k_B T} = k_{pre} \cdot e^{-\Delta G_c^{nc\ddagger}/k_B T}, \\ k_d^{nc} &= k_{pre} \cdot e^{-\left(G_d^{nc\ddagger} - G_{es}^{nc0}\right)/k_B T} = k_{pre} \cdot e^{-\Delta G_d^{nc\ddagger}/k_B T}. \end{aligned} \tag{3.9b}$$

From Eqs. (3.6) and (3.9) it follows that the *d*-values can be expressed in standard free energies as:

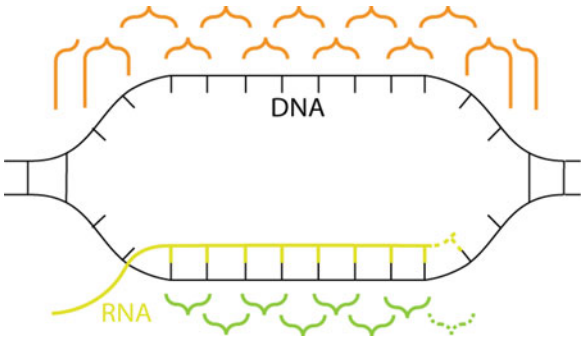


Fig. 3.2 A transcription bubble with nearest neighbour pairs indicated. DNA pairs are shown in *orange* and RNA pairs in *green*. The presence of a nucleotide at the active site (*dashed*) depends on the state of the system (Color figure online)

$$
\begin{aligned}
a &= e^{\left(\left(G_c^{c\ddagger}-G_{es}^{c0}\right)-\left(G_d^{c\ddagger}-G_{es}^{c0}\right)\right)/k_BT} = e^{\left(\left(\Delta G_c^{c\ddagger}\right)-\left(\Delta G_d^{c\ddagger}\right)\right)/k_BT}, \\
d_a &= e^{\left(\left(G_a^{nc\ddagger}-G_s^{nc0}\right)-\left(G_a^{c\ddagger}-G_s^{c0}\right)\right)/k_BT} = e^{\left(\left(\Delta G_a^{nc\ddagger}\right)-\left(\Delta G_a^{c\ddagger}\right)\right)/k_BT} = e^{\Delta\Delta G_a^{\ddagger}/k_BT}, \\
d &= e^{\left(\left(G_c^{nc\ddagger}-G_s^{nc0}\right)-\left(G_c^{c\ddagger}-G_s^{c0}\right)\right)/k_BT} = e^{\left(\left(\Delta G_c^{nc\ddagger}\right)-\left(\Delta G_c^{c\ddagger}\right)\right)/k_BT} = e^{\Delta\Delta G_c^{\ddagger}/k_BT}
\end{aligned}
\tag{3.10}
$$

The interpretation of the intrinsic or maximal accuracy d in Eq. (3.10) is that it reflects the standard free energy difference between the transition state for product formation and the free substrate state for then non-cognate reaction minus the corresponding difference for the cognate reaction.

We note that the intrinsic or maximal accuracy parameter d can also be written as $d = d_a d_{dc}$, where (Fig. 3.1)

$$
d_{dc} = e^{\left(\left(G_c^{nc\#}-G_d^{nc\#}\right)-\left(G_c^{c\#}-G_d^{c\#}\right)\right)/k_BT} = e^{\Delta\Delta G_{dc}^{\#}/k_BT} \tag{3.11}
$$

When $d_a = 1$, then $d = d_{dc}$ and

$$
1 < A < d \tag{3.12}
$$

It follows from Eqs. (3.5) and (3.8) that there is a linear trade-off between the efficiency of the cognate reaction and the accuracy (Fig. 3.2)

$$
k_{cat}^c / K_m^c = k_a^c (d - A)/(d - d_a) \tag{3.13}
$$

That is, when a varies from zero to infinity at constant k_a^c, d and d_a, the accuracy increases from its smallest value of 1.0, implying no substrate selectivity, to its asymptotically largest value of d. At the same time, the kinetic efficiency k_{cat}/K_m decreases from its largest value of k_a^c to zero. This is the inevitable trade-off between rate and accuracy in all types of enzymatic selections, implying that enzymes

operating close to their maximal accuracy limit have very low kinetic efficiency. Accordingly, enzymes are tuned so that their selectivity is high, but not too close to its maximal value, and so that their efficiency is high, but not too close to its maximal value. Another way to view this trade-off is as follows. The flows of cognate and non-cognate product formation obey the following relations

$$j^c = E\,[S^c]\,k_c^c/K_m^c = ES^c k_c^c = E\,[S^c]\left(\delta^c/K_D^c\right)k_c^c,$$

$$j^{nc} = E\,[S^{nc}]\,k_c^{nc}/K_m^{nc} = ES^{nc} k_c^{nc} = E\,[S^{nc}]\left(\delta^{nc}/K_D^{nc}\right)k_c^{nc} \tag{3.14}$$

Here, δ^c and δ^{nc} signify how much the complexes ES^c and ES^{nc}, respectively, are shifted below equilibrium with their free states in the steady state reaction. Furthermore, we obtain

$$A = \frac{k_{cat}^c/K_m^c}{k_{cat}^{nc}/K_m^{nc}} = d\cdot\delta^c/\delta^{nc} = d\cdot\frac{K_D^c}{K_m^c}\cdot\frac{K_m^{nc}}{K_D^{nc}} = d\,\frac{a}{1+a}\cdot\frac{1+d_{dc}a}{d_{dc}a} = \frac{d_a+da}{1+a} \tag{3.15}$$

This relation, equivalent to Eq. (3.5) above, shows that maximal accuracy $A = d$ is obtained when both ES^c and ES^{nc} are in equilibrium with their free states, so that $\delta^c = \delta^{nc} = 1$.

3.3 Multiple Step Substrate Selection by Proofreading Enzymes

Cognate and non-cognate substrates are often quite similar, which will create an upper limit of the d-value that separates their product formation flows by the selecting enzyme system. It is, for instance, essential for cell viability that the error frequency per base ($1/A$) in genome replication is smaller than the inverse of the effective genome size [6]. For the human genome with three billion bases this would require a d-value of DNA replication larger than one billion for a DNA polymerase with single step selectivity as in scheme (3.4). If the d-value for base separation by the DNA polymerase is 10,000, this selection scheme would limit the human genome size to about 10,000 bases, and thus make the evolution of *Homo sapiens* impossible. A fundamental question is therefore if it is possible for an enzyme to achieve an accuracy parameter, A, larger than the d-value. This question was formulated and answered by Hopfield [11] and Ninio [20]): yes, it is possible by addition of one (or several) free energy driven proofreading step(s) after an initial selection step as in scheme (3.4):

$$
\begin{array}{c}
E + Co\text{-}S + S^c \underset{k_d^c}{\overset{k_a^c}{\rightleftharpoons}} ECo\text{-}SS^c \xrightarrow{k_c^c} ECo\text{-}PS^c \xrightarrow{k_p^c} E + Co\text{-}P + P^c \\
\downarrow q_d^c \\
E + Co\text{-}P + S^c \\
E + Co\text{-}S + S^{nc} \underset{k_d^{nc}}{\overset{k_a^{nc}}{\rightleftharpoons}} ECo\text{-}SS^{nc} \xrightarrow{k_c^{nc}} ECo\text{-}PS^{nc} \xrightarrow{k_p^{nc}} E + Co\text{-}P + P^{nc} \\
\downarrow q_d^{nc} \\
E + Co\text{-}P + S^{nc}
\end{array}
\tag{3.16}
$$

In this scheme a second exit step has been added, with dissociation rate constant q_d^c for the cognate and q_d^{nc} for the non-cognate substrate. Product is formed from the ECo-PS state, with rate constant k_p^c for the cognate and k_p^{nc} for the non-cognate substrate. In the initial selection step, the substrate S and co-substrate Co-S dissociates from the enzyme or the co-substrate is transformed to co-product Co-P. The co-substrate to co-product formation or an equivalent reaction is fundamental to drive the proofreading reaction. For this, co-substrate must be present at a concentration far above equilibrium with co-product:

$$
[Co\text{-}S] = [Co\text{-}P]\, K_{Co\text{-}SCo\text{-}P}\, \gamma \tag{3.17}
$$

For proofreading to work it is necessary that $\gamma \gg 1$, so that every time a co-substrate Co-S is transformed to co-product Co-P, free energy corresponding to $k_B T \log \gamma$ is dissipated. When $\gamma = 1$ there is no energy dissipation and the detailed balance constraint forbids neglect of inflow over the exit steps in the proofreading part of the mechanism [5]. In fact, when $\gamma = 1$ the additional selection step in scheme (3.16) does not increase the accuracy, A, of substrate selection [5]. As before, we define the accuracy A as the ratio between the k_{cat}/K_m-values for cognate and non-cognate substrates:

$$
A = \left(k_a^c P_{prod}^c\right) \Big/ \left(k_a^{nc} P_{prod}^{nc}\right) = \frac{k_a^c P_I^c}{k_a^{nc} P_I^{nc}} \frac{P_F^c}{P_F^{nc}} = I \cdot F \tag{3.18}
$$

P_I is the probability that substrate and co-substrate dissociates from the enzyme before co-substrate transformation to co-product. P_F is the probability that product is formed and released from the state ECo-PS in scheme (3.16). I is the accuracy provided by initial selection and F is the accuracy provided by proofreading selection of substrate. The overall accuracy A written in terms of the rate constants in scheme is given by:

$$
A = I \cdot F = \frac{k_a^c k_c^c/(k_d^c + k_c^c)}{k_a^{nc} k_c^{nc}/(k_d^{nc} + k_c^{nc})} \cdot \frac{k_p^c/(q_d^c + k_p^c)}{k_p^{nc}/(q_d^{nc} + k_p^{nc})} = \frac{d_a + a_I d_I}{1 + a_I} \cdot \frac{1 + a_F d_F}{1 + a_F} \tag{3.19}
$$

Here, d_a, a_I and d_I are defined as d_a, a and d in Eq. 3.6, respectively, while a_F and d_F are defined as

$$a_F = q_d^c / k_p^c,\ d_F = \left(q_d^{nc} / k_p^{nc}\right) \Big/ \left(q_d^c / k_p^c\right) \tag{3.20}$$

It is seen that when a_I and a_F vary between zero and infinity, A varies between its lowest value d_a and its highest value $d_I d_F$:

$$d_a < A < d_I d_F \tag{3.21}$$

From this follows that A can be larger than the single step selectivity d_I of scheme (3.4). The trade-off between the efficiency, k_{cat}/K_m, for cognate product formation as given for a single step mechanism in Eq. 3.13 can in the case of a single proofreading step be generalized to

$$k_{cat}^c / K_m^c = k_a^c \frac{(d_I - I)}{(d_I - d_a)} \cdot \frac{(d_F - F)}{(d_F - 1)} \tag{3.22}$$

It is in principle possible to have several proofreading exit steps following the transformation of a single co-substrate to co-product. In general therefore, for a mechanism with n proofreading steps the accuracy A obeys the inequality [11]:

$$d_a < A < d_I d_F^n \tag{3.23}$$

The proofreading mechanism requires that the co-substrate is shifted above equilibrium with co-product. Neglect of inflow over the proofreading steps is a valid approximation only as long as the proofreading selection, F, remains smaller than the shift, γ, of co-substrate above equilibrium with co-product. Therefore, there is yet another inequality for A that is always valid [5]:

$$d_a < A < d_I \gamma \tag{3.24}$$

We note that the number, f_c, of co-substrates transformed to co-products per cognate product is equal to the inverse of the probability, P_F^c, that a cognate substrate survives the proofreading step(s) and ends up as product. Furthermore, that the corresponding number, f_{nc}, of co-substrates dissipated per non-cognate product is equal to the inverse of the probability, P_F^{nc}, that a non-cognate substrate survives the proofreading step(s). This means that the accuracy contribution by proofreading can always be related to the excess dissipation of co-substrates in the cognate and non-cognate reaction pathways:

$$F = \frac{P_F^c}{P_F^{nc}} = \frac{f_{nc}}{f_c} \tag{3.25}$$

In general, these excess dissipation parameters can be measured experimentally and thus used to identify a putative proofreading mechanism by the criterion that $F \gg 1$ and to determine the accuracy contribution by proofreading through an

estimate of F. Indeed, this strategy was first used by Hopfield et al. [12] to identify a proofreading mechanism for the enzyme IleRS, which uses ATP as co-substrate as it aminoacylates $tRNA^{Ile}$ with the amino acid isoleucine to Ile-$tRNA^{Ile}$. Hopfield et al. found that about 1.5 molecules of ATP were hydrolyzed for every molecule of Ile-$tRNA^{Ile}$ formed ($f_c = 1.5$) and that about 300 molecules of ATP were hydrolyzed for every non-cognate aminoacylation of $tRNA^{Ile}$ with the amino acid valine to Val-$tRNA^{Ile}$ ($f_{nc} = 300$). From this experiment they could conclude that IleRS uses proofreading ($F = f_{nc}/f_c \gg 1$) and that proofreading contributes by a factor of about 200 (300/1.5) to the overall accuracy of the enzyme.

3.4 Accuracy-Demanding Biological Systems

The above theoretical description of substrate selection in a biochemical system can typically be applied to any enzymatic reaction. However, certain reaction systems come with a higher demand for accuracy, and this is where the accuracy computations are of particular relevance. These systems involve catalysis of polymerizations, in which an erroneous product means not only that the processing of that specific substrate was futile, but that the whole sequence of thousands or even millions of repeat units could be corrupted. As already mentioned in the introduction, the most notable examples of such processes are replication, translation and transcription. In the next section, we use the prokaryotic transcription system from *E. coli* for exemplification.

3.5 Accuracy in Transcription

In the ternary complex of transcription, the enzyme is RNA polymerase, the template is a double stranded DNA sequence, locally denatured to form a transcription "bubble" with its base sequence exposed (Fig. 3.2). The substrates are the four canonical nucleoside triphosphates A, U, G and C, that form the RNA transcript by Watson-Crick base-pairing with the template bases T, A, C and G, respectively. A transcription error occurs when a nucleotide becomes phosphodiester bonded to a nascent RNA chain in erroneous, not Watson-Crick, base pair with the opposite template base.

The achievement of a low transcription error frequency is complicated by the fact that the nucleobases are turned towards each other, so that the polymerase interacts only with their backbones. Hence, hydrogen bonds and other interactions in-between bases must be recognized by the effect they have on the total stability of the ternary complex.

Yager and von Hippel [31] quantified the stability of the transcription elongation complex as the sum of the free energy terms for separating the double stranded DNA, $\Delta G^0_{DNA/DNA}$, forming an RNA/DNA hybrid with the template

strand, $\Delta G^0_{RNA/DNA}$, and the binding interactions between the polymerase and the nucleotide sequences within the transcription bubble, $\Delta G^0_{polymerase}$. In this description the current standard free energy of the transcription complex, ΔG^0, is given by

$$\Delta G^0 = \Delta G^0_{DNA/DNA} + \Delta G^0_{RNA/DNA} + \Delta G^0_{polymerase} \tag{3.26}$$

The nucleic acids in the transcription complex consist of 2×12 complementary DNA nucleotides that form the DNA bubble and 8–9 base pairs of RNA/DNA hybrid that stabilize it, thus giving each polymerase position along the operon a characteristic, sequence specific transcription bubble standard free energy. As the polymerase translocates one base along the template sequence, one forward base pair has to melt and one backward base pair has to form as the bubble is updated. The standard free energy of the nucleotide sequences is calculated by taking into account the sum of the nearest-neighbour parameters of each of the constituent nucleotide pairs, including the stacking energies between them, which vary considerably between neighbouring base pairs [25–27].

Figure 3.2 shows the nearest neighbour pairs of the transcription bubble, with DNA pairs in orange and RNA/DNA hybrid pairs in green. The last nearest neighbour pair of the hybrid (dashed) is situated in the catalytic site of the polymerase. When the polymerase translocates one step forward, this position becomes empty until the arrival of the next nucleotide.

The kinetics of transcript elongation as modelled by Bai et al. [1], following the description of polymerase movement from Guajardo and Sousa [10], uses the free energy exactly in this way. The movement of the polymerase is viewed as a "Brownian ratchet" mechanism, where the polymerase can translocate stochastically between a number of states. The incoming nucleoside triphosphate (NTP) and the phosphodiester bond formation capture and lock the forward translocated state, supplying a forward bias. Tadigotla et al. [29] introduced a statistical mechanical approach to the problem, where the probability of reaching the available states are given by Boltzmann factors, including the free energy of the available positions and the energy barrier of the reaction between them. These contributions were combined by Dennis et al. (2009) into the transcription model used here.

3.6 Initial Selection

The initial selection of transcription represents the enzyme's ability to distinguish between cognate and non-cognate substrates on their path to RNA chain elongation by phosphodiester bond formation. This path includes the arrival of the substrate to the active site, hydrogen bond formation with the opposite template nucleotide and catalysis of the phosphodiester bond reaction. While the stationary states are defined by the stability of the constituent parts of the complex, the transition state of bond formation also includes a reaction barrier (Fig. 3.3).

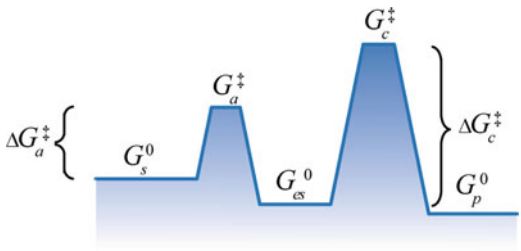


Fig. 3.3 Standard free energies of the initial selection in transcription, as presented in scheme (3.27)

A new nucleotide enters the polymerase in post-translocated state, where the catalytic site is unoccupied and the length of the stabilizing RNA/DNA hybrid is eight base pairs (Fig. 3.2). The interaction between the incoming substrate and the template DNA will thus stabilize the complex, but the magnitude of stabilization will vary, depending on the template, the substrate and the previously inserted base. Assuming Michaelis-Menten kinetics, the schemes of cognate and non-cognate product formation are analogous to those in Eq. (3.4) above, but here we indicate the position of the transcription bubble and the length of the nascent RNA chain

$$\begin{aligned} ED_jR_i + NTP^c &\underset{k_d^c}{\overset{k_a}{\rightleftharpoons}} ED_jR_i - NTP^c \xrightarrow{k_c^c} ED_jR_{i+1}^{i+1=c} + PP_i \\ ED_jR_i + NTP^{nc} &\underset{k_d^{nc}}{\overset{k_a}{\rightleftharpoons}} ED_jR_i - NTP^{nc} \xrightarrow{k_c^{nc}} ED_jR_{i+1}^{i+1=nc} + PP_i \end{aligned} \tag{3.27}$$

The RNA polymerase (E) in complex with the DNA transcription bubble in position j along the sequence (D_j) and the RNA transcript of length i (R_i) is called ED_jR_i. The substrate is a nucleoside tri-phosphate (NTP), which is ester-bonded with the nascent RNA chain as NMP in a reaction accompanied by pyrophosphate (PP_i) release. In analogy with the general system for one-step selectivity above, the accuracy is defined through the probabilities of product formation, which in turn are determined by the rate constants for the cognate and non-cognate reactions. Here, however, we will take the analysis one step further by specifying the cognate substrate (X = A, U, G or C), its non-cognate competitor (Y ≠ X) and the identity of the previously incorporated nucleotide (N = A, U, G or C). Accordingly, we write the normalized initial selection parameter as

$$I_{NY}^X = \frac{1 + a_{NX} d_{NY}^X}{1 + a_{NX}} \tag{3.28a}$$

To simplify, we have assumed that the polymerase does not discriminate between correct and non-correct substrate by their association rate constants ($k_a^{nc} = k_a^c$). The discard parameter, a_{NX}, depends on the identity, X, of the cognate substrate since, for instance, a G-C base pair with three H-bonds is expected to be more stable

than an A-T pair with two H-bonds. The stacking energy provided by the nearest neighbour RNA base will depend both on the identity of the cognate base and its previously inserted neighbour, and hence the cognate discard parameter will also depend on *N*. The intrinsic discrimination, *d*, will depend on the identity, X, of the cognate, the identity, Y, of the non-cognate and the identity, *N*, of the neighbour base. To obtain the form (3.28a) of the initial selection and well-defined stacking energies also for non-cognate substrates, we note that I_{NY}^{X} can also be written as

$$I_{NY}^{X} = \frac{1 + a_{NY} d_{Y}^{X}}{1 + a_{NX}} \tag{3.28b}$$

Here, a_{NY} is the cognate discard parameter for the base *Y* when selected by its matching template base with N as its nearest neighbour and d_{Y}^{X} is the intrinsic discrimination against *Y*-incorporation when its cognate template base is swapped for the cognate template base of X in the same neighbour context *N*. From this we obtain

$$I_{NY}^{X} = \frac{1 + a_{NY} d_{Y}^{X}}{1 + a_{NX}} = \frac{1 + a_{NX} \left(\frac{a_{NY}}{a_{NX}}\right) d_{Y}^{X}}{1 + a_{NX}} = \frac{1 + a_{NX} d_{NY}^{X}}{1 + a_{NX}}, \tag{3.29}$$

where d_{NY}^{X} is defined by

$$d_{NY}^{X} = \frac{a_{NY}}{a_{NX}} d_{Y}^{X} \tag{3.30}$$

Accordingly, discrimination parameter d_{NY}^{X} has a *context independent* part, d_{Y}^{X}, and a *context dependent* part, a_{NY}/a_{NX}. The former depends on base-pair interactions between mismatched bases, and the latter is determined from well-defined stacking energies related to cognate template interactions. The validity of this approach requires that each non-cognate substrate has the same sterical configuration in the transition state for phosphodiester bond formation as it would have with a cognate template base in the same neighbour context. The rationale for this assumption is that catalysis is likely to require a precise sterical configuration, similar to that associated with the corresponding cognate reaction. In terms of standard free energies, the parameters a_{NY}, a_{NX}, d_{Y}^{X} and d_{NY}^{X} can be written as

$$\begin{aligned}
a_{NX} &= e^{\left(G_c^{c\ddagger} - G_d^{c\ddagger}\right)/k_B T} = e^{\left(\left(G_{es}^{c0} + \Delta G_c^{c\ddagger}\right) - \left(G_s^{c0} + \Delta G_a^{\ddagger}\right)\right)/k_B T}, \\
d_{NY}^{X} &= e^{\left(\left(\left(G_{es}^{nc0} + \Delta G_c^{nc\ddagger}\right) - \left(G_s^{0} + \Delta G_a^{\ddagger}\right)\right) - \left(\left(G_{es}^{c0} + \Delta G_c^{c\ddagger}\right) - \left(G_s^{0} + \Delta G_a^{\ddagger}\right)\right)\right)/k_B T} \\
&= e^{\left(\left(G_{es}^{nc0} + \Delta G_c^{nc\ddagger}\right) - \left(G_{es}^{c0} + \Delta G_c^{c\ddagger}\right)\right)/k_B T} \\
&= e^{\left(\left(G_{es}^{nc0} - G_{es}^{c0}\right) + \left(\Delta G_c^{nc\ddagger} - \Delta G_c^{c\ddagger}\right)\right)/k_B T} = \frac{a_{NY}}{a_{NX}} d_{Y}^{X}
\end{aligned} \tag{3.31}$$

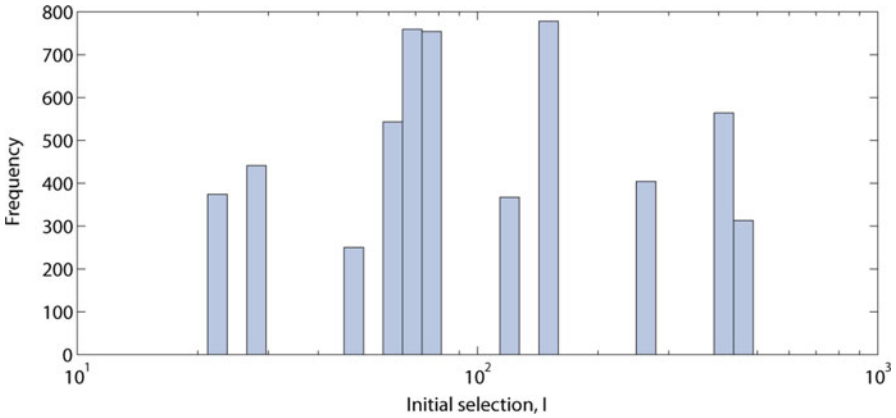


Fig. 3.4 Histogram over the initial selection in *E. coli rrnC*. The initial selectivity varies by a factor of 25. Parameters: $k_c = 2{,}000\ s^{-1}$, $\Delta G_a = 1RT$, $d_Y^X = 1{,}000$

Figure 3.4 shows the initial selection for every position in the *E. coli* rRNA operon C, calculated from its base sequence using Eqs. (3.29) and (3.31). Assuming equimolar amounts of all four nucleotide triphosphates, the non-cognate processing probability will be the sum of probabilities of product formation for each of the three non-cognate substrates to give the total initial selection.

Many positions share the same initial accuracy, since the sequence dependence rely on only two positions, the incorporated nucleotide and its nearest neighbour. Even though the absolute magnitude of the accuracy is arguable, the considerable variation in accuracy (a factor 25 in a 5,600 bp operon) stems from the a-values with experimentally well-defined parameters.

The d_Y^X-values are more difficult to predict, but a clue to the discriminating power of transcriptional proofreading can be found in the newly determined discrimination parameters for some translation errors [14]. The order of magnitude of the d-values is likely to be similar for all the systems based on nucleotide pairing, since the structural contingencies of the substrates are the same.

3.7 Transcriptional Proofreading

The surrounding sequence, apart from the bases in the active site, will define the importance of the stability of the next position, which is relevant for the chance of proofreading – the only way to rectify an incorporation of a non-cognate substrate.

Transcript cleavage in *E. coli* is always preceded by backtracking of the polymerase along the DNA. When the polymerase backtracks, the just extended 3′-end of the nascent transcript is extruded "backwards" along with the downstream DNA. Hydrogen bonds between template and transcript are broken at the 3′-end, and

the entire transcription bubble is shifted upstream, which means that the proximal upstream DNA base-pair is re-opened while the last base-pair at the other end of the bubble is closed [10].

In the backtracked state, the polymerase possesses endonuclease activity, and can cleave off the extruding transcript. The cleavage occurs at the position right next to the active site, leaving the active site open for the next incorporation, suggesting that the cleavage product is at least two bases in length [33]. It has been experimentally observed that the probability of the translocation events is shifted towards backtracking after incorrect incorporations, instead of towards the forward translocation followed by the next nucleotide addition [7].

The existence of kinetic proofreading in transcription would ideally be proved the same way as translational proofreading; by separate observation of the cognate, f_c, and non-cognate, f_{nc}, consumed per irreversibly formed phosphodiester bond formed per consumed nucleoside triphosphate. This has not been done, but the existence of a transcript cleavage mechanism in RNA polymerase still makes transcriptional proofreading a strong suggestion. Transcript cleavage leads to re-incorporation of nucleotides and a new cycle of initial selection. If misincorporations have an increased rate of transcript cleavage, accuracy is consequently enhanced. The reaction scheme of transcriptional proofreading would be partly the same as for initial selection in scheme (3.27), but with the extension of the consecutive transcript elongation step:

$$
\begin{aligned}
&ED_jR_i + NTP^c \underset{k_d^c}{\overset{k_a^c}{\rightleftharpoons}} ED_jR_i - NTP^c \xrightarrow{k_c^c} \\
&ED_jR_{i+1}^{i+1=c} + PP_i \xrightarrow{k_p^c} ED_{j+1}R_{i+2}^{i+1=c} + PP_i \\
&\qquad \downarrow q_d^c \\
&ED_{j-1}R_{i-1} + NMP - NMP^c + PP_i \\
&ED_jR_i + NTP^{nc} \underset{k_d^{nc}}{\overset{k_a^{nc}}{\rightleftharpoons}} ED_jR_i - NTP^{nc} \xrightarrow{k_c^{nc}} \\
&ED_jR_{i+1}^{i+1=nc} + PP_i \xrightarrow{k_p^{nc}} ED_{j+1}R_{i+2}^{i+1=nc} + PP_i \\
&\qquad \downarrow q_d^{nc} \\
&ED_{j-1}R_{i-1} + NMP - NMP^{nc} + PP_i
\end{aligned}
\tag{3.32}
$$

For the sake of simplification, the scheme above and the following discussion assumes that the polymerase backtracks only one step and that only two nucleotides are cleaved, but it is possible that the polymerase backtracks even further to remove longer parts of the transcript (Dennis et al. 2009).

From the position immediately after nucleotide incorporation, the system can either backtrack along the DNA and cleave the transcript with reaction rates q_d, or

go into another round of transcript elongation, which eventually secures the previous incorporation. Reaction rates k_p congregate all the reaction steps in the consecutive transcript elongation, including the association of the next substrate and the forward translocation where the D-position is shifted.

A notable feature of transcriptional proofreading is that here, as well as in DNA-replication, the co-substrate is actually part of the substrate itself. Accordingly, proofreading requires that the parameter, γ, describing how much XTP is shifted above equilibrium with XMP and PP_i must be much larger than one:

$$[XTP]\,[H_20] = \gamma K\,[XMP]\,[P\,P_i] \tag{3.33}$$

The "extra" phosphates on the nucleotide backbone provide the di-phosphate (pyrophosphate) co-product. When XTP is a cognate substrate, the rate of production of XMP in the proofreading selection is expected to be smaller than the rate of synthesis of phosphodiester bonds. If, however, XTP is a cognate substrate, there should be many free XMPs produced per phosphodiester bond, conforming the existence of proofreading and telling us how much it contributes to the accuracy of transcription.

The final accuracy of transcription follows in principle Eqs. (3.19, 3.20, and 3.21) above, but is expected to display very large context variation due to the standard free energy variation as the transcription bubble moves along the template. Again, the accuracy of the proofreading is associated with discard parameters and discrimination parameters.

The discard parameters a_F, the ratio of the rate constants q_d and k_p, show very large variation. The reason is that both reactions involve a translocation event among their sub-steps, backwards or forwards. As explained above for the backtracking, the rate of translocation is determined by the free energy difference between two adjacent transcription bubbles, meaning the energy cost to denaturate that specific sequence of DNA and the energy gained from hybrid formation. This means that the sequence dependence of proofreading concern not only the nucleotides at the active site, but all the base-pairs in the transcription bubble, which infers a large variability. The logical conclusion is that the sequence of the entire transcription bubble will determine the a_F-parameter for each position, and thus give each position a characteristic selection sensitivity. This can be expected from the large sequence dependent variability in transcription speed (Dennis et al. 2009).

Discrimination parameters, d_Y^X, are introduced in both of these translocations on the assumption that a mismatch affects the stability of the entire bubble, but are compiled in a single d_{NY}^X for each non-cognate substrate and position. Figure 3.5 shows the proofreading accuracy for each position as calculated from the sequence of *E. coli* rRNA operon C. The d_Y^X-values are tuned to give a mean accuracy that together with initial selection gives a total accuracy that corresponds to experimental estimates of the overall transcriptional accuracy. The probability of non-cognate product formation used in the accuracy calculation is the sum of the processing probabilities of the three non-cognate substrates in each template directed selection step.

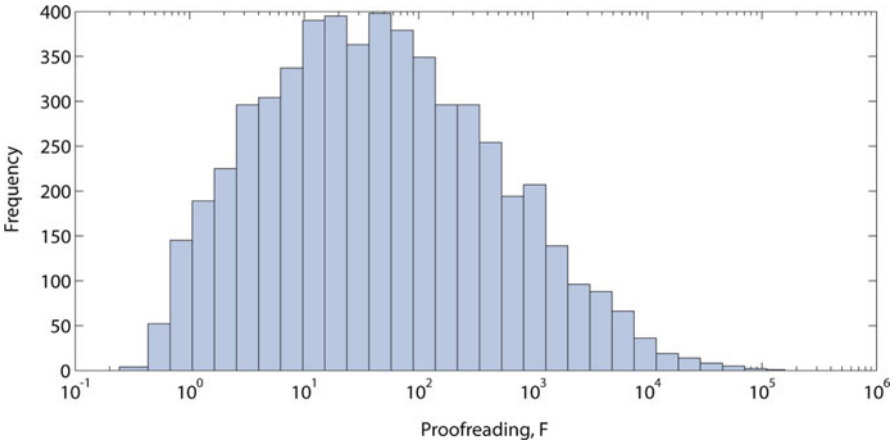


Fig. 3.5 Histogram over the proofreading selection in *E. coli rrnC*. The proofreading selectivity varies by almost a factor 10^6. Parameters: $k_c = 2{,}000\ s^{-1}$, $\Delta G_a = 1RT$, $d_Y^X = 1{,}000$ and 100, $\Delta G_{translocation} = 2RT$

The variability in proofreading selection is almost a factor 10^6, and this variation is, interestingly, due to the sequence specificity in the well-defined a-values. Compared to the initial selection, the variability of the proofreading is magnified by the contributions of all the base-pairs in the transcription bubble. The participation of the a-values in the parameter d_{NY}^X extends the range of the accuracy beyond the limits set in Eq. (3.12). Another effect is that the sequence specificity is enhanced so that no two positions in the operon share exactly the same value of proofreading accuracy. For these reasons, the large variability is expected to be robust to changes in the externally adjustable parameter values.

The total accuracy, the product of initial and proofreading selection, of the example system is shown (Fig. 3.6). The total accuracy varies by more than a factor 10^6.

3.8 Conclusion

The theory presented above provides a practical framework for the analysis of accuracy control in biochemical systems. Initial selection of substrate is separated from the later culling of misincorporation of proofreading. The driving forces behind both accuracy control systems are in this framework described in terms of the kinetic efficiency, equilibrium and free energy.

Furthermore, the effective selectivity is divided into a context dependent and a context independent part. The context independent part sets the limits of the accuracy control, according to theory, and the context dependent part determines where in this range that the system operates. However, in the example of transcription that we give, we show that when the context dependent discard parameter

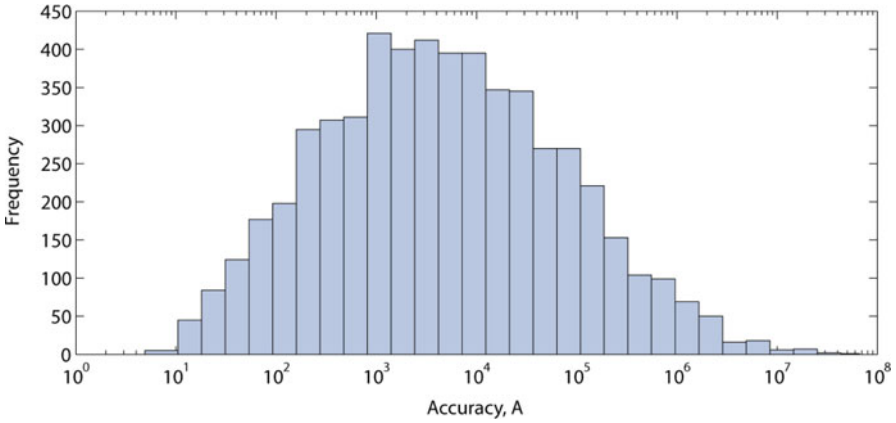


Fig. 3.6 Histogram over the total acuracy in *E. coli rrnC*. The total accuracy varies by more than a factor 10^6. The total accuracy is, for each position, calculated as the product of the initial and proofreading selection as presented in Figs. 3.4 and 3.5

differs between the cognate and the non-cognate substrate, it can inflate or reduce the accuracy beyond the limits set by the context independent discrimination parameters.

The biological implication of this conclusion is that the error rate will vary greatly among the template sequences. There will be a corresponding variation in k_{cat}/K_m of the process, due to the relation of the two (Eq. 3.13). Moreover, this reasoning can be extended to the processes of translation and replication. Replication is affiliated with the same sequence depending stacking effects as transcription, and in translation, the different tRNAs could give incommensurable a-values to different codons.

As this theory makes direct predictions of the accuracy variation for a given sequence, it raises interesting questions for further investigation. One of these is experimental, *i.e.* the theory must be validated by carefully designed *in vitro* or *in vivo* experiments on transcriptional accuracy. Another question relates to molecular evolution: the predicted variation of transcriptional accuracy suggests for instance that certain error prone sequences are counterselected, and this may impact codon usage in all organisms by a mechanism separate from the currently assumed impact on codon usage mainly by translational demands [2].

References

1. Bai L, Shundrovsky A, Wang MD (2004) Sequence-dependent kinetic model for transcription elongation by RNA polymerase. J Mol Biol 344(2):335–349
2. Berg OG, Kurland CG (1997) Growth rate-optimised tRNA abundance and codon usage. J Mol Biol 270(4):544–550

3. Brutlag D, Kornberg A (1972) Enzymatic synthesis of deoxyribonucleic acid. J Biol Chem 247(1):241–248
4. Dennis P, Ehrenberg M, Fange D, Bremer H (2009) Varying rate of RNA chain elongation during rrn transcription in Escherichia coli. J Bacteriol 191(11):3740–3746
5. Ehrenberg M, Blomberg C (1980) Thermodynamic constraints on kinetic proofreading in biosynthetic pathways. Biophys J 31(3):333–358
6. Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften 58(10):465–523
7. Erie DA, Hajiseyedjavadi O, Young MC, von Hippel PH (1993) Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription. Science 262(5135):867–873
8. Fersht A (1999) Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding. W.H. Freeman and Company, New York
9. Freter RR, Savageau MA (1980) Proofreading systems of multiple stages for improved accuracy of biological discrimination. J Theor Biol 85(1):99–123
10. Guajardo R, Sousa R (1997) A model for the mechanism of polymerase translocation. J Mol Biol 265(1):8–19
11. Hopfield JJ (1974) Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proc Natl Acad Sci U S A 71(10):4135–4139
12. Hopfield J, Yamane T, Yue V, Coutts S (1976) Direct experimental evidence for kinetic proofreading in amino acylation of tRNAIle. Proc Natl Acad Sci U S A 73(4):1164–1168
13. Ibba M, Söll D (1999) Quality control mechanisms during translation. Science 286(5446):1893–1897
14. Johansson M, Zhang J, Ehrenberg M (2012) Genetic code translation displays a linear trade-off between efficiency and accuracy of tRNA selection. Proc Natl Acad Sci 109(1):131–136
15. Kurland C (1978) The role of guanine nucleotides in protein biosynthesis. Biophys J 22(3):373–392
16. Kurland C (1992) Translational accuracy and the fitness of bacteria. Annu Rev Genet 26(1):29–50
17. Kurland CG, Hughes D, Ehrenberg M (1996) Limitations of translational accuracy. American Society for Microbiology, Washington, DC, pp 979–1004
18. Libby RT, Nelson JL, Calvo J, Gallant JA (1989) Transcriptional proofreading in Escherichia coli. EMBO J 8(10):3153–3158
19. McCulloch SD, Kunkel TA (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell Res 18(1):148–161
20. Ninio J (1975) Kinetic amplification of enzyme discrimination. Biochimie 57(5):587–595
21. Reardon JT, Sancar A (2004) Thermodynamic cooperativity and kinetic proofreading in DNA damage recognition and repair. Cell Cycle (Georgetown, Tex) 3(2):141–144
22. Reynolds NM, Lazazzera BA, Ibba M (2010) Cellular mechanisms that control mistranslation. Nat Rev Microbiol 8(12):849–856
23. Roberts JD, Bebenek K, Kunkel TA (1988) The accuracy of reverse transcriptase from HIV-1. Science 242(4882):1171–1173
24. Ruusala T, Ehrenberg M, Kurland C (1982) Is there proofreading during polypeptide synthesis? EMBO J 1(6):741–745
25. SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci 95(4):1460–1465
26. SantaLucia J, Hicks D (2004) The thermodynamics of DNA structural motifs. Rev Biophys Biomol Struct 33:415–440
27. Sugimoto N et al (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry 34(35):11211–11216
28. Sydow JF, Cramer P (2009) RNA polymerase fidelity and transcriptional proofreading. Curr Opin Struct Biol 19(6):732–739
29. Tadigotla VR et al (2006) Thermodynamic and kinetic modeling of transcriptional pausing. Proc Natl Acad Sci U S A 103(12):4439–4444

30. Thompson RC, Stone PJ (1977) Proofreading of the codon-anticodon interaction on ribosomes. Proc Natl Acad Sci U S A 74(1):198–202
31. Yager TD, Von Hippel PH (1991) A thermodynamic analysis of RNA transcript elongation and termination in Escherichia coli. Biochemistry 30(4):1097–1118
32. Yamane T, Hopfield J (1977) Experimental evidence for kinetic proofreading in the aminoacylation of tRNA by synthetase. Proc Natl Acad Sci U S A 74(6):2246–2250
33. Zenkin N, Yuzenkova Y, Severinov K (2006) Transcript-assisted transcriptional proofreading. Science 313(5786):518–520

Chapter 4
Structures of Novel HIV-Inactivating Lectins

Leonardus M.I. Koharudin and Angela M. Gronenborn

Abstract To elucidate the structural basis for recognition of high-mannose glycans on HIV-1 gp120, we investigated sugar binding of two novel lectins, OAA and PFA, using NMR spectroscopy and X-ray crystallography. We solved crystal structures of these proteins, free and complexed with α3,α6-mannopentaose, the minimal mannosyl oligosaccharide substructure of the gp120-attached mannoses that is recognized by these lectins. Our data provide the first atomic details of how the core epitope of Man-9 is recognized by the OAA-class of lectins.

4.1 Introduction

Over the last 30 years, extensive and comprehensive research efforts have been directed at understanding the lifecycle of HIV in the fight against AIDS, and, indeed, HIV is perhaps one of the best-understood viruses at the present time. As a result, several effective therapies and drugs that target different phases of the viral life cycle have become available. Despite these major advances, the AIDS pandemic continues to pose serious global concerns. According to the latest statistics (end of 2009), over 33 million people around the world are living with HIV/AIDS, and more than 55 million globally have been infected with the virus since AIDS was first reported more than 30 years ago. The current death toll stands at more than 25 million people, and, in 2009, an estimated 2.7 million people became newly infected with HIV (that is 14,000 individuals per day), and two million died. Approximately 70% of all people living with HIV reside in sub-Saharan Africa, more than 50% of all infected individuals are women, and 91% of new HIV infections occur in

L.M.I. Koharudin • A.M. Gronenborn (✉)
Department of Structural Biology, University of Pittsburgh School of Medicine,
3501 Fifth Avenue, 1051 BST3, Pittsburgh, PA 15261, USA
e-mail: amg100@pitt.edu

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_4,

children. Treatment of HIV-infected individuals generally involves a combination, or 'cocktails', of several classes of drugs, directed predominantly against the key HIV enzymes. This Highly Active Antiretroviral Therapy (or HAART) has proved highly successful and has dramatically reduced mortality. New HIV infections have been reduced by 17% over the past 8 years world-wide, with large declines in sub-Saharan Africa, East Asia, and South East Asia by 15 25, and 10%, respectively. Overall, there is no doubt that remarkable progress has been made in combating AIDS and in understanding the details of the viral life cycle as well as how it interacts with the host's cellular machinery.

While significant progress has been made on treating individuals with HIV, the development of strategies to *prevent* individuals from becoming infected with HIV has progressed more slowly. Nearly two decades ago, new initiatives to prevent sexual transmission of the virus via antiviral agents were encouraged. In particular, approaches involving microbicide use could be particularly helpful in curbing the escalating rate of HIV infection among women – in many societies, women can neither control sexual encounters nor can they insist on protective measures, such as abstinence or mutual monogamy. Such agents, when incorporated into vaginal and rectal gels, foams and suppositories, sponges or rings for *ex vivo* or topical use, can potentially decrease or prevent the sexual transmission of HIV. Promising candidates for such barrier applications are substances that directly interact with HIV virions, thus preventing viral entry into and fusion with the target cells (e.g. Carraguard®, Cyanoviran®, cellulose sulphate, PRO 2000®), or that act by enhancing the natural vaginal defense mechanism of maintaining an acidic pH (e.g. Acidform®, BufferGel® and Lactobacillus crispatus). (See http://www.who.int/hiv/topics/microbicides/microbicides/en/ for updates).

A number of these microbicide products have gone through various stages of clinical development. Carraguard®, a product of the Population Council, underwent a Phase III effectiveness trial in a general population of women in South Africa and Botswana. It was found to be safe, although the trial did not show that it was effective in preventing HIV transmission during vaginal intercourse. Phase II/III studies of BufferGel® and PRO 2000® were conducted in India, Malawi, South Africa, United Republic of Tanzania, and Zimbabwe. Unfortunately, little or no protective effects were seen. In 2010, however, the CAPRISA 004 microbicide trial established that 1% tenofovir (a nucleotide analogue reverse transcriptase inhibitor) gel reduced women's risk of acquiring HIV from their male sexual partners by an estimated 39% overall. Anthony Fauci, Director of the U.S. National Institute of Allergy and Infectious Disease (NIAID), called these findings 'an exciting scientific achievement that moves us one step forward to gaining another effective tool to prevent HIV infection'. However, since it is very unlikely that this single approach will universally work, other avenues need to be pursued for HIV prevention, including basic scientific strategies designed to discover or create novel agents that may lead to finding a cure for HIV/AIDS.

Candidates for scientific study and possibly pharmacological development are mannose-binding lectins. The first lectin that was discovered to exhibit potent HIV-inactivating properties was Cyanovirin-N (CV-N), a 101 amino acid protein. It was

originally isolated in the Laboratory of Dr. Michael R. Boyd (NCI) from an aqueous extract of the cyanobacterium *Nostoc ellipsosporum* in a screen of materials in the U.S. National Cancer Institute's (NCI) Natural Products Repository [1]. The protein's activity was linked to its specific and tight binding to the HIV envelope glycoprotein, gp120, that is pivotal for viral attachment to the host cell [2, 3]. This viral protein is remarkably rich in high mannose N-linked sugars [4] that contribute ~50% to its overall molecular weight [5]. Spurred by the finding that glycan binding by CV-N was the cause for its anti-HIV activity, the search for other lectins uncovered DC-SIGN [6] Scytovirin [7], Griffithsin [8], MVL [9], and Actinohivin [10], all of which bind to the high mannose glycans on gp120, thereby exerting anti-HIV activity. Interestingly, the binding modes and target epitopes on high-mannoses, Man-8 and Man-9, the predominant mannose glycans of gp120, are distinct for the different lectins. CV-N specifically recognizes Manα(1–2)Man linked mannose substructures, in particular the D1 or D3 arms of Man-9 [11, 12], DC-SIGN preferentially interacts with the Manα(1–3)Manα(1–6)Man trisaccharide [13], Griffithsin binds to single mannose units [14], and MVL specifically interacts with the Manα(1–6)Manβ(1–4)GlcNAcβ(1–4)GlcNAc tetrasaccharide [15]. For Scytovirin and Actinohivin, specificities towards Manα(1–2) Manα(1–6) Manα(1–6)Man [16] and Manα(1–2)Man [17], respectively, were established via sugar binding studies.

Another such protein was recently discovered (named *Oscillatory Agardhii* Agglutinin; OAA) in the cyanobacterium *Oscillatory Agardhii,* and its gene sequence permitted its placement into a phylogenetic tree composed of eight other homologous hypothetical proteins [18, 19]. We determined the three-dimensional crystal structures of OAA as well as that of a homolog from *Pseudomonas fluorescens* (named *Pseudomonas fluorescens* Agglutinin; PFA). The structures revealed a novel, compact, β-barrel-like architecture [20] with two specific binding sites for the α3,α6-mannopentaose core unit of Man-9. This makes the OAA-like lectins unique, both with respect to their structure as well as specific carbohydrate binding.

4.2 Structures

The atomic structures of OAA and PFA were determined by X-ray crystallography for proteins comprising residues A2-T133 (OAA) and S2-E133 (PFA) (see Fig. 4.1a). The first methionine is completely removed by the *E.coli* N-terminal methionine aminopeptidase during protein expression (verified by NMR and mass spectrometry). For consistency, however, we follow the numbering scheme of Sato and Hori [19].

The overall architecture of OAA and PFA is compact, β-barrel-like, and contains a continuous ten-stranded, anti-parallel β-sheet (Fig. 4.1b). Each of the amino acid sequence repeats (Fig. 4.1a) folds into five β-strands, denoted as β1 to β5 (colored in grey and green for OAA and PFA, respectively) and β6 to β10 (colored in light blue and purple for OAA and PFA, respectively). A very short linker, comprising residues G67-N69, connects the two-sequence repeats (colored in orange).

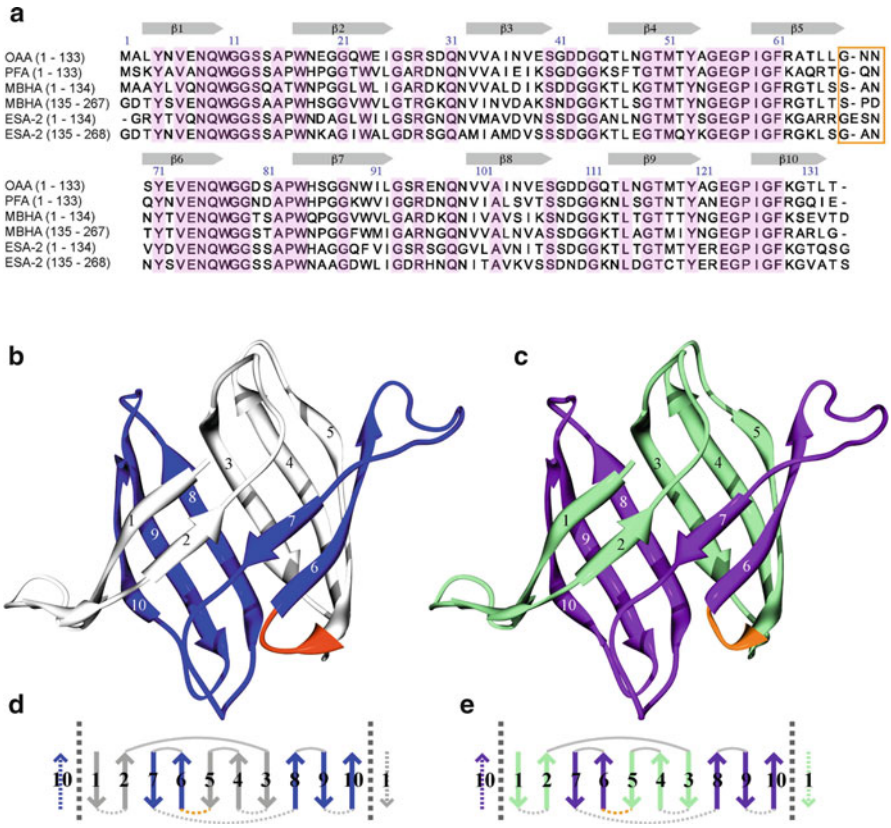


Fig. 4.1 Amino acid sequence alignment of OA-related lectins (**a**), X-ray structures of OAA (**b**) and PFA(**c**), and arrangement of strands in the β-barrels of the OAA (**d**) and PFA (**e**)

The first two β-strands of each sequence repeat (β1-β2 and β6-β7) and the next three β-strands (β3-β4-β5 and β8-β9-β10) are positioned on opposite sides of the barrel (Fig. 4.1b), and the linkers connecting strands β2 and β3, and β7 and β8, respectively, cross at the top or the bottom of the barrel. As a result, the first two β-strands (β1-β2) in the first sequence repeat are located in the barrel between the β-strands from the second repeat, namely strands β7 and β6 on one side and strands β10, β9 and β8 on the other side (Fig. 4.1b, c). Similarly, strands β3, β4 and β5 of the first repeat are flanked by β8 and β6, respectively, of the other sequence repeat (Fig. 4.1b, c). The swap of β-strands between the two sequence repeats creates an almost perfect C2 symmetric arrangement, rendering the conformation of the five β strands in each sequence repeat extremely similar. This can be easily appreciated from the backbone atomic r.m.s.d. values of 0.66 and 0.70 Å for the OAA and PFA repeats, respectively.

The β-barrel structure is stabilized through numerous aliphatic side chain interactions on the inside of the barrel that contribute to the hydrophobic core of the protein. For OAA, these are provided by residues Y4, V6, W23, I25, V33, I36, V38, L47, M51, Y53, I59, F61, A63, and L65 from the first sequence repeat and residues Y71, V73, W90, L92, V100, I103, V105, L114, M118, Y120, I126, F128, G130, and L132 from the second sequence repeat. Apart from two positions (I25/L92 and A63/G130), these residues are invariant in the two sequence repeats and they also exhibit identical χ1 and χ2 angles. For PFA, these residues are Y4, V6, W23, L25, V33, I36, I38, F47, M51, Y53, I59, F61, and A63 from the first repeat and Y71, V73, W90, I92, V100, L103, V105, L114, N118, Y120, I126, F128, and G130 from the second sequence repeat. Positions with different amino acids in the two sequence repeats are L25/I92, I36/L103, I38/V105, F47/L114, M51/N118, and A63/G130.

The 2D ^{1}H-^{15}N HSQC spectra of OAA and PFA display well-dispersed and narrow resonances, as expected for a perfectly folded protein (Fig. 4.2), and complete backbone assignments were obtained using experiments commonly used in our laboratory, such as 3D HNCACB, CBCA(CO)NH, and ^{1}H-^{15}N NOESY HSQC. All expected amide backbone resonances were observed, except for residue N69 of OAA, which is broad and cannot be detected at 25°C. Interestingly, the two amide resonances of G26 and G93 are significantly up-field shifted in their proton frequencies and resonate at 3.38/111.7 ppm and 2.93/104.2 ppm in OAA and at 3.08/107.6 ppm and 3.26/110.4 ppm in PFA, respectively, due to their positioning above the side chains of W90 and W23 (insets in the spectra). Having structures and NMR assignments available, it is easily possible to directly test for glycan binding by ^{1}H-^{15}N HSQC NMR titrations. Titration data sets were recorded for ^{15}N-labeled OAA and PFA, and spectra in the absence or presence of α3,α6-mannopentaose at 1:3 M ratio were used for chemical shift mapping (Fig. 4.2c, d). Based on this data, the glycan binding sites on the proteins can be easily delineated. α3,α6-mannopentaose binding is in slow exchange on the chemical shift scale (new bound resonances appear), suggesting relatively tight binding. In OAA, resonances with chemical shift differences of >0.1 ppm (or higher than 1x standard deviation of ~0.095) include residues L3, E7-S13, A15, W17-W23, E95, L114, T117, M118, Y120, E123-F128 in binding site 1 and residues I25, S27, R28, V34, M51-A54, E56-F61, A63, E72, Q76-D80, A82, W84, S86, G88, and L92 in binding site 2. In PFA, the affected resonances similarly belong to the corresponding residues in OAA. Note that in both proteins, site 1 is formed by loops between β-strands from the N- and C-terminal ends of the polypeptide (between β1/β2 and β9/β10), and site 2 comprises residues in the middle of the chain in loops between β4/β5 and β6/β7.

Our initial efforts aimed at crystallizing a complex between OAA and Man-9 or α3,α6-mannopentaose, either through soaking or co-crystallization, were unsuccessful. We determined that the reason for our failure was the presence of N-cyclohexyl-3-aminopropane-sulfonic acid (CAPS) in the crystallization buffer. CAPS was a critical component in the original crystallization conditions, and the solved crystal structure of OAA alone revealed two bound CAPS molecules (Fig. 4.3b). We, therefore, focused further co-crystallization efforts towards finding additional conditions for obtaining OAA crystals without any bound ligand and

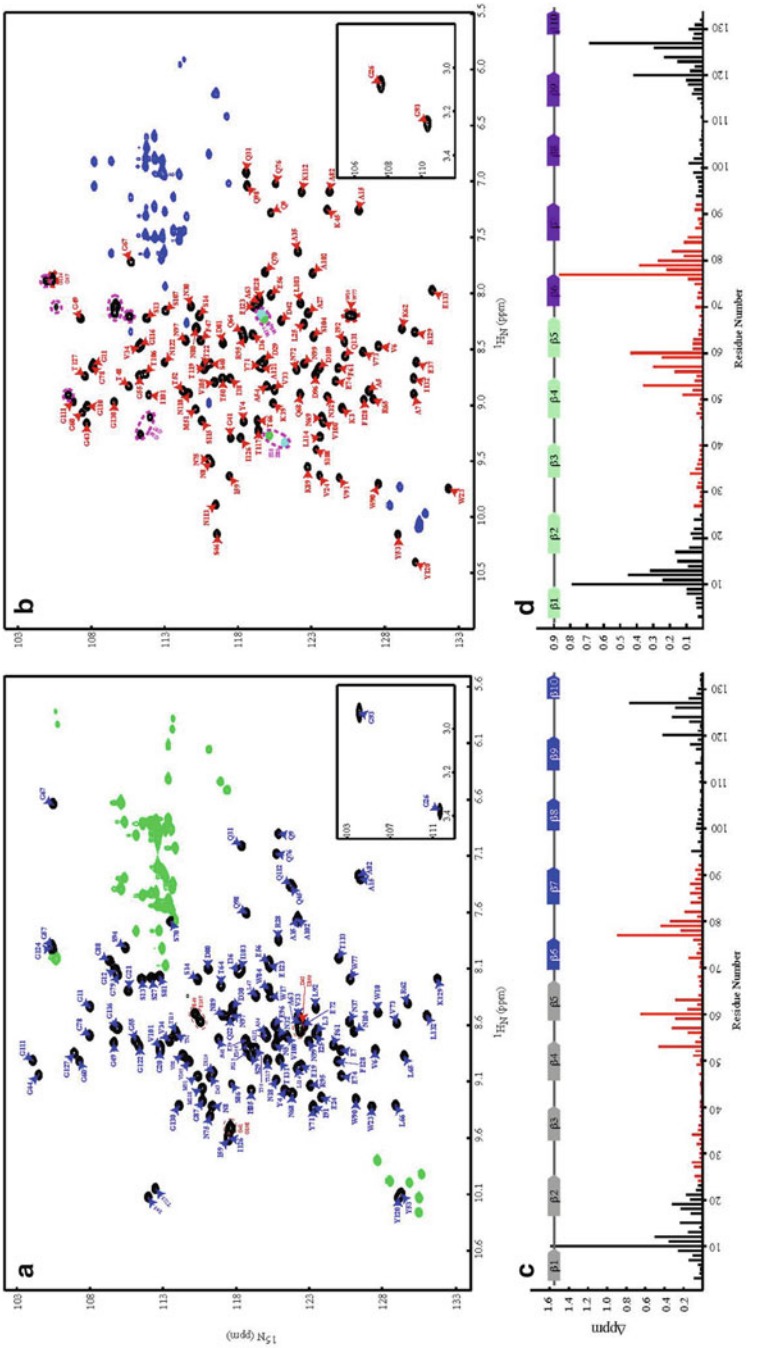


Fig. 4.2 ^{1}H-^{15}N HSQC spectra of OAA (**a**) and PFA (**b**), and chemical shift mapping of α3,α6-mannopentaose binding to OAA (**c**) and PFA (**d**)

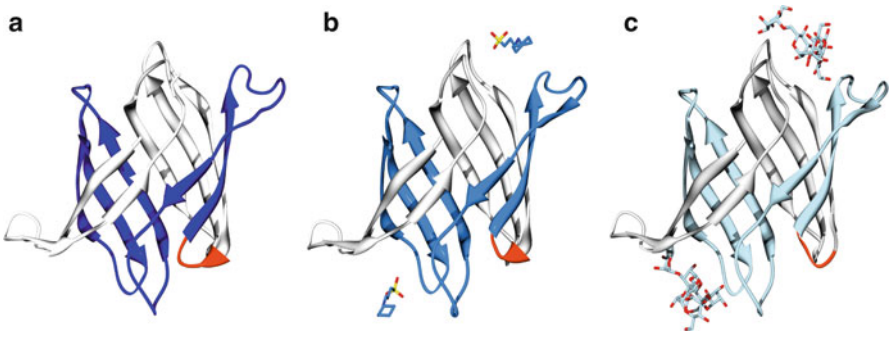


Fig. 4.3 X-ray structures of apo-OAA (**a**), CAPS-bound OAA (**b**) and a3,a6-mannopentaose-bound OAA (**c**)

complex crystals with α3,α6-mannopentaose. A modification of our previous crystallization conditions, removing CAPS, and a change in the precipitant composition (2.0 M $(NH_4)_2SO_4$ and 0.1 M Tris.HCl, pH 8.5) resulted in crystals of both apo- and α3,α6-mannopentaose-bound OAA, with the protein crystallizing in $P2_12_12_1$, with one protein molecule per asymmetric unit. This is different from the previously determined structure of CAPS-bound OAA with two protein chains per asymmetric unit in P_1. The final models of the apo-OAA structure, the structure with bound CAPS molecules, and the complex structure between OAA and α3,α6-mannopentaose were refined to 1.55 Å resolution with R = 18.2% and R_{free} = 19.6%, 1.2 Å resolution with R = 14.0% and R_{free} = 16.9%, and 1.65 Å resolution with R = 18.9% and R_{free} = 22.8%, respectively.

Superpositions between the apo-, the CAPS- and the α3,α6-mannopentaose-bound OAA structures yield r.m.s.d. values of 0.56 Å (apo vs CAPS-bound), 0.55 Å (apo vs glycan-bound), and 0.51 Å (CAPS-bound vs glycan-bound) for backbone atoms and 1.02 Å (apo vs CAPS-bound), 0.81 Å (apo vs glycan-bound), and 0.95 Å (CAPS-bound vs glycan-bound) for all heavy atoms.

The sugar binding pockets of site 1 and site 2 on OAA are very similar and the carbohydrate recognition sites comprise short clefts, residing between the loops on the surface of the protein. These two loops are in direct contact with the carbohydrate. The M3α(1–6)M4′ disaccharide unit of the α3,α6-mannopentaose is in closest contact with the protein (Fig. 4.3c), and, of the five-mannose carbohydrate moieties, the M4α(1–3)M3 disaccharide is located inside the binding pocket, while the M5′α(1–3)[M5″α(1–6)]M4′ trisaccharide unit is pointing towards solvent. In the M4α(1–3)M3 unit, the pyranose ring of M3 is stacked on top of the indole ring of tryptophan side chains in in both binding sites. The pyranose ring of M4 is flanked by the long side chains of two arginines. On the other side of the cleft, where the trisaccharide is located, the pyranose ring of M5′ is flanked by residues in the loop connecting strands β1-β2 and β6-β7 in site 1 and 2, respectively, while the pyranose rings of M5″ and M4′ are flanked by residues in the β9-β10 and β4-β5 loops for site 1 and 2, respectively. Of all the contacts in the binding sites, it appears that

the hydrophobic interaction between the aromatic side chain of W10 in site 1 and of W77 in site 2 with the pyranose ring of M3 plays a critical role. In addition, several polar interactions are also observed. In particular, hydrogen bonds between the hydroxyl groups of the carbohydrate and main chain amide groups are present, augmented by several contacts with side chains.

Comparison of the carbohydrate binding sites in the apo-structure and the sugar-bound structure reveals some interesting differences. In the free protein, the orientation of the peptide bond between W10 and G11 (binding site 1) and W77 and G78 (binding site 2) is flipped, with the carbonyl oxygens of W10 and W77 pointing in opposite directions. Upon sugar binding, the conformation in binding site 1 is essentially unchanged, i.e. the protein is already in a "bound" conformation, even in the absence of sugar, while α3,α6-mannopentaose binding to site 2 changes the orientation of the W77/G78 peptide plane from the "free" to the "bound" conformation. We, therefore, evaluated whether conformational selection or an induced fit model would best explain this difference. Relaxation data for the free and bound protein yield essentially identical values for the equivalent residues, indicating that the loops associated with sugar binding are equally flexible in both sites. In addition, we explored different temperatures for X-ray and NMR data collection, since differences between the NMR and X-ray results could have been caused by 'freezing out' conformations at cryogenic temperatures in the crystal. None of our data suggested that in the cryogenic X-ray structure the bound conformation was fortuitously selected. We, therefore, conclude that both loop regions in the free protein are flexible and that crystal packing effects around the carbohydrate binding site 1 forces the conformation of the β1-β2 loop into the conformation seen in the carbohydrate-bound structure.

All together, our combined NMR spectroscopic and X-ray crystallographic results provide the first atomic details of how the core epitope of Man-9 is recognized by the OAA-class of lectins and explain the protein's potent anti-HIV activity. In addition, our structural data adds important novel information to the growing body of knowledge about antiviral lectins that, in turn, may be exploited for glycan targeting on gp120 in the fight against HIV transmission.

Acknowledgements The authors are indebted to all past and present members of AMG's group. This work was funded by the National Institutes of Health (R01GM080642).

References

1. Boyd MR, Gustafson KR, McMahon JB, Shoemaker RH, O'Keefe BR, Mori T, Gulakowski RJ, Wu L, Rivera MI, Laurencot CM, Currens MJ, Cardellina JH 2nd, Buckheit RW Jr, Nara PL, Pannell LK, Sowder RC 2nd, Henderson LE (1997) Discovery of cyanovirin-N, a novel human immunodeficiency virus-inactivating protein that binds viral surface envelope glycoprotein gp120: potential applications to microbicide development. Antimicrob Agents Chemother 41:1521–1530
2. Freed EO, Martin MA (1995) The role of human immunodeficiency virus type 1 envelope glycoproteins in virus infection. J Biol Chem 270:23883–23886

3. Eckert DM, Kim PS (2001) Mechanisms of viral membrane fusion and its inhibition. Annu Rev Biochem 70:777–810
4. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA (1998) Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 393:648–659
5. Ji X, Chen Y, Faro J, Gewurz H, Bremer J, Spear GT (2006) Interaction of human immunodeficiency virus (HIV) glycans with lectins of the human immune system. Curr Protein Pept Sci 7:317–324
6. Pohlmann S, Baribaud F, Lee B, Leslie GJ, Sanchez MD, Hiebenthal-Millow K, Munch J, Kirchhoff F, Doms RW (2001) DC-SIGN interactions with human immunodeficiency virus type 1 and 2 and simian immunodeficiency virus. J Virol 75:4664–4672
7. Bokesch HR, O'Keefe BR, McKee TC, Pannell LK, Patterson GM, Gardella RS, Sowder RC 2nd, Turpin J, Watson K, Buckheit RW Jr, Boyd MR (2003) A potent novel anti-HIV protein from the cultured cyanobacterium Scytonema varium. Biochemistry 42:2578–2584
8. Mori T, O'Keefe BR, Sowder RC 2nd, Bringans S, Gardella R, Berg S, Cochran P, Turpin JA, Buckheit RW Jr, McMahon JB, Boyd MR (2005) Isolation and characterization of griffithsin, a novel HIV-inactivating protein, from the red alga Griffithsia sp. J Biol Chem 280:9345–9353
9. Yamaguchi M, Ogawa T, Muramoto K, Kamio Y, Jimbo M, Kamiya H (1999) Isolation and characterization of a mannan-binding lectin from the freshwater cyanobacterium (blue-green algae) Microcystis viridis. Biochem Biophys Res Commun 265:703–708
10. Chiba H, Inokoshi J, Nakashima H, Omura S, Tanaka H (2004) Actinohivin, a novel anti-human immunodeficiency virus protein from an actinomycete, inhibits viral entry to cells by binding high-mannose type sugar chains of gp120. Biochem Biophys Res Commun 316: 203–210
11. Barrientos LG, Matei E, Lasala F, Delgado R, Gronenborn AM (2006) Dissecting carbohydrate-Cyanovirin-N binding by structure-guided mutagenesis: functional implications for viral entry inhibition protein. Eng Des Sel 19:525–535
12. Sandstrom C, Berteau O, Gemma E, Oscarson S, Kenne L, Gronenborn AM (2004) Atomic mapping of the interactions between the antiviral agent cyanovirin-N and oligomannosides by saturation-transfer difference NMR. Biochemistry 43:13926–13931
13. Feinberg H, Mitchell DA, Drickamer K, Weis WI (2001) Structural basis for selective recognition of oligosaccharides by DC-SIGN and DC-SIGNR. Science 294:2163–2166
14. Ziolkowska NE, O'Keefe BR, Mori T, Zhu C, Giomarelli B, Vojdani F, Palmer KE, McMahon JB, Wlodawer A (2006) Domain-swapped structure of the potent antiviral protein griffithsin and its mode of carbohydrate binding. Structure 14:1127–1135
15. Williams DC Jr, Lee JY, Cai M, Bewley CA, Clore GM (2005) Crystal structures of the HIV-1 inhibitory cyanobacterial protein MVL free and bound to Man3GlcNAc2: structural basis for specificity and high-affinity binding to the core pentasaccharide from n-linked oligomannoside. J Biol Chem 280:29269–29276
16. McFeeters RL, Xiong C, O'Keefe BR, Bokesch HR, McMahon JB, Ratner DM, Castelli R, Seeberger PH, Byrd RA (2007) The novel fold of scytovirin reveals a new twist for antiviral entry inhibitors. J Mol Biol 369:451–461
17. Tanaka H, Chiba H, Inokoshi J, Kuno A, Sugai T, Takahashi A, Ito Y, Tsunoda M, Suzuki K, Takenaka A, Sekiguchi T, Umeyama H, Hirabayashi J, Omura S (2009) Mechanism by which the lectin actinohivin blocks HIV infection of target cells. Proc Natl Acad Sci USA 106:15633–15638
18. Sato Y, Okuyama S, Hori K (2007) Primary structure and carbohydrate binding specificity of a potent anti-HIV lectin isolated from the filamentous cyanobacterium Oscillatoria agardhii. J Biol Chem 282:11021–11029
19. Sato T, Hori K (2009) Cloning, expression, and characterization of a novel anti-HIV lectin from the cultured cyanobacterium. Oscillatoria agardhii Fish Sci 75:743–753
20. Koharudin LM, Furey W, Gronenborn AM (2011) Novel fold and carbohydrate specificity of the potent anti-hiv cyanobacterial lectin from Oscillatoria Agardhii. J Biol Chem 286: 1588–1597

Chapter 5
Fluorescence Resonance Energy Transfer Studies of Structure and Dynamics in Nucleic Acids

David M.J. Lilley

Abstract Fluorescence spectroscopy is highly sensitive, and can be performed on single molecules. Using fluorescence resonance energy transfer (FRET) distances can be estimated in biological macromolecules. This has provided significant structural and dynamic information of DNA and RNA molecules. It has even been able to allow us to observe the catalytic function of a ribozyme in real time.

5.1 Fluorescence Resonance Energy Transfer

Fluorescence resonance energy transfer (FRET) [1, 2] provides a spectroscopic way of estimating distances over a size range that is suitable for biological macromolecules. It is based upon fluorescence, a highly sensitive spectroscopic method. Because of this it can be performed on single molecules, and this has greatly expanded the importance of FRET in biology in the last decade.

Molecules interact with the electric component of light because of the change in electronic distribution between the ground state and an excited state, measured by the transition dipole moment vector. Thus the vector for a transition from states m to n ($\hat{d}$) can be written in Dirac bra.ket notation as:

$$\hat{d} = \langle \Psi_n | \Re | \Psi_m \rangle \tag{5.1}$$

where Ψ_m and Ψ_n are the wavefunctions of the initial and final states, and $\Re$ is the dipole moment operator. A molecule that is in an excited electronic state due to

D.M.J. Lilley (✉)
Cancer Research UK Nucleic Acid Structure Research Group, The University of Dundee, MSI/WTB Complex, Dundee DD1 5EH, UK
e-mail: d.m.j.lilley@dundee.ac.uk

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_5,

absorption of a photon rapidly loses energy to reach the lowest vibrational level of that state. It can then emit a photon of longer wavelength than the one that excited it to reach the ground electronic state. This is fluorescent emission, and the loss of vibrational energy is observed as the Stokes shift.

The transition dipoles of two different fluorescent molecules may interact together in a process that leads to transfer of energy from one (the donor) to the other (the acceptor). Resonance energy transfer arises from the dipolar coupling between the oscillating transition dipoles of the donor and acceptor fluorophores. FRET is strongly dependent upon the physical separation between the two, and by tethering two small-molecule fluorescent probes (fluorophores) to a biomolecule of interest at known positions, we can monitor the distance between these two points.

FRET is a resonance between singlet-singlet electronic transitions of the donor and acceptor fluorophores, giving a transfer of excitation energy from the donor to the acceptor. It results from dipolar coupling between the emission transition moment of the donor and the absorption transition moment of the acceptor. FRET can be observed in a number of ways. These include a lowering of the fluorescent quantum yield of the donor and a corresponding shortening of the lifetime of the excited state of the donor. It also leads to an increased fluorescent emission from the acceptor (if fluorescent, so-called sensitized emision).

The rate of the energy transfer process depends on the inverse sixth power of the distance between the two fluorophores [2], and this is the basis of the use of the technique to provide structural information. In practical terms, the efficiency of energy transfer (E_{FRET}) is normally determined. This is the proportion of donor excitation events that lead to excitation of the acceptor by dipolar coupling – a quantum yield of FRET.

$$E_{\mathrm{FRET}} = 1 - \tau_{\mathrm{DA}} / \tau_{\mathrm{D}} = \left\{1 + (r_{\mathrm{DA}} / R_0)^6\right\}^{-1} \tag{5.2}$$

where τ_{DA} and τ_{D} are the fluorescent lifetime in the presence and absence of the acceptor respectively, r_{DA} is the separation between the donor and acceptor [3]. This results in the dependence on distance shown graphically in Fig. 5.1.

R_0 is the Förster length for a given donor-acceptor pair of fluorophores, given by:

$$R_0^6 = 0.529 \cdot \Phi_D \cdot \kappa^2 \cdot J(\lambda) \,/\, N \cdot n^4 \tag{5.3}$$

where the units of R_0 and the wavelength λ are in cm. Φ_{D} is the fluorescent quantum yield of the donor in the absence of the acceptor and N is Avogadro's number. κ depends on the relative orientation of the donor and acceptor transition moments, defined below. n is the refractive index of the medium in which the electric fields of the transition dipoles extend, and $J(\lambda)$ is the normalized spectral overlap between donor emission and acceptor absorption, given by:

$$J(\lambda) = \int_0^\infty \phi_D(\lambda) \cdot \varepsilon_A(\lambda) \cdot \lambda^4 d\lambda \Big/ \int_0^\infty \phi_D(\lambda) d\lambda \tag{5.4}$$

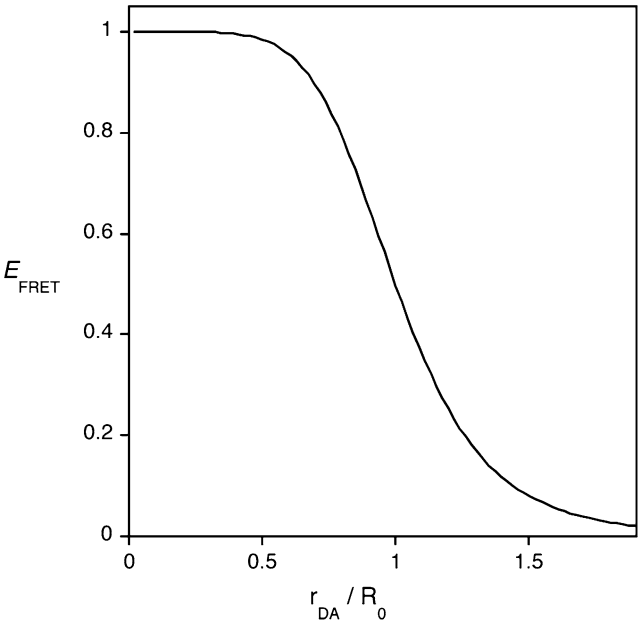


Fig. 5.1 Plot of FRET efficiency (E_{FRET}) as a function of inter-fluorophore separation (r_{DA}) in units of Förster length R_0, assuming that R_0 is constant. This is calculated using Eq. (5.2)

where $\boldsymbol{\phi}_{\mathrm{D}}$ is the spectral shape function for the donor emission and ε_{A} is that for acceptor absorption (in $\mathrm{M}^{-1}\ \mathrm{cm}^{-1}$). From Eq. (5.2) it can be seen that when $r_{\mathrm{DA}} = R_0$, the efficiency of FRET is 0.5. R_0 values are frequently calculated on the basis of an assumption (stated or otherwise) that $\kappa^2 = 2/3$, and this is sometimes written as R_0 (2/3).

However, the orientation dependence of energy transfer may be both a complication for the analysis of FRET, and a potential source of structural information [3]. The rate at which the acceptor is excited by FRET is proportional to the square of the scalar product of its transition dipole with the local electric field of the donor transition dipole. This is the origin of the κ^2 term in Eq. (5.3), and is given by:

$$\begin{aligned} \kappa^2 &= (\hat{P}_D \bullet \hat{P}_A - 3 \cdot (\hat{P}_D \bullet \hat{r}_{DA}) \cdot (\hat{r}_{DA} \bullet \hat{P}_A))^2 \\ &= (\cos \Theta_T - 3 \cdot \cos \Theta_D \cdot \cos \Theta_A)^2 \end{aligned} \tag{5.5}$$

where $\hat{P}_D$ and $\hat{P}_A$ are the donor and acceptor transition dipole moment vectors, $\hat{r}_{DA}$ is the vector between their centers, and the angles Θ_{T}, Θ_{D} and Θ_{A} are defined in Fig. 5.2a. κ^2 can take values between 0 and 4, as shown in Fig. 5.2b. If the value is not known, it becomes very hard to extract distances from FRET efficiencies. However, if at least one fluorophore is flexible, so that it experiences many orientations during the lifetime of the excited state, then κ^2 should average

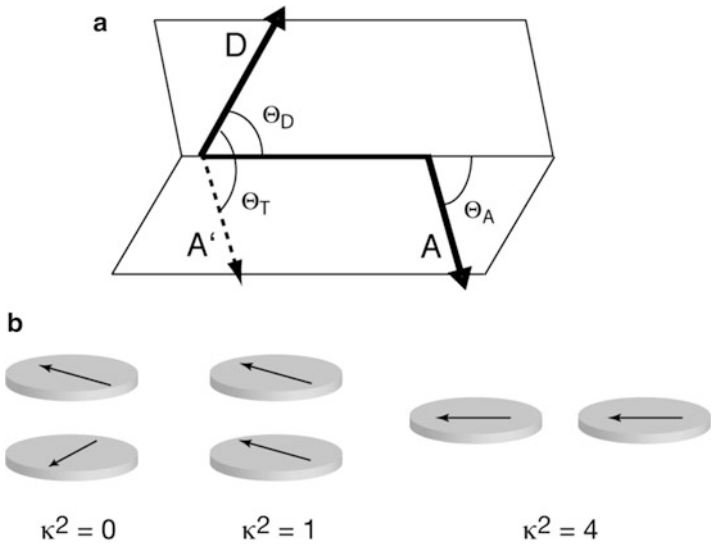


Fig. 5.2 The orientation dependence of FRET. (**a**) definition of the κ^2 parameter. The transition moments of the donor and acceptor are indicated by **D** and **A**, and **A′** is a translation of vector **A** to the origin of **D**. (**b**): Some specific relative orientations of the donor and acceptor fluorophores that lead to integral values of κ^2

to 2/3. Where fluorescein is tethered to a phosphate group of RNA via a flexible linker this is probably a good approximation. The negative charge is repelled by the backbone so that the fluorophore freely rotates in a cone, and the fluorescent anisotropy of fluorescein attached to DNA is typically low (usually ~ 0.1).

By contrast, the cyanine fluorophores (fluorophores that are popular for single-molecule FRET studies) interact strongly with DNA and RNA. We have shown that both Cy3 [4] and Cy5 [5] when attached to 5′-termini via C_3 linkers (as often generated when coupled as phosphoramidites during synthesis) stack upon the ends of double-helical DNA very much in the manner of an additional basepair. Using a series of DNA and DNA-RNA duplexes we demonstrated a modulation of FRET efficiency due to the changing relative orientation of the fluorophore transition moments due to the helical periodicity [6] (Fig. 5.3). The data could be simulated by a model based on the geometric properties of standard B- and A-form helices (corresponding to DNA and RNA-DNA respectively), and the positions of the fluorophores determined by NMR with a significant lateral averaging [6], with or without a fraction of unstacked fluorophore as indicated by time-resolved studies [6, 7]. This is a clear confirmation of the expected orientational dependence of Förster resonance energy transfer. It also provides a warning that such effects could significantly affect the interpretation of FRET data in terms of distance with a commonly-used pair of fluorophores. Within the duplex species, distances could be in error by as much as 12 Å if it is assumed that $\kappa^2 = 2/3$. If the fluorophores are not constrained to lie in parallel planes the errors could be even greater, because κ^2 could reach a value of four under these circumstances.

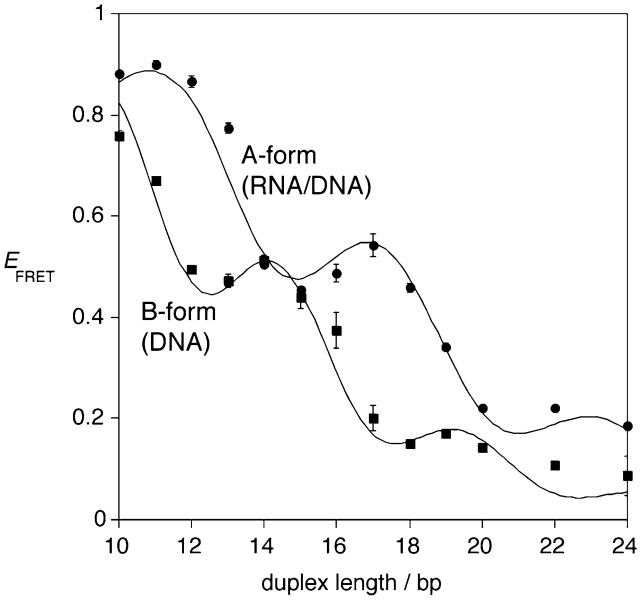


Fig. 5.3 Efficiency of energy transfer for Cy3 (donor), Cy5 (acceptor)-labelled hybrid DNA and RNA/DNA duplexes as a function of duplex length [6]. E_{FRET} was measured for each duplex species as phospholipid vesicle-encapsulated single molecules. The E_{FRET} values are plotted for the DNA (*squares*) and DNA/RNA duplexes (*circles*) as a function of helix length, with the estimated errors. The *lines* show simulation of the data, using a model in which the fluorophores were mainly stacked onto the helix undergoing lateral motion. For the DNA duplexes this was based on standard B geometry with a periodicity 10.5 bp/turn and a helical rise of 3.6 Å/bp step; 31% of the fluorophore was allowed to be freely mobile (based on time resolved analysis) with $\kappa^2 = 2/3$, while the remaining fluorophore underwent lateral motion with a Gaussian half-width of 42°. For the DNA/RNA duplexes the simulation was based on standard A geometry with a periodicity 12 bp/turn and a helical rise of 3 Å/bp step; 12% of the fluorophore was allowed to be freely mobile (based on time resolved analysis) with $\kappa^2 = 2/3$, while the remaining fluorophore underwent lateral motion with a Gaussian half-width of 42°. The fluorescent quantum yield for Cy3 was 0.30 attached to DNA and 0.35 attached to DNA/RNA

5.2 Dynamic Conformer Exchange in the Four-Way DNA Junction

The four-way DNA (Holliday) junction is the central intermediate of genetic recombination. In the presence of metal ions the junction folds by the pairwise coaxial stacking of helical arms [8], to adopt the stacked X-structure [9]. This structure was first demonstrated using comparative gel electrophoresis, followed shortly by steady-state fluorescence in what was probably the first paper in the modern era of the application of FRET to nucleic acids [10]. The structure of the junction in free solution was found to be antiparallel, in contrast to earlier assumptions and the commonly-found parallel depiction in text books. The detailed

structure has been subsequently confirmed by X-ray crystallography [11–13]. More recent single-molecule studies showed that the probability of parallel forms of the junction with a lifetime greater than 1 ms is less than 2×10^{-5} [14].

Formation of the stacked X-structure lowers the symmetry from four- to two-fold, and there are therefore two conformers of the structure possible, differing in the choice of stacking partner. This suggested that the two forms might interconvert, and this was supported by somewhat indirect studies using *Mbo*II cleavage of junctions [15, 16], and by NMR, showing that a minor conformer existed for one junction that was in rapid exchange on the NMR timescale [17].

Since it is not possible to synchronize conformer exchange, we were unable to confirm this process directly before the advent of single-molecule methods. However, FRET studies of single junction molecules showed that exchange between junction conformers [18] does occur, at a surprisingly slow rate in the presence of 50 mM Mg^{2+} ions (Fig. 5.4). The rate of conformer exchange was found to increase as the ionic strength of the solution was lowered [14]. This indicates that the transition state is destabilized by metal ions, suggesting that it resembles the open form of the junction that exists in the absence of metal ions. The lifetime of the open species was estimated to be <1 ms.

Conformer exchange has been recently studied by the application of stretching force [19] in single-molecule FRET experiments. It was possible to determine the position of the transition state along the reaction coordinate by measuring the gradient of the rate of conformer exchange rates as a function of applied force. Using relatively low forces (<4 pN) we determined that the open structure is a shallow intermediate, flanked by transition states on either side. The application of stretching force tilts the entire energy landscape, thereby altering the current rate determining step.

5.3 Helical Junctions in RNA

Helical junctions exert a major influence on the 3D structures of complex RNA species [8]. Their geometry orients helices relative to one another, allowing long-range tertiary contacts to occur. Analysis of the structures of the 16S and 23S rRNA species within the ribosome reveals many helical junctions [20–24]. The structure of some autonomously-folding smaller RNA species can be almost entirely determined by component junctions [25, 26], as exemplified by the VS ribozyme where the structural core of the ribozyme comprises five helical segments related by two three-way junctions [27, 28]. Helical junctions can also generate the catalytic centre of ribozymes, such as the hammerhead ribozyme [29], and ligand binding pockets of great selectivity as found in a number of riboswitches [30–32].

Perfect four-way (4H) RNA junctions are more polymorphic than their DNA equivalents [33, 34]. The 4H junction derived from the hairpin ribozyme was studied by single-molecule fluorescence spectroscopy [35]. In the presence of Mg^{2+} ions, the junction was found to sample both parallel and antiparallel conformations and

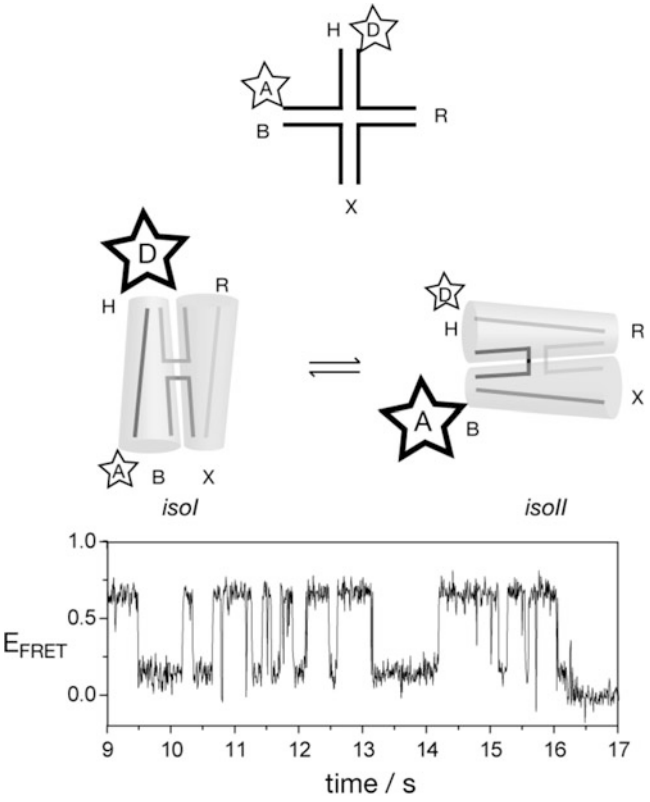


Fig. 5.4 Conformer exchange in a four-way DNA junction. The junction comprises four helical arms, labelled B, H, R and X. Donor and acceptor fluorophores were attached to the 5′-termini of the B and H arms, so that there is low FRET efficiency in the *iso*I (*left*), but high FRET efficiency in the *iso*II conformer (*right*). A record of FRET efficiency (E_{FRET}) as a function of time for a single junction molecule shows repeated interconversion between the low efficiency ($E_{\mathrm{FRET}} = 0.1$) and the high efficiency ($E_{\mathrm{FRET}} = 0.6$) states [18]

both stacking conformers, with a bias towards one antiparallel stacking conformer. A continual interconversion occurred between the various conformers, at rates of several transitions per second under physiological conditions. The rate of interconversion became slower at higher Mg^{2+} ion concentrations, suggesting that interconversion proceeds via an open intermediate.

Perfectly-paired 4H RNA junctions are actually quite rare. Most natural four-way RNA junctions contain one or more formally unpaired nucleotides at the point of strand exchange. The structural and dynamic properties would be expected to be altered by the extra nucleotides. The hepatitis C-virus (HCV) RNA contains an IRES element that allows initiation of translation to occur independently of a 5-terminus with a CAP. Its structure is based on a number of junctions, including a $2HS_2 2HS_1$ four-way junction. In contrast to 4H RNA junctions, the IRES junction adopted an

extended structure in the absence of added metal ions. On addition of divalent metal ions the junction folded by pairwise coaxial stacking of arms, adopting the stacking conformer that places the extra nucleotides onto the exchanging strands [36].

5.4 Structure of a Three-Way Helical Junction in the HCV IRES Element

Very few natural three-way junctions in RNA are perfectly-paired 3H junctions. Most include one or more formally unpaired nucleotides in the linker regions connecting the helices. Such junctions are extremely common in natural RNA molecules.

The HCV IRES contains a formally $3HS_4$ helical junction that determines the trajectory of helix IIId that interacts with the 40S ribosomal subunit (Fig. 5.5). In principal the junction has the possibility of adopting two alternative stacking conformers. In addition the junction could also undergo two steps of branch migration that would form $2HS_1HS_3$ and $2HS_2HS_2$ junctions. Comparative gel electrophoresis and ensemble FRET studies showed that the junction is induced to fold by the presence of Mg^{2+} ions in low micromolar concentrations, and suggest that the structure adopted is based on coaxial stacking of the two helices that do not terminate in a hairpin loop [37]. Single-molecule FRET studies confirm

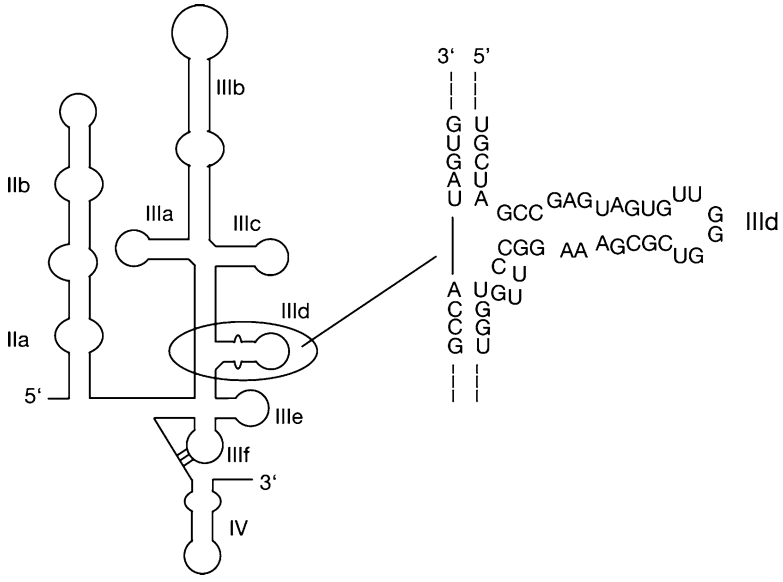


Fig. 5.5 A schematic illustration of the secondary structure of the HCV IRES. The sequence of the three-way helical junction (*circled*) is shown

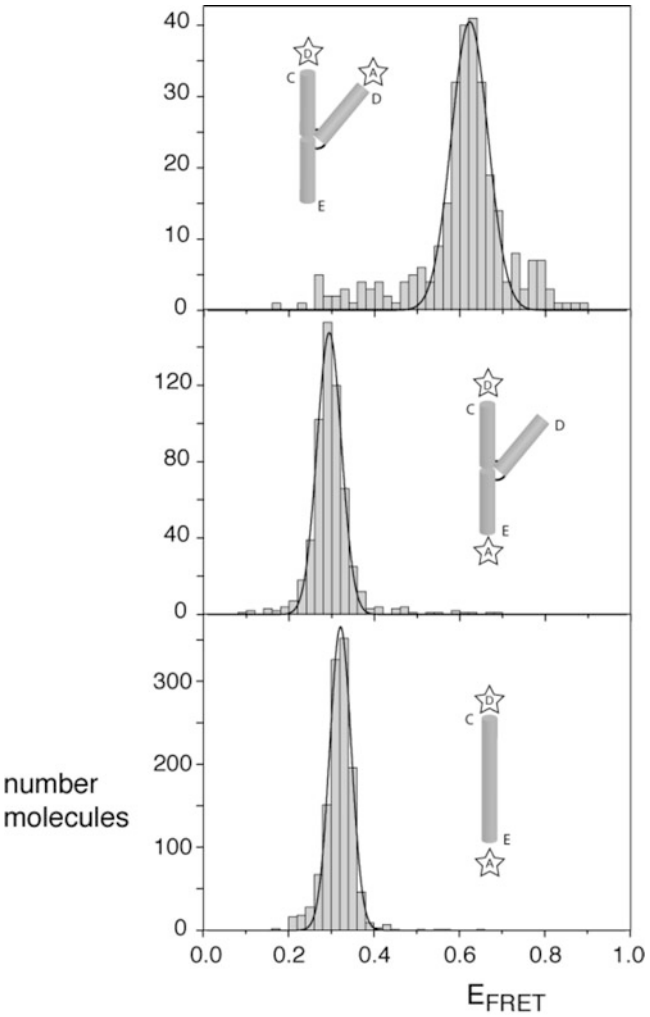


Fig. 5.6 Population distributions of FRET efficiencies for Cy3-Cy5-labelled IRES three-way junctions studied as single-molecules. Two different end-labelled vectors are shown, together with a Cy3-Cy5-labelled duplex that provides a model for coaxially-stacked C and E arms. The different species were encapsulated in phospholipid vesicles and studied by total internal reflection fluorescence microscopy. Hundreds of single molecules were studied for each species, and fluorescent intensities at Cy3 and Cy5 emission wavelengths recorded for a number of minutes with 100 ms resolution. These were plotted as the histograms shown, that are fitted to Gaussian distributions

this conclusion, and indicate that there is no minor conformer present based an alternative choice of helical stacking partners (Fig. 5.6). Moreover, analysis of single-molecule FRET data at 8 ms resolution fail to reveal evidence for structural transitions. It seems probable that this junction adopts a single conformation as a unique and stable fold.

5.5 Cleavage and Ligation Cycles in the Hairpin Ribozyme

The hairpin ribozyme is one of the nucleolytic ribozymes [38]. It comprises two internal loops that are present on adjacent arms of a four-way helical junction (Fig. 5.7). All the key nucleotides required for activity reside in the loops. It was surmised that the loops would interact to generate the local environment in which catalysis could occur, and this was confirmed in solution using FRET between fluorophores attached to the ends of the two arms [25, 39]. A minimal version of the ribozyme (termed the hinged form) in which the four-way junction is replaced by a simple phosphodiester linkage is active [40], but requires a 1,000-fold higher Mg^{2+} ion concentration to induce folding [41, 42]. Moreover, the internal equilibrium between cleavage and ligation is shifted relative to the natural form [43]. The crystal structure of the hairpin ribozyme in its four-way junction form was solved by Ferré d'Amaré [44]. In agreement with the earlier solution studies [25], the junction exhibited coaxial stacking of helices A on D and B on C, and was rotated in an antiparallel manner to allow intimate association between the two loops.

The four-way helical junction of the hairpin ribozyme spontaneously adopts the conformation that juxtaposes the loops of the A and B arms [45, 46], allowing the ribozyme to fold in physiological ionic conditions [25, 39, 41, 46–48].

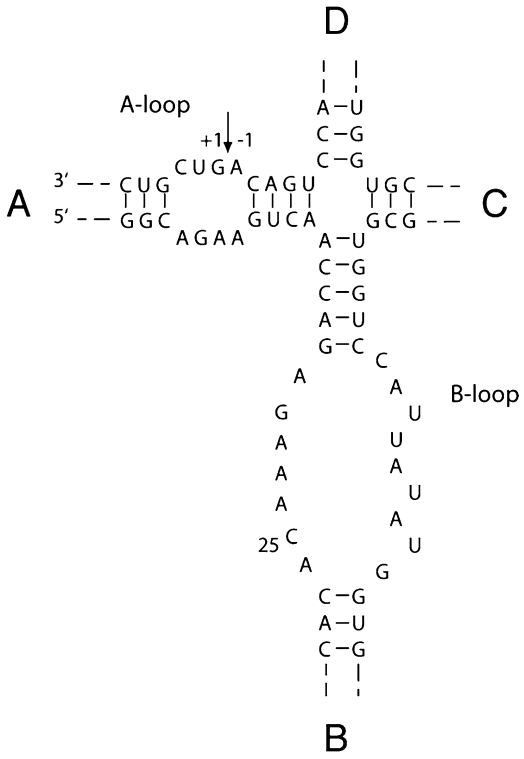


Fig. 5.7 The nucleotide sequence of the hairpin ribozyme. The ribozyme is shown in its junction form, with arms sequentially labelled A through D around the four-way junction. The position of cleavage or ligation within loop A is indicated by the *arrow*. In the folded conformation of the ribozyme, helices A and D are coaxially stacked, as are helices B and C

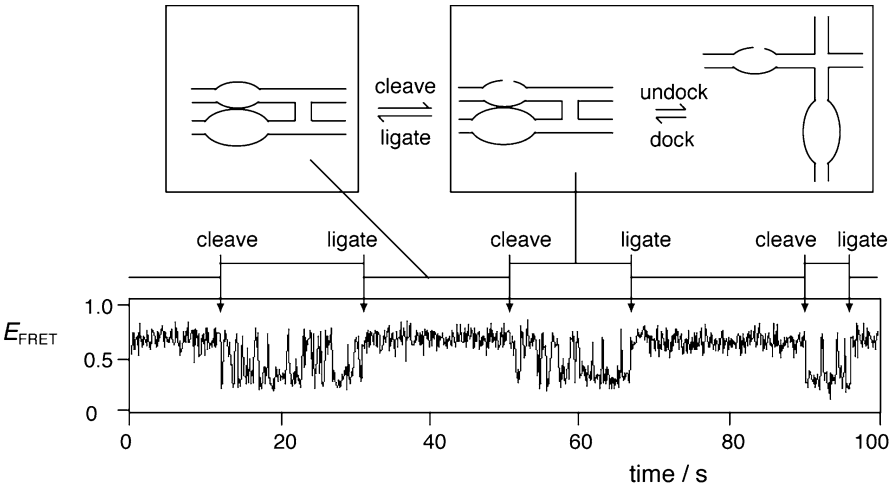


Fig. 5.8 Multiple cycles of cleavage and ligation in a single hairpin ribozyme in the presence of 1 mM Mg^{2+}. The ribozyme has a terminal helix of arm A comprising 7 bp, ensuring retention of the product of cleavage that can serve as the substrate for ligation. The ligated form of the ribozyme is stably docked under the conditions of the experiment. Upon cleavage the product ribozyme can undergo multiple docking/undocking transitions. Since the molecule was generated by ribozyme cleavage it has a cyclic 2′3′ phosphate, and is therefore competent in ligation. Ligation events may be identified by the change in dynamics back to a stable docked conformation

Single-molecule studies showed that the junction was structurally very polymorphic, and the structure observed in bulk reflected rapid interconversion between several different conformations [35].

Single-molecule FRET has been used to follow the cleavage and ligation reactions in the hairpin ribozyme, by exploiting the distinct dynamics of the intact ribozyme and its cleaved product [49]. Individual ribozyme molecules can be observed switching between two dynamic modes (Fig. 5.8). In the ligated state the molecule remains stably docked for a period before undergoing a cleavage reaction, whereupon it undergoes rapid docking and undocking. A ligation reaction then leads to the restoration of the stable docked state. The docking and undocking rates within the cleaved state are $k^C{}_{\text{dock}} = 2.5\ \text{s}^{-1}$ and $k^C{}_{\text{undock}} = 2.3\ \text{s}^{-1}$; these rates are significantly slower than the junction dynamics under the same conditions [46]. Using data from many single molecules gave the internal conversion rates for the ribozyme. The cleavage rate (k_C) is $\sim 0.6\ \text{min}^{-1}$ in the presence of 1 mM Mg^{2+}, while the ligation rate (k_L : the rate of ligation in the docked ribozyme) is $\sim 0.3\ \text{s}^{-1}$. Thus the reaction is biased towards ligation, with an internal equilibrium constant of

$$K_{\text{int}} = k_L/k_C = 34$$

This property maintains the integrity of the circular (−) strand allowing it to act as a template for (+) strand synthesis. However, cleavage of the concatenated product of replication is made possible by of the rapid undocking that follows cleavage.

Acknowledgements I thank my co-workers in Dundee over a number of years, especially Tim Wilson, Jo Ouellet, Alastair Murchie and Asif Iqbal, and our collaborators in the University of Illinois Taekjip Ha, Sean McKinney, Sungchul Hohng and Michelle Nahas. We thank Cancer Research UK for financial support.

Note Added in Proof In the 2 years since this article was written there has been significant progress in the analysis of orientation effects using the cyanine fluorophores attached to double-stranded nucleic acids. It was found that the orientation effects were not lost when Cy3 and Cy5 were attached by long, 13-atom tethers, suggesting that these fluorophores possess an intrinsic tendency to undergo terminal stacking (A). However, repeating the experiment with a series of DNA duplexes of varying length led to a phase shift (e.g. compared to the data shown in Fig. 5.3) that was interpreted in terms of a reorientation of each fluorophore by 30°, so that the transition moment lay parallel to the long axis of the terminal basepair. This was subsequently confirmed in NMR studies (B), providing proof of principle that such FRET data can provide reliable orientational information in nucleic acids.

(A) Ouellet J, Schorr S, Iqbal A, Wilson TJ, Lilley DMJ (2011) Orientation of cyanine fluorophores terminally attached to DNA via long, flexible tethers. Biophys J 101:1148–1154

(B) Urnavicius L, McPhee SA, Lilley DMJ, Norman DG (2012) The structure of sulfoindocarbocyanine 3 terminally attached to dsDNA via a long, flexible tether. Biophys J 102:561–569

References

1. Perrin F (1932) Théorie quantique des transferts d'activation entre molécules de méme espèce. Cas des solutions fluorescentes. Ann Chim Phys 17:283–314
2. Förster T (1948) Zwischenmolekulare Energiewanderung und Fluoreszenz. Ann Phys 2:55–75
3. Lilley DMJ (2009) The structure and folding of branched RNA analyzed by fluorescence resonance energy transfer. Method Enzymol 469:159–187
4. Norman DG, Grainger RJ, Uhrin D, Lilley DMJ (2000) The location of Cyanine-3 on double-stranded DNA; importance for fluorescence resonance energy transfer studies. Biochemistry 39:6317–6324
5. Iqbal A, Wang L, Thompson KC, Lilley DMJ, Norman DG (2008) The structure of cyanine 5 terminally attached to double-stranded DNA: implications for FRET studies. Biochemistry 47:7857–7862
6. Iqbal A, Arslan S, Okumus B, Wilson TJ, Giraud G, Norman DG, Ha T, Lilley DMJ (2008) Orientation dependence in fluorescent energy transfer between Cy3 and Cy5 terminally-attached to double-stranded nucleic acids. Proc Natl Acad Sci USA 105:11176–11181
7. Sanborn ME, Connolly BK, Gurunathan K, Levitus M (2007) Fluorescence properties and photophysics of the sulfoindocyanine Cy3 linked covalently to DNA. J Phys Chem B 111:11064–11074
8. Lilley DMJ (2000) Structures of helical junctions in nucleic acids. Q Rev Biophys 33:109–159
9. Duckett DR, Murchie AIH, Diekmann S, von Kitzing E, Kemper B, Lilley DMJ (1988) The structure of the Holliday junction and its resolution. Cell 55:79–89
10. Murchie AIH, Clegg RM, von Kitzing E, Duckett DR, Diekmann S, Lilley DMJ (1989) Fluorescence energy transfer shows that the four-way DNA junction is a right-handed cross of antiparallel molecules. Nature 341:763–766
11. Ortiz-Lombardía M, González A, Erijta R, Aymamí J, Azorín F, Coll M (1999) Crystal structure of a DNA Holliday junction. Nat Struct Biol 6:913–917
12. Eichman BF, Vargason JM, Mooers BHM, Ho PS (2000) The Holliday junction in an inverted repeat DNA sequence: sequence effects on the structure of four-way junctions. Proc Natl Acad Sci USA 97:3971–3976

13. Thorpe JH, Gale BC, Teixeira SC, Cardin CJ (2003) Conformational and hydration effects of site-selective sodium, calcium and strontium ion binding to the DNA Holliday junction structure d(TCGGTACCGA)$_4$. J Mol Biol 327:97–109
14. Joo C, McKinney SA, Lilley DMJ, Ha T (2004) Exploring rare conformational species and ionic effects in DNA Holliday junctions using single-molecule spectroscopy. J Mol Biol 341:739–751
15. Murchie AIH, Portugal J, Lilley DMJ (1991) Cleavage of a four-way DNA junction by a restriction enzyme spanning the point of strand exchange. EMBO J 10:713–718
16. Grainger RJ, Murchie AIH, Lilley DMJ (1998) Exchange between stacking conformers in a four-way DNA junction. Biochemistry 37:23–32
17. Carlström G, Chazin WJ (1996) Sequence dependence and direct measurement of crossover isomer distribution in model Holliday junctions using NMR spectroscopy. Biochemistry 35:3534–3544
18. McKinney SA, Déclais A-C, Lilley DMJ, Ha T (2003) Structural dynamics of individual Holliday junctions. Nat Struct Biol 10:93–97
19. Hohng S, Zhou R, Nahas MK, Yu J, Schulten K, Lilley DMJ, Ha T (2007) Fluorescence-force spectroscopy maps two-dimensional reaction landscape of the Holliday junction. Science 318:279–283
20. Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V (2000) Structure of the 30S ribosomal subunit. Nature 407: 327–339
21. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 289:905–920
22. Lescoute A, Westhof E (2006) Topology of three-way junctions in folded RNAs. RNA 12: 83–93
23. de la Pena M, Dufour D, Gallego J (2009) Three-way RNA junctions with remote tertiary contacts: a recurrent and highly versatile fold. RNA 15:1949–1964
24. Laing C, Schlick T (2009) Analysis of four-way junctions in RNA structures. J Mol Biol 390:547–559
25. Murchie AIH, Thomson JB, Walter F, Lilley DMJ (1998) Folding of the hairpin ribozyme in its natural conformation achieves close physical proximity of the loops. Mol Cell 1:873–881
26. Walter F, Murchie AIH, Thomson JB, Lilley DMJ (1998) Structure and activity of the hairpin ribozyme in its natural junction conformation; effect of metal ions. Biochemistry 37: 14195–14203
27. Lafontaine DA, Norman DG, Lilley DMJ (2002) The global structure of the VS ribozyme. EMBO J 21:2461–2471
28. Lipfert J, Ouellet J, Norman DG, Doniach S, Lilley DMJ (2008) The complete VS ribozyme in solution studied by small-angle X-ray scattering. Structure 16:1357–1367
29. Martick M, Horan LH, Noller HF, Scott WG (2008) A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. Nature 454:899–902
30. Serganov A, Yuan YR, Pikovskaya O, Polonskaia A, Malinina L, Phan AT, Hobartner C, Micura R, Breaker RR, Patel DJ (2004) Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. Chem Biol 11:1729–1741
31. Garst AD, Heroux A, Rambo RP, Batey RT (2008) Crystal structure of the lysine riboswitch regulatory mRNA element. J Biol Chem 283:22347–22351
32. Serganov A, Huang L, Patel DJ (2008) Structural insights into amino acid binding and gene control by a lysine riboswitch. Nature 455:1263–1267
33. Duckett DR, Murchie AIH, Lilley DMJ (1995) The global folding of four-way helical junctions in RNA, including that in U1 snRNA. Cell 83:1027–1036
34. Walter F, Murchie AIH, Duckett DR, Lilley DMJ (1998) Global structure of four-way RNA junctions studied using fluorescence resonance energy transfer. RNA 4:719–728
35. Hohng S, Wilson TJ, Tan E, Clegg RM, Lilley DMJ, Ha T (2004) Conformational flexibility of four-way junctions in RNA. J Mol Biol 336:69–79

36. Melcher SE, Wilson TJ, Lilley DMJ (2003) The dynamic nature of the four-way junction of the hepatitis C virus IRES. RNA 9:809–820
37. Ouellet J, Melcher SE, Iqbal A, Ding Y, Lilley DMJ (2010). Structure of the three-way helical junction of the hepatitis C virus IRES element. RNA 16:1597–1609
38. Lilley DMJ, Eckstein F (eds) (2008) Ribozymes and RNA catalysis. Royal Soc. Chemistry, Cambridge
39. Walter NG, Burke JM, Millar DP (1999) Stability of hairpin ribozyme tertiary structure is governed by the interdomain junction. Nat Struct Biol 6:544–549
40. Berzal-Herranz A, Simpson J, Chowrira BM, Butcher SE, Burke JM (1993) Essential nucleotide sequences and secondary structure elements of the hairpin ribozyme. EMBO J 12:2567–2574
41. Zhao Z-Y, Wilson TJ, Maxwell K, Lilley DMJ (2000) The folding of the hairpin ribozyme: dependence on the loops and the junction. RNA 6:1833–1846
42. Wilson TJ, Lilley DMJ (2002) Metal ion binding and the folding of the hairpin ribozyme. RNA 8:587–600
43. Fedor MJ (1999) Tertiary structure stabilization promotes hairpin ribozyme ligation. Biochemistry 38:11040–11050
44. Rupert PB, Ferré-D'Amaré AR (2001) Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. Nature 410:780–786
45. Walter F, Murchie AIH, Lilley DMJ (1998) The folding of the four-way RNA junction of the hairpin ribozyme. Biochemistry 37:17629–17636
46. Tan E, Wilson TJ, Nahas MK, Clegg RM, Lilley DMJ, Ha T (2003) A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. Proc Natl Acad Sci USA 100:9308–9313
47. Klostermeier D, Millar DP (2001) Tertiary structure stability of the hairpin ribozyme in its natural and minimal forms: different energetic contributions from a ribose zipper motif. Biochemistry 40:11211–11218
48. Pljevaljcic G, Millar DP, Deniz AA (2004) Freely diffusing single hairpin ribozymes provide insights into the role of secondary structure and partially folded states in RNA folding. Biophys J 87:457–467
49. Nahas MK, Wilson TJ, Hohng S, Jarvie K, Lilley DMJ, Ha T (2004) Observation of internal cleavage and ligation reactions of a ribozyme. Nat Struct Mol Biol 11:1107–1113

Chapter 6
An Introduction to Macromolecular Crystallography Through Parable and Analogy

Alexander McPherson

6.1 An Analogy

Let us assume that we want to determine the structure of some object which is invisible. It has the supernatural property that it is non responsive to any electromagnetic radiation such as light. But, to make the example more concrete, let's assume it is an invisible, cream colored, 1984 Alpha Romeo Spider (curiously, exactly like the author's). If we have never seen such a glorious object before, how can we learn of its structure? How can we visualize it?

One way we might approach this problem is to take advantage of the fact that although the Alpha Romeo is impervious to light, it retains all of its other physical properties. We might, for example, take a basketball and throw it at the invisible object from some direction $\overline{k}_O$, and note which direction $\overline{k}$ it bounces off the object. More informatively, we might throw a hundred or a thousand balls at the invisible automobile and note how many balls bounce in all directions $\overline{k}$. Some will hit the fender, others the hood, others the windshield, etc. and depending on the orientation of the car with respect to the balls thrown along $\overline{k}_O$, some directions $\overline{k}$ for the reflected balls will be much favored over others. If the direction $\overline{k}_O$ corresponds to one aiming directly at the front of the car, for example, balls bouncing off the hood and windshield will be strongly favored.

Let's assume, however, that we can walk around the invisible Alpha Romeo and hurl the basketballs from many, in fact all possible directions $\overline{k}_O$, and each time, we note carefully how many balls bounce in which direction. Then, ultimately, we will know for every direction of our incoming stream of basketballs, $\overline{k}_O$, how many are reflected in every direction $\overline{k}$.

A. McPherson (✉)
Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA 92697-3900, USA
e-mail: amcphers@uci.edu

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_6,

This assembly of observations, $\overline{k}_O$, $\overline{k}$ and the number, or intensity of balls in the direction $\overline{k}$ contains information about the structure of the invisible object, the directions of its various external planes (doors, windows, hoods, fenders, etc.) from which the balls bounce. Now the question is, can we, from the observations, deduce the shape of the object which gave rise to the pattern of reflected basketballs? The answer is yes. Mathematical procedures do indeed exist for extracting the shape of a 1984 Alpha Romeo from a scattering pattern of basketballs. We might even invent some analogue device which we could place in a manner that it accumulated automatically the reflected balls and somehow translated the pattern into an image of the object. We would call such a device a lens.

Now basketballs are rather large objects (probes), and when they bounce from a surface plane, they are rather insensitive to its finer details such as windshield wipers, door handles, bolt heads, etc. We could, however, make our investigation more sensitive by using, instead of basketballs, tennis balls, and even more sensitive still by using ping-pong balls, or even marbles (no, let's not use marbles, they would damage the paint job). Then the direction in which our probe deflected would very closely reflect the undulations of the hood, and the presence of door handles. That is, we would obtain a more refined, higher resolution image. In other words, the resolution of detail we could obtain would be a function of the dimensions of the probe.

The approach illustrated here is not exactly what is done in X-ray diffraction, but it is similar. For example, we don't learn anything about the shape of the engine in the example of the automobile because our various probes cannot penetrate the interior of the car, while x-radiation can penetrate and reflect from the internal atoms of molecules. But in many other ways it is quite the same.

Let us alter our analogy a bit and now assert that the reason our 1984 Alpha Romeo is invisible is because it is too small to see. It is smaller than the wavelength of visible light. We could in principle, however, carry out the same experiment of walking around the minute car and directing a probe at it from all directions $\overline{k}_O$ and noting in every case what intensity of reflected probes was observed for all directions $\overline{k}$. If the probe we used sometimes penetrated into the interior parts of the object (and struck the transmission, for example) so much the better, for, although our pattern of diffracted probes would be considerably more complex, we would then learn about the structures of things inside the car as well. As long as the size of the probe was comparable to the sizes of the molecular features we wished to see then we could do as we did with the basketballs or ping-pong balls.

In organic molecules, the distances between bonded atoms are usually 1–2 Å, hence the size of our probe must be comparable. The wavelength, $\boldsymbol{\lambda}$, of x-rays used in diffraction experiments are usually between 1 and 2 Å. $\mathrm{CuK}_{\boldsymbol{\alpha}}$ radiation produced by most conventional laboratory sources, for example, is 1.54 Å wavelength. $\boldsymbol{\lambda}$ is exactly analogous to the probe size, the shorter $\boldsymbol{\lambda}$, the smaller the diameter of the ball we are using.

Now a single, molecule sized Alpha Romeo, impressive though it might otherwise be, would nonetheless be very, very small and, in practice, it would be

impossible to hit it with enough balls (or probes, or waves) from a particular direction $\overline{k}_O$ and measure the intensity of reflected waves, in a particular direction, $\overline{k}$. How could we amplify the effect so that we could measure it?

Consider an enormous parking lot full of identical 1984 Alpha Romeo Spiders, all perfectly parked by their drivers so that every car is identically oriented and placed in exact order. That is, they form a vast periodic array of cars. If we now direct millions of basketballs at this Alpha Romeo array from the same direction $\overline{k}_O$, then every car, having identically the same disposition, would reflect the balls exactly the same. The signal, or reflected pattern of probes, would be amplified by the number of cars in the parking lot, and the end result, which we call the signal, would be far more easily detected because of its strength. In our diffraction experiment, which is what we are really doing here, the automobiles are the molecules, the basketballs analogous to X-radiation, and the numbers of basketballs scattered in each direction are the intensities of the diffracted waves. Instead of an automobile parking lot, we have a molecular parking lot, a crystal.

A single voice, in a coliseum, though shouting, cannot be heard at a distance. Even a stadium full of voices cannot be heard far away if each individual is shouting a different cheer at random times. But if every voice in the stadium (or at least those favoring one particular team) are united in time in a single mighty cheer, the sound echoes far and wide. It is this cooperative effort of many individuals united in space and time, as occurs in a crystal that makes a molecular diffraction experiment possible.

Now clearly, in this analogy, we have simplified things a bit to get at the essentials, but the details can come later. It is important, to emphasize, however, that in carrying out our experiments on vast, ordered arrays of structurally unknown objects, we do sacrifice some information that might have been obtained from a single individual. In addition, because our probes in X-ray crystallography are not particles, or balls, but are waves, an additional complication is introduced. This is because waves add together, or interfere with one another, in a manner quite different than single particle probes. The good news is that we know how to deal with waves and we have the mathematical tools to overcome the complexities. Thus we ultimately must consider the diffraction pattern from our molecular array, or crystal, as sums of waves. Later, this sacrifice of information will emerge as what is known as "**The Phase Problem**" in X-ray crystallography.

6.2 A Lens

In our common experience, we rarely even think of waves and how they add together, even though we depend on light (wavelength $\boldsymbol{\lambda} = 3{,}500$–$6{,}000$ Å) for visualizing nearly everything. We use our eyes, microscopes, telescopes, cameras, and other optical devices that depend on waves of light, yet we never, it seems, have to deal directly with waves. The reason is that we have lenses that gather the light waves scattered by objects together, and focus them into an image of the original

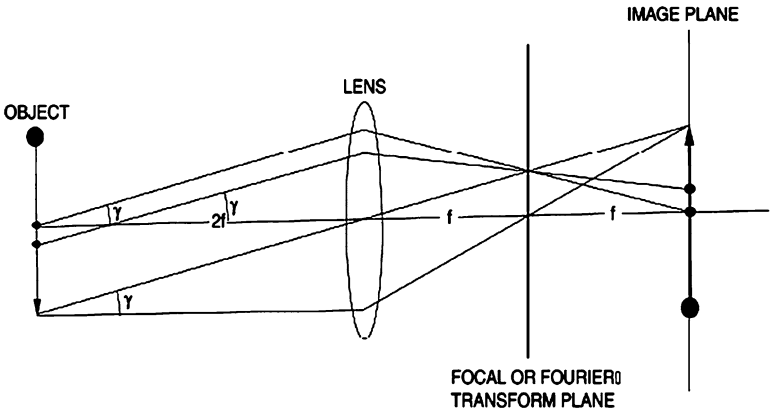


Fig. 6.1 Formation of the Fourier transform (or diffraction pattern) of an object by a lens having focal length f. The rays leaving the object are caused, by the refractive properties of the lens, to converge at both the image plane ($2f$) and at a second "focal" plane (f). The rays converging at each point on this transform plane at f are those that form a common angle with the plane of the scattering object, denoted here by $\boldsymbol{\gamma}$; that is, they have a common scattering direction

object. The lens of our eye focuses light waves scattered by an object at a distance into an image of that object on our retinas. The lens of a microscope focuses the light scattered by a minute object in the path of a light beam into an image of the object, and magnifies it for us at the same time.

Figure 6.1 illustrates the essential features of image formation by a lens using a simple ray diagram. There are two unique planes where the rays emitted by the light-scattering object intersect after passage through the lens. One plane is twice the focal length of the lens (**2f**). There, an inverted image of the object is formed by the summation of all rays scattered or diffracted from discrete points on the object converging at corresponding points on the image plane. The rays converge in a different manner, however, on a second plane at a distance **f** between the lens and the image plane. In that plane, rays intersect that do not originate at the same point on the scattering object, but rays converge that have the same direction (defined by the angle $\boldsymbol{\gamma}$) in leaving the object. The convergence of the various sets of rays, each having a different direction parameter $\boldsymbol{\gamma}$, forms in this plane a second kind of image, which is called the diffraction pattern of the object. The diffraction pattern is known mathematically as the Fourier transform of the object.

What does the diffraction pattern of an object look like? We can visualize the diffraction pattern, the Fourier transform, of an object by making a mask about the object and then passing a collimated beam of light through the mask and onto a lens. The lens, as in Fig. 6.1, then creates the diffraction pattern at a distance **f**, which we can view on a screen, or record on a film.

Figure 6.2 is a simple example. The object is the square shown on the mask in (a). If we look at a distance **f** behind the lens, then we see the diffraction pattern of the square in (b). A second example is shown in Fig. 6.3. With the possible exception

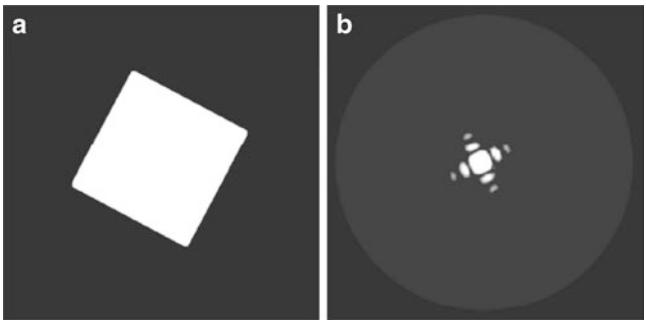


Fig. 6.2 If a mask containing the *square* in (**a**) has a parallel beam of light directed through it and onto a lens, as in Fig. 6.3, then if a screen were placed at one focal length (***f***) behind the lens, the diffraction pattern in (**b**) would be observed. At twice the focal length behind the lens, an image of the original *square* would appear. The pattern of *light* and *dark* seen in (**b**) is both the optical diffraction pattern of the *square*, and, in mathematical terms, it is the Fourier transform of the *square*

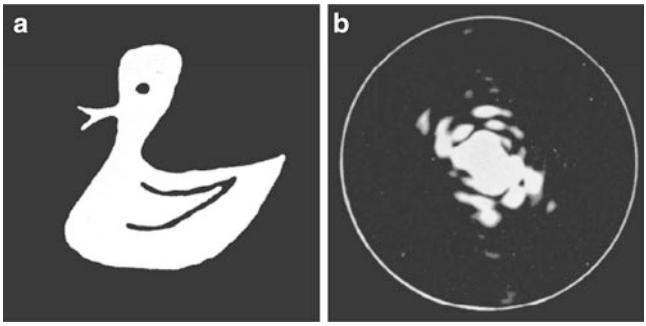


Fig. 6.3 If the object mask contains the image of a duck, as in (**a**), its optical diffraction pattern, or Fourier transform, seen at ***f***, is the seemingly meaningless pattern of *light* and *dark* in (**b**). It is important to note that the object, the duck, is a continuous object of basically arbitrary placed points (in a mathematical, not a biological sense), and as a consequence, the diffraction pattern is a continuous function of intensity consisting of patches and islands of *light*. Any continuous object, such as the duck seen here, might be expected to yield a diffraction pattern having these characteristics

of the diffraction pattern of DNA, this is probably the most reproduced diffraction pattern in the history of X-ray crystallography (see [1] for its origin). Here, the object in (a) is a duck and in (b) we see the duck's Fourier transform, its diffraction pattern. Significantly, if we were to place the diffraction pattern in (b) in the place of the object in (a), then at **f** behind the lens we would now see the duck. In other words, (b) is the Fourier transform of (a), but (a) is also the Fourier transform of (b). The transform is symmetrical, and it tells us that either side of the Fourier transform contains all the information necessary to recreate the other side. Thus, if we can record the diffraction pattern (or Fourier transform) of an object, we can always recreate the object itself or at least its image.

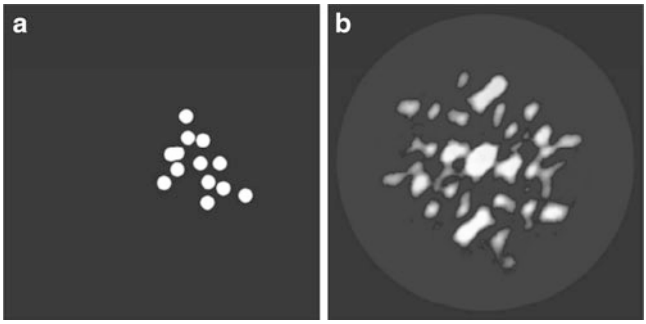


Fig. 6.4 In (**a**), the object for the optical transform is not a continuous object, but is a set of points arbitrarily distributed in space, as we might expect to find in a molecule made up of discrete atoms. That is, they bear no fixed mathematical relationship to one another. In (**b**) the optical diffraction pattern of the set of points is again a continuum consisting of islands of light and dark. Indeed, the transform is typical of one we might expect from any conventional organic molecule. The locations of the *light* and *dark* areas in the transform are dependent only on the **x, y** positions of the individual points in the object. If a point in the object were moved, the transform would change. If the entire set of points were rotated in the plane, the transform would undergo a corresponding rotation

Figures 6.2a and 6.3a were what we call continuous objects in that they were composed of a continuum of points covering a defined area, *i.e.,* a square or the surface of a duck. The diffraction patterns were similarly continuous. Molecules, however, are not really continuous; they are composed of atoms, which serve as discrete scattering points. In Fig. 6.4, for example, we have an arbitrary distribution of scattering points, like atoms in a molecule, and in (b) we see the diffraction pattern of the atom set. Note that even though the object is composed of unique scattering points, the diffraction pattern is nonetheless still continuous. Thus we should expect the diffraction pattern of a single molecule to be continuous, even if the molecule itself is not.

A final example, but of a different kind of object, is shown in Fig. 6.5. This object is a discrete set of points distributed over the surface of a mask in a periodic (uniformly repetitive) array. We call such a periodic point array in space a lattice. In (b) is the diffraction pattern of the lattice in (a), and *vice versa*. The diffraction pattern in (b) is also a lattice composed of discrete points (it is what we call a discrete transform), but the spacings between the points are quite different than for the lattice in (a). It can readily be shown that the distances between lattice points in (a) and (b) are reciprocals of one another. The Fourier transform of a lattice then is a reciprocal lattice.

It is possible to combine the two kinds of transforms illustrated here, the continuous transform of a molecule with the periodic, discrete transform of a lattice. In so doing, we will create the Fourier transform, the diffraction pattern of a crystal composed of individual molecules (sets of atoms) repeated in three dimensional space according to a precise and periodic point lattice.

We can get some idea as to what to expect by again using optical diffraction. In Fig. 6.6a is a pattern of scattering points having no internal symmetry or periodicity.

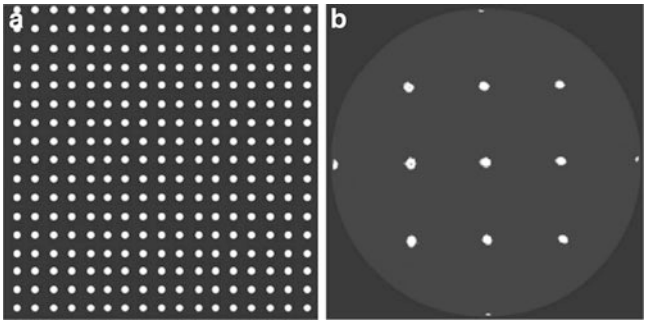


Fig. 6.5 In (**a**) the object of the mask, again exposed to the parallel beam of light, is not a continuous object or an arbitrary set of points in space, but is a two dimensional periodic array of points. That is, the **x, y** positions of the points are not arbitrary, but bear the same fixed, repetitive relationship to all others. One need only define a starting point along with two translation vectors along the *horizontal* and *vertical* directions to generate the entire array. We call such an array a lattice. The periodicity of the points in the lattice is its crucial property, and as a consequence of the periodicity, its transform, or diffraction pattern in (**b**) is also a periodic array of discrete points, i.e., a lattice. Notice, however, that the spacings between the spots in the diffraction pattern are different than in the object. We will see presently that there is a reciprocal relationship between distances in object space (which we also call real space), and in diffraction space (which we also call Fourier space, or sometimes, reciprocal space)

It might well represent the set of atoms in a molecule. Its diffraction pattern is seen in Fig. 6.6b. If the molecular motif in 6.6a is repeated in a periodic manner in two dimensions, that is, according to a point lattice, then we can generate the array seen in 6.6c. And what is the diffraction pattern, the Fourier transform of the periodic distribution in 6.6c? It is shown in 6.6d.

The diffraction pattern of the molecular array in Fig. 6.6d is also periodic, but the spacings between the diffraction intensities are reciprocals of the molecular point lattice. The intensities in 6.6d vary from point to point as well, unlike the example shown in Fig. 6.5. If the diffraction pattern in 6.6d were superimposed on that in 6.6b, we would find that the intensities at the discrete points in 6.6d are identical to the intensities of the corresponding points in the continuous transform in 6.6b, which they overlay. The discrete lattice according to which the molecules in Fig. 6.6a are periodically arrayed has the effect of allowing us to see, or sample the transform (diffraction pattern) of the individual molecules at periodic, specific points. The lattice appearing in diffraction space, having reciprocal spacings between points is again the reciprocal lattice.

To discriminate, or resolve, individual points in an object, as we saw in the Alpha Romeo Spider parable, one must utilize a radiation of wavelength comparable to the distances between the scattering points. Thus, we can use microscopy with light to resolve detail within an object that is on the order of a few thousand angstroms. We can use radio waves, as in radar, to resolve details measured in meters. If the objective is to produce an image of a macromolecule composed of atoms separated by an average bond length of about 1.5 Å, then one is obligated to use a

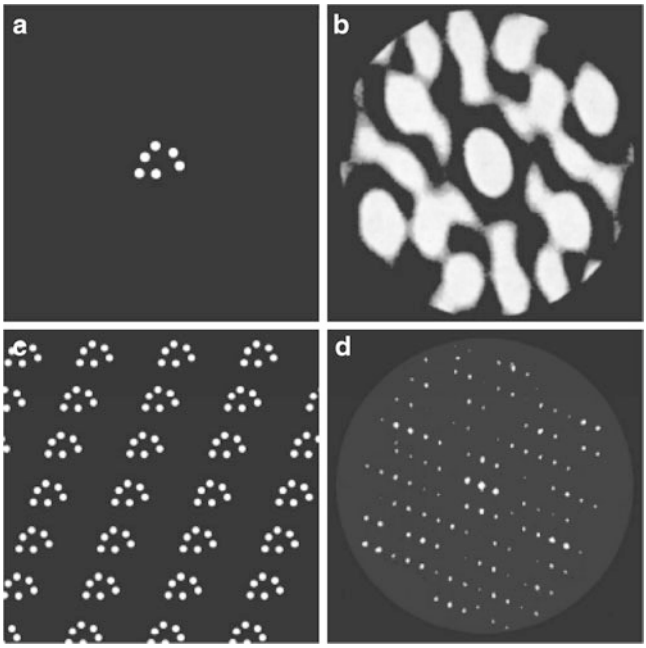


Fig. 6.6 In (**a**) is an arbitrary set of points that might represent the atoms in a molecule, and in (**b**) is the optical diffraction pattern of that set of points. It is a continuum of *light* and *dark* over the whole surface of the screen. The mask (object) in the optical diffraction experiment in (**c**) is the periodic arrangement of the fundamental set of points in (**a**) in two dimensions, i.e., the repetition of the object according to the instruction of a lattice. The diffraction pattern of (**c**) is shown in (**d**). We would find that if we superimposed the point array in (**d**) upon the continuous transform in (**b**), that the intensity at each point in (**d**) corresponded to the value of the continuous transform beneath. That is, the diffraction pattern in (**d**) samples the continuous transform in (**b**) at specific points determined by the periodic lattice of (**c**)

radiation of comparable wavelength. Conveniently, the characteristic X-radiation produced by the collision of high energy electrons with a number of different metal targets is of the range 1–3 Å, precisely what is required. Less convenient is the unfortunate reality that nature has not provided us with any known lens or mechanism for the focusing of scattered X-rays.

Unlike light, which, because of its refractive properties, can be focused by a properly ground glass lens, and unlike electrons, which, because of their charge, can be focused by electromagnetic fields, X-rays have no properties that permit an analogous process. Thus, X-radiation can be scattered from the electrons of an object, just as light or electrons are scattered by a specimen as they pass through it, but contrary to the situation we enjoy with a microscope, no lens can be interposed between specimen and observer to gather the scattered radiation and focus it into a meaningful image.

The crystal lattice, however, plays a second role. It not only amplifies the diffraction signal from individual molecules, it also serves as half a lens.

The X-rays scattered by the atoms in a crystal combine together, by virtue of the periodic distribution of their atomic sources, so that their final form is precisely the Fourier transform, that is, the diffraction pattern that we would ordinarily observe at **f** if we did in fact have an X-ray lens. Thus, the situation is not intractable, only difficult. We find from X-ray crystallography that while we cannot record the image plane, we can record what appears at the diffraction plane. It is then up to us to figure out what is on the image plane from what we see on the diffraction plane. Our computers serve as the other functional half of the lens, and Fourier mathematics provides the mechanism.

6.3 How X-Ray Diffraction Works

If a collimated beam of monochromatic X-rays is directed through an object, such as a macromolecule, the rays are scattered in all directions by the electrons of every atom in the object with a magnitude proportional to the size of its electron complement. This is the fundamental experiment that we perform in a diffraction investigation, and it is illustrated in Fig. 6.7. If the object were composed of more or less arbitrarily placed atoms, as they are in a single macromolecule, then at any point

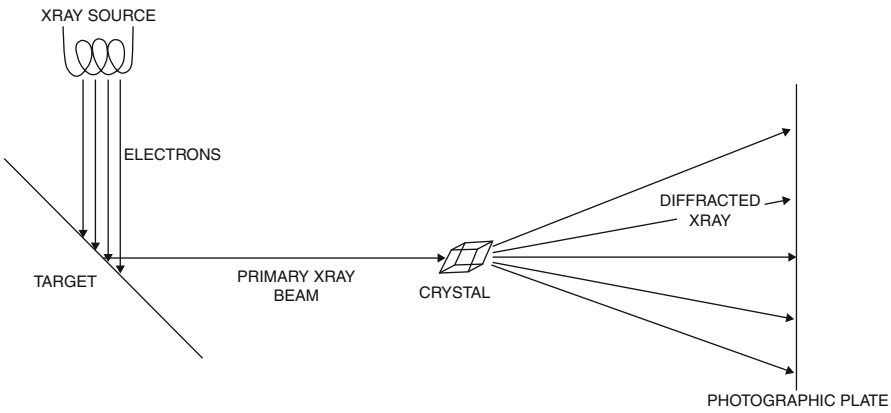


Fig. 6.7 The basic X-ray diffraction experiment is shown here schematically. X-rays, produced by the impact of high velocity electrons on a target of some pure metal, such as copper, are collimated so that a parallel beam is directed on a crystal. The electrons surrounding the nuclei of the atoms in the crystal scatter the X-rays, which subsequently combine (interfere) with one another to produce the diffraction pattern on the film or electronic detector face. Each atom in the crystal serves as a center for scattering of the waves, which then form the diffraction pattern. The magnitudes and phases of the waves contributed by each atom to the interference pattern (the diffraction pattern) is strictly a function of each atom's atomic number and its position (**x,y,z**) relative to all other atoms. Because atomic positions (**x,y,z**) determine the intensities of the diffraction pattern, or Fourier transform, then the diffraction pattern, conversely, must contain information specific to the relative atomic positions. The objective of an X-ray diffraction analysis is to extract that information and determine the relative atomic positions

in space about the isolated object a measurable amount of scattered radiation would be expected to be recorded by an observer. That is, the distribution of scattered rays would be a continuum of varying intensity. This was illustrated earlier. The variability of intensity throughout this continuous scattering distribution, which is again the Fourier transform, or diffraction pattern of the macromolecule, would depend on the relative positional coordinates and atomic numbers of the atoms in the object, and would ostensibly be independent of any other property of the object.

If a number of identical objects were arranged in three-dimensional space in such a way that they form a periodically repeating array, the scattering distribution, or diffraction pattern from the collection of objects would tend to be less continuous, taking on observable values at some points and approaching zero elsewhere. This was illustrated in Fig. 6.6. When the number of objects in the array becomes very large, as it does in a crystal, the scattering distribution, the diffraction pattern, becomes absolutely discrete.

Now as discussed already, the scattering of X-rays from a single protein or nucleic acid molecule would be immeasurably small. Due to its size alone, such an object could not be directly imaged. If, however, a vast number of such molecules are organized into an array, so that their scattering contributions are cooperative, then the resultant radiation can be observed and quantitated as a function of direction in space. This is precisely what is provided by macromolecular crystals, or in fact, any crystals.

Some examples of typical macromolecular crystals (see [2]) are shown in Fig. 6.8. While objects of beauty perhaps, their regular features only begin to suggest the degree of their internal order. In Fig. 6.9 we see with electron microscopy, and in Fig. 6.10, with atomic force microscopy evidence of their exquisite periodic nature. In Figs. 6.9 and 6.10 the individual molecules that comprise the crystals are aligned in rows and columns, indeed in all three dimensions, in perfect register, every molecule identically disposed, every molecule in precisely the same environment as any other. It is the molecular equivalent of our parking lot for Alpha Romeos, but in three dimensions.

The resultant radiation scattered, or diffracted, in specific directions create the intensities we see, precisely arranged, on an X-ray diffraction photograph. Because of the uniformity of orientation and periodic position imposed on the molecules by the crystal lattice, the scattered X-radiation, being waves, constructively interferes in unique directions dictated by the parameters that define the periodicity of the crystal lattice. It destructively interferes and sums to zero in all other directions. Hence we observe that the diffraction patterns from the ordered arrays that exist in crystals are absolutely discrete and that the observable diffraction pattern is an array of intensities that falls on a regular net or lattice (a reciprocal lattice). The spacings between the intensities, or reflections, and the symmetry properties that govern their distribution are manifestations of the periodic disposition of the molecules in the crystal. Because the physical relationship between the diffracted rays and the crystal lattice is well understood, mathematical expressions, such as Bragg's law, can be written that describe the correspondence.

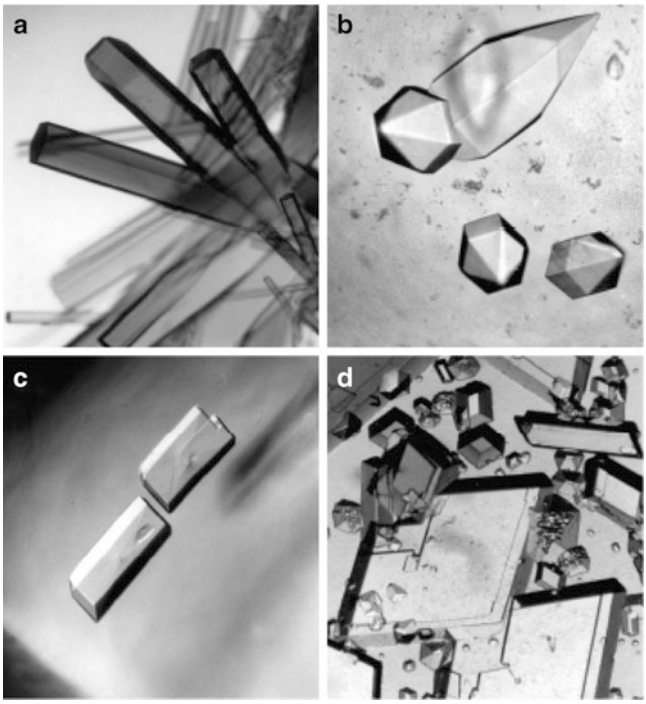


Fig. 6.8 Crystals of a variety of proteins. In (**a**), hexagonal prisms of beef liver catalase. In (**b**) crystals of α_1 — acid glycoprotein, (**c**) Fab fragments of a murine IgG antibody, and in (**d**) rhombohedral crystals of the plant seed protein canavalin

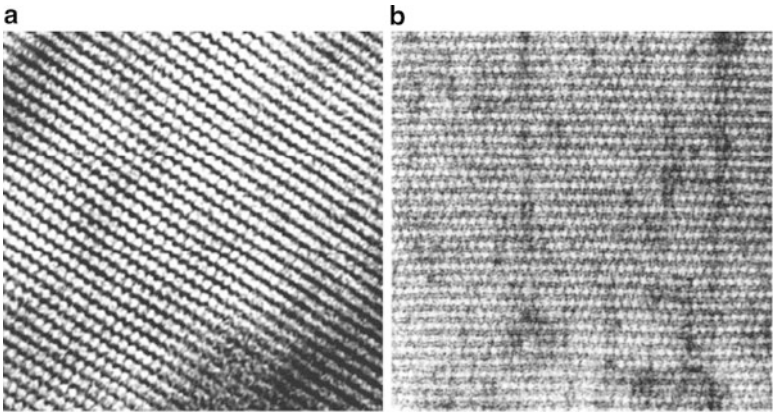


Fig. 6.9 Electron micrographs of negatively stained crystals of (**a**) pig pancreatic α amylase, and (**b**) beef liver catalase. The *dark* areas represent solvent filled areas in the crystal, which are replaced by the dense heavy metal stain, the *light* areas correspond to protein molecules where stain is excluded. The underlying periodicity of the crystals is evident here, even after dehydration and staining

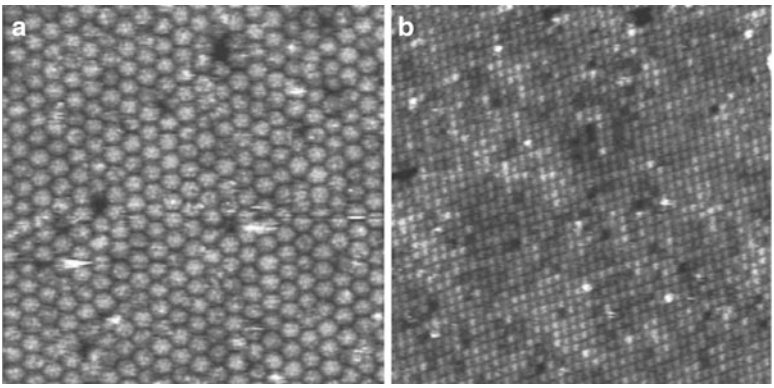


Fig. 6.10 Atomic force microscopy (AFM) also reveals the fundamental periodicity of macromolecular crystals. In (**a**) is the surface layer of a crystal of brome mosaic virus, a particle having a diameter of about 280 Å. In (**b**) is an AFM image of a monoclinic crystal of duodecahedral complexes of IgG antibodies which have a diameter of about 230 Å

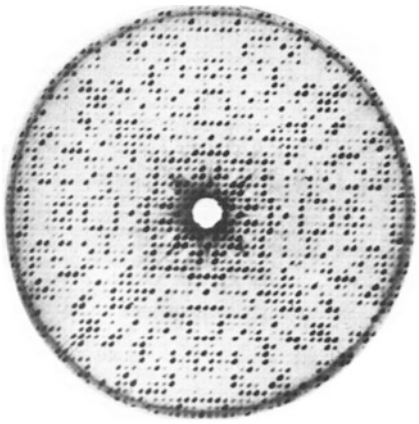


Fig. 6.11 The **hk0** zone diffraction image from a tetragonal crystal of lysozyme, an enzyme from hen egg *white*. Here, the fourfold symmetry of the pattern is striking, and it reflects the fourfold symmetry of the arrangement of the protein molecules in the unit cells of the crystal. Again, the intensities fall on a very regular, periodic net, or reciprocal lattice. The net is based on a tetragonal axis system

It is clear from looking at diffraction patterns obtained from real crystals, such as those in Figs. 6.11 and 6.12, that all of the reflections are not equal. They span a broad range of intensity values from very strong to completely absent. The variation in intensity from reflection to reflection is a direct function of the atomic structure of the macromolecules that comprise the crystal and occupy its lattice points. That is, the relative intensities of the reflections that make up the three dimensional diffraction pattern, or Fourier transform, of a crystal are directly related to the relative $\mathbf{x}_j$, $\mathbf{y}_j$, $\mathbf{z}_j$ coordinates of the nuclei of all of the atoms **j** that define an individual molecule, and to the relative strength, Z_j, with which the different atoms scatter X-rays. Z_j is the electron complement of each atom, and is, therefore, its atomic number.

The complete diffraction pattern from a protein crystal is not limited to a single planar array of intensities like those seen in Figs. 6.11 and 6.12. These images

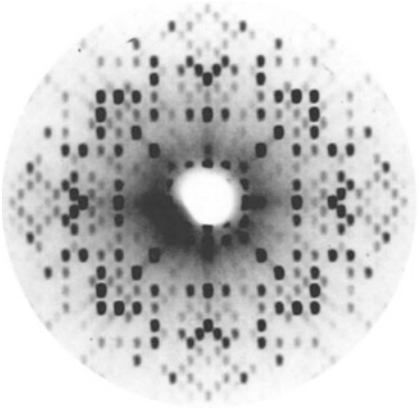


Fig. 6.12 Seen here is the **hk0** zone diffraction pattern from a crystal of M4 dogfish lactate dehydrogenase obtained using a precession camera. It is based on a tetragonal crystal system and, therefore, exhibits a fourfold axis of symmetry. The hole at *center* represents the point where the primary X-ray beam would strike the film (but is blocked by a *circular* beamstop). Note the very predictable positions of the diffraction intensities. All of the intensities, or reflections, fall at regular intervals on an orthogonal net, or lattice. This lattice in diffraction space is called the reciprocal lattice

represent, in each case only a small part of the complete diffraction pattern. Each photo corresponds to only a limited set of orientations of the crystal with respect to the X-ray beam. In order to record the entire three-dimensional X-ray diffraction pattern, a crystal must be aligned with respect to the X-ray beam in all orientations, and the resultant patterns recorded for each. From many two-dimensional arrays of reflections, corresponding to cross-sections through diffraction space, the entire three-dimensional diffraction pattern composed of ten to hundreds of thousands of reflections is compiled.

Because the diffraction pattern from a macromolecular crystal is the Fourier transform of the crystal, a precise mathematical expression can be set down that relates the diffracted waves to the distribution of atoms in the crystal. This expression, called the electron density equation, is a three-dimensional spatial transform that we refer to as a Fourier synthesis. It is sufficient here to simply understand that it is a summation of terms, one for each reflection observed in the diffraction pattern, and that the relative intensity of each reflection is the absolute magnitude of one of the terms in the series.

6.4 The Phase Problem

The situation, in truth, is somewhat more involved than the explanation thus far would lead you to believe. The individual reflections of the diffraction pattern are the interference sum of the waves scattered by all of the atoms in the crystal

in a particular direction. Being waves they have not only a resultant amplitude, but also a unique phase angle associated with each of them. This too depends on the distribution of the atoms, their $\mathbf{x}_j$, $\mathbf{y}_j$, $\mathbf{z}_j$. The phase angle is independent of the amplitude of the reflection, but most importantly, it is an essential part of the individual terms that make up the Fourier summation, the electron density equation. Unfortunately, the phase angle of a reflection cannot be recorded as we record the intensity. In fact, we have no empirical way to directly measure it at all. But, without the phase information, no Fourier summation can be computed. In about 1960, however, it became possible, with persistence, skill and patience (and luck), to recover this elusive phase information for protein crystals, thus permitting the calculation of Fourier summations and hence images of macromolecules. The technique, which is known as multiple isomorphous replacement, is based on the chemical derivitization of protein crystals with heavy metal atoms such as mercury. Its development [3–5] was the major breakthrough in modern crystallography which ultimately made possible the determination of macromolecular structures.

In this technique, the heavy atom, whose position in the crystal can be determined by what are known as Patterson techniques, provides a reference wave. In a derivatized crystal, the resultant diffraction intensities represent the sum of this heavy atom-produced reference wave interfering with the wave arising from all of the other atoms in the protein. Just as the relative phase of a specific sound wave can be deduced by 'beating' it against a reference sound wave of known phase, or for light waves using interferometry, the same is done for the native diffracted wave. The mathematical construct for obtaining the phase information requires measurement of the native diffraction intensities and the equivalent intensities from crystals derivatized at least at two unique sites in the crystal. It is known as a Harker diagram [6]. From Harker diagrams for each of the reflections that comprise a complete diffraction pattern, all of the necessary phase information, or at least reasonable approximations, can be obtained.

References

1. Taylor CA, Lipson H (1964) Optical transforms: their preparation and application to X-ray diffraction problems. Cornell University Press, Ithaca
2. McPherson A (1999) The crystallization of biological macromolecules. Cold Spring Harbor Press, Cold Spring Harbor
3. Blow DM, Crick FHC (1959) The treatment of errors in the isomorphous replacement method. Acta Cryst 12:794
4. Boyes-Watson J, Davidson E, Perutz MF (1947) An X-ray study of horse methemoglobin. Proc Roy Soc (Lond) Ser A 191:83
5. Bragg WL, Pertuz MF (1954) The structure of hemoglobin. VI. Fourier projections on the 010 plane. Proc Roy Soc A225:315
6. Harker D (1956) The determination of the phases of the structure factors of noncentro-symmetric crystals by the method of double isomorphous replacement. Acta Cryst 9:1

Chapter 7
Using NMR to Determine the Conformation of the HIV Reverse Transcription Initiation Complex

Elisabetta Viani Puglisi and Joseph D. Puglisi

Abstract Initiation of reverse transcription of genomic RNA is a key early step in replication of the human immunodeficiency virus upon infection of a host cell. Viral reverse transcriptase (RT) initiates from a specific RNA-RNA complex formed between a host transfer RNA ($tRNA_3^{Lys}$) and region at the 5′end of genomic RNA; the 3′ end of the tRNA acts as a primer for reverse transcription of genomic RNA. We determined the secondary structure of the 50 kDa complex between HIV genomic RNA and human $tRNA_3^{Lys}$ by nuclear magnetic resonance (NMR) spectroscopy. We show that both RNAs undergo large-scale conformational changes upon complex formation. Formation of an 18 base pair primer helix with the 3′ end of $tRNA_3^{Lys}$ drives large conformational rearrangements of the tRNA at the 5′ end, while maintaining the anticodon loop for potential loop-loop interactions. HIV RNA forms an intramolecular helix adjacent to the intermolecular primer helix. This helix, which must be broken by reverse transcription, likely acts as a kinetic block to reverse translation.

7.1 Introduction

Reverse transcription of the human immunodeficiency virus (HIV) RNA genome occurs immediately upon viral entry into a host cell (Fig. 7.1). This process is catalyzed by a virally-encoded RNA/DNA-dependent DNA polymerase, reverse transcriptase (RT) [38]. HIV RT consists of two subunits, p66 and p51, and its structure has been determined by x-ray crystallography; RT adopts a standard

E.V. Puglisi (✉) • J.D. Puglisi
Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305-5126, USA
e-mail: epuglisi@stanford.edu; puglisi@stanford.edu

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_7,

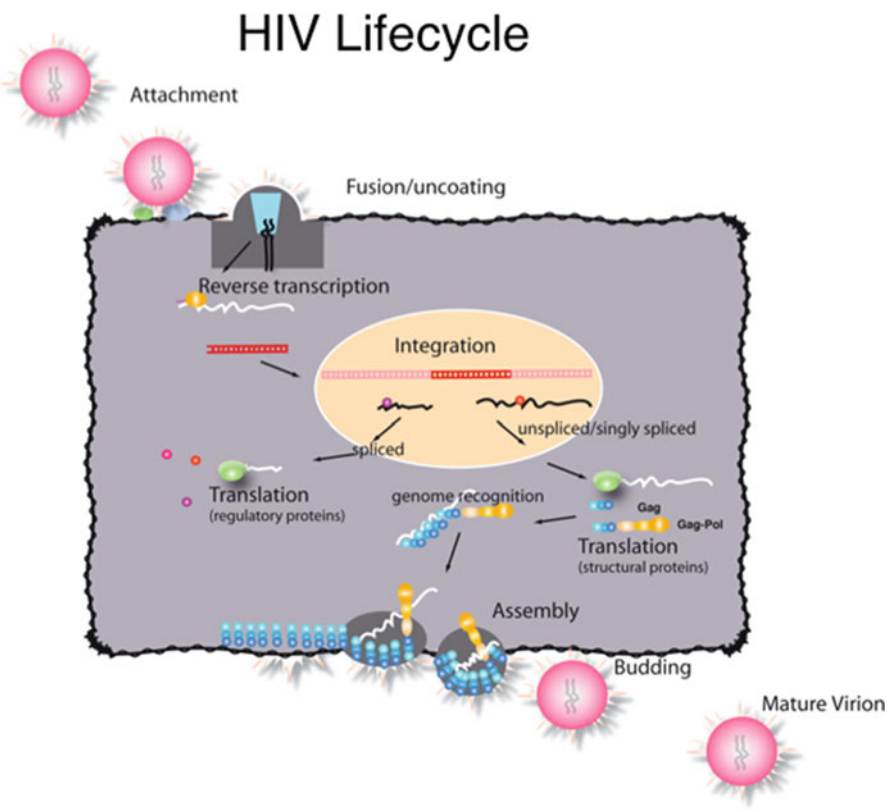


Fig. 7.1 Schematic of the HIV infection cycle. HIV binds to surface receptors, and enters the host cell. Reverse transcription of the viral RNA genome occurs upon uncoating in the cytoplasm. The viral DNA genome is transported to the nucleus, where it is integrated into the host genome. Early gene expression leads to synthesis of regulatory proteins, with subsequent synthesis of genomic RNA and structural proteins. Viral assembly and budding then occur along the cell membrane

polymerase fold, with an additional RNase H domain, which cleaves RNA-DNA hybrids. RT is a major target of therapeutic intervention in treatment of Acquired Immunodeficiency (AIDS).

RT initiates reverse transcription from a defined complex that is preassembled in the virion. The initiation complex consists of a host $tRNA_3^{Lys}$ bound to a defined region (primer binding site, PBS) near the 5′-end of the HIV genomic RNA [14, 15, 17] (shown schematically in Fig. 7.2a). RT copies the 5′-end of viral RNA into DNA, using its RNAase H activity to digest the RNA template within a product DNA-RNA duplex. Continuation of this process, strand jumping, and minus-strand DNA synthesis lead to a double-stranded DNA copy that is competent for integration into the host genome.

Here we focus on the specific RNA-RNA complex formed between host $tRNA_3^{Lys}$ and HIV genomic RNA [34] that directs RT initiation. The HIV genomic

PBS forms an 18 bp duplex with the 3′ end of $tRNA_3^{Lys}$, but sequences outside the simple PBS-tRNA base-pairing region are required for efficient initiation [19, 41–44]. Chemical probing of the viral RNA focused on a *ca.* 100 nt region whose reactivity to chemical probes changed upon binding of the viral RNA [14, 17]. Mutations within this region can change both the efficiency of tRNA-viral RNA complex formation and the rate of initiation of reverse transcription *in vitro* [15, 16] and *in vivo* [21]. Different viral isolates have sequence changes in this region [9–11].

Biochemical experiments have defined global features of the initiation complex. The complex between host $tRNA_3^{Lys}$ and HIV PBS is >40 kDa in size, and its secondary structure has been characterized by chemical and enzymatic probing [14, 15, 18]. Modified nucleotides in $tRNA_3^{Lys}$, in particular a modified 2-thioduridine at position 34 in the anticodon, stabilizes loop-loop interaction with an A-rich sequence in HIV genomic RNA [1, 3, 17]. A secondary structure for the tRNA-HIV RNA complex has been proposed based on these investigations, but detailed structural data have been lacking. Here, we discuss our structural studies by NMR to determine the conformation of the initiation complex.

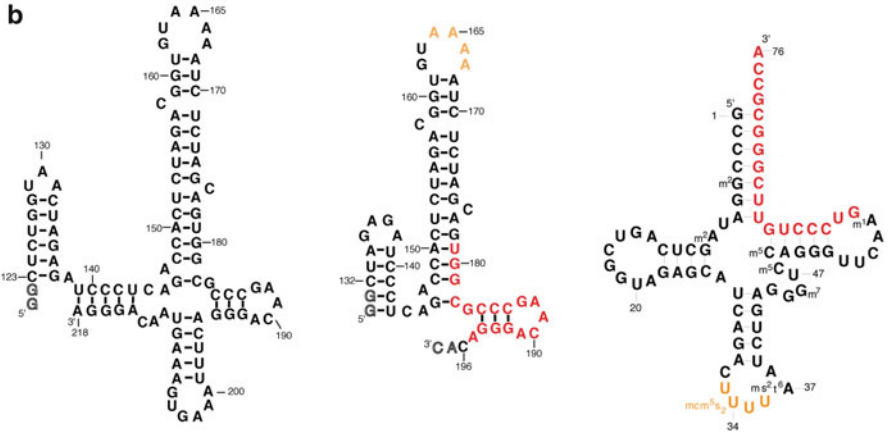
RNA structure guides key functions in HIV infection and replication. The viral genome is RNA, and RNA-RNA and RNA-protein interactions dominate the viral life cycle. Structural characterization of these RNA complexes has lagged behind those of individual protein components. RNAs are often difficult to prepare and conformationally heterogeneous, whether alone or in complex with other RNAs or proteins. Hindered by these problems, structural methods, such as NMR and x-ray crystallography, still struggle to solve RNA complexes.

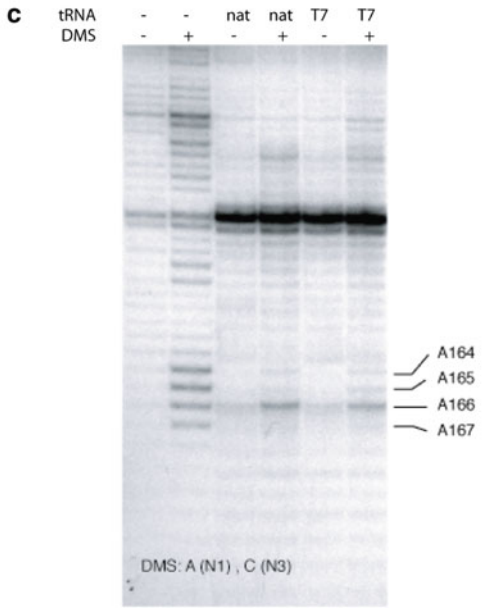
Here, we will review the NMR methods available to delineate RNA conformation [8]. We apply these approaches to probe the structure of the $tRNA_3^{Lys}$-HIV RNA complex. NMR has defined the conformation of a key hairpin loop in the HIV genomic RNA [35], and the conformation of $tRNA_3^{Lys}$ [36]. Biochemical analyses, coupled with prior experiments, defined the region of HIV RNA required for interaction with human $tRNA_3^{Lys}$, and allowed investigation of the RNA-RNA complex by NMR. We show that large-scale conformational changes occur in both the HIV genomic RNA and host $tRNA_3^{Lys}$ upon formation of a binary RNA complex. These results support prior models of the initiation complex and are a starting point for further mechanistic and structural studies of HIV reverse transcription.

7.2 Defining an RNA System to Explore the Initiation Complex

To explore a biological RNA by NMR, a tractable biochemical system must be created. The HIV genomic RNA is >9,000 nts, so the RNA must be minimized for structural study. We defined a system to investigate the initiation complex *in vitro*. Prior groundbreaking work by Marquet and co-workers defined the essential


a
tRNA Lys 3
RT
5′
3′
5′
3′
R
U5
PBS
U3
R


b


c
tRNA
DMS
nat
nat
T7
T7
A164
A165
A166
A167
DMS: A (N1) , C (N3)

features of the HIV RNA-RNA initiation complex using biochemical and chemical probing (discussed above). Their results defined a 300 nt fragment of the HIV-1 RNA genome, spanning the PBS region that was required for complex formation [2, 14, 15, 17, 18]. A *ca.* 100 nt sub-region (120–220 in the Mal isolate) formed a specific complex with modified native $tRNA_3^{Lys}$. We initially explored the conformation by NMR of a 99nt HIV fragment spanning the PBS (Fig. 7.2b), but its spectral properties (linewidths, number of peaks) were poor. The segment of HIV viral RNA between nts 130–200 directs $tRNA_3^{Lys}$ hybridization. A 69mer RNA (Fig. 7.2b) that corresponds to this region did not hydrolyze under standard conditions, and NMR spectroscopy suggested a single conformation for this RNA. We prepared unlabeled and ^{13}C, ^{15}N-labeled versions of the 69mer, and confirmed a general secondary structure as proposed from chemical probing experiments.

To observe whether the HIV RNA constructs could form a specific initiation complex with $tRNA_3^{Lys}$, we performed chemical probing experiments. Natural, fully-modified $tRNA_3^{Lys}$ was purified from bovine liver as described previously [45], and was used as a control in comparison to transcribed $tRNA_3^{Lys}$. RNAs were hybridized at 1:1 stoichiometry at 90°C for 3 min in low salt (10 mM Na phosphate or cacodylate, pH 7), followed by cooling to 60°C, addition of 100 mM NaCl, and slow cooling to room temperature. $MgCl_2$ was only added at room temperature, to avoid hydrolysis of the RNAs. Hybridization of tRNA and HIV RNA was readily confirmed by native gel electrophoresis.

RNA structure in the free and complexed HIV 69mer was probed using chemical reactivity with dimethyl sulfate, which reacts with the N1 position of A and N3 position of C [33]. In the absence of tRNA, the 69nt RNA shows chemical reactivity within the A-rich loop region between A164-167 (Fig. 7.2c). Upon addition of 1:1 stoichiometry of natural $tRNA_3^{Lys}$ or transcribed $tRNA_3^{Lys}$, similar protections are observed in the A-rich loop region, indicating formation of RNA structure in this

Fig. 7.2 (**a**) Schematic of the reverse transcription initiation complex within the HIV genome. The RNA genome of HIV RNA has the standard structure for retroviral genomes, with long terminal repeat regions (U5, U3, R), and the primer binding site (PBS) near the 5′ end. Host $tRNA_3^{Lys}$ interacts with the PBS through an 18 base-pair interaction, providing the 3′-OH group to initiate reverse transcription by viral reverse transcriptase (RT). (**b**) RNA oligonucleotides corresponding to HIV-1 genomic RNAs used in the current study (*left*) a 99 nt and (*right*) 69 nt RNA with numbering according to the Mal isolate. Additional nucleotides added at the 5′ end for transcription with T7 RNA polymerase are indicated as *outlines*. The primer binding site region of 18 bp of complementarity with tRNA is highlighted in *red*, whereas the A-rich loop is in *yellow*. (**c**) Secondary structure of human $tRNA_3^{Lys}$ with modified nucleotides indicated. Coloring as in (**b**). (**d**) Chemical probing experiments on HIV initiation complexes with 69nt model HIV1 genomic RNA. Accessibility of adenosine N1 and cytosine N3 positions to reaction with dimethyl sulfate (DMS) was detected in the presence or absence of 1:1 stoichiometry of either pure bovine native $tRNA_3^{Lys}$ (nat) or T7 RNA polymerase transcript (T7). Reactivity with DMS was detected using primer extension with reverse transcription using a DNA primer in the absence (−) and presence (+) of dimethyl sulfate reaction. Reactions were performed at room temperature as described (color figure online)

region. These results confirmed that tRNA transcripts or native tRNAs form similar complexes with the 69nt HIV genomic RNA construct. We thus used the transcribed tRNA, which is readily labeled with ^{15}N and ^{13}C, for subsequent NMR experiments on the RNA-RNA complex.

7.3 RNA NMR Spectroscopy

NMR spectroscopy allows investigation of RNA secondary and tertiary structures [6, 8, 25, 40]. Base-pair formation can be monitored through the exchangeable imino proton resonances; these protons are protected from solvent exchange upon base-pair formation, and appear in a distinct ^{1}H frequency in an NMR spectrum. Isotopic labeling of RNAs with ^{13}C and ^{15}N facilitate assignment and analysis of RNA NMR spectra [23, 25, 40]. High-resolution structures of RNAs up to 100–120 nts have been painstakingly achieved by using a combination of distance constraints and residual dipolar couplings obtained through heteronuclear NMR [4, 5, 24, 32]. Large RNA complexes suffer from line-broadening caused by slow molecular tumbling, although TROSY methods greatly improve linewidths for RNA base ^{1}H resonances [25]. Secondary structure and overall folds for large RNAs (>25 kDa) can be achieved by analysis of experiments on imino ^{1}H resonances [12, 36].

Detailed structural determination of RNAs by NMR requires spectral assignments and subsequent analysis of conformational constraints. Assignment of resonances in RNA molecules is hindered by high spectral overlap; RNAs consist of only 4 chemically similar nucleotides, whereas proteins have 20 amino acids of distinct chemical nature. RNAs have about 1/3rd the proton/unit mass density as proteins, leading to lower interproton distance restraint densities for RNAs vis a vis proteins. RNAs also have 6 dihedral angles that define backbone conformation, compared to 2 for proteins. Thus each nucleotide has much greater conformational freedom, and fewer restraints, than for an equivalent protein. Despite these problems, NMR is still a powerful method to probe RNA conformation.

Assignment of NMR resonances in RNA begins with the exchangeable imino proton resonances (Fig. 7.3). Both uracil and guanine have imino protons. When engaged in canonical Watson-Crick A-U or G-C pairing, exchange of these imino protons with surrounding solvent H2O is suppressed and resonances are observed downfield (10–15 ppm) in an uncluttered region of a ^{1}H NMR spectrum. Each Watson-Crick base pair in an RNA structure will give rise to a single imino resonance in the NMR spectrum; base pairs at ends of helices are often broadened due to solvent exchange. Because RNAs are highly hydrated, solvent exchange rates are still considerable for base-paired protons (lifetimes 50–1,000 ms) [20]. Thus sophisticated solvent suppression methods are required to allow observation of imino ^{1}H resonances without saturation transfer [39].

Heteronuclear NMR experiments on ^{15}N-labeled of ^{13}C, ^{15}N-labeled RNAs facilitate assignment of exchangeable RNA resonances. Using^{15}N-heteronuclear

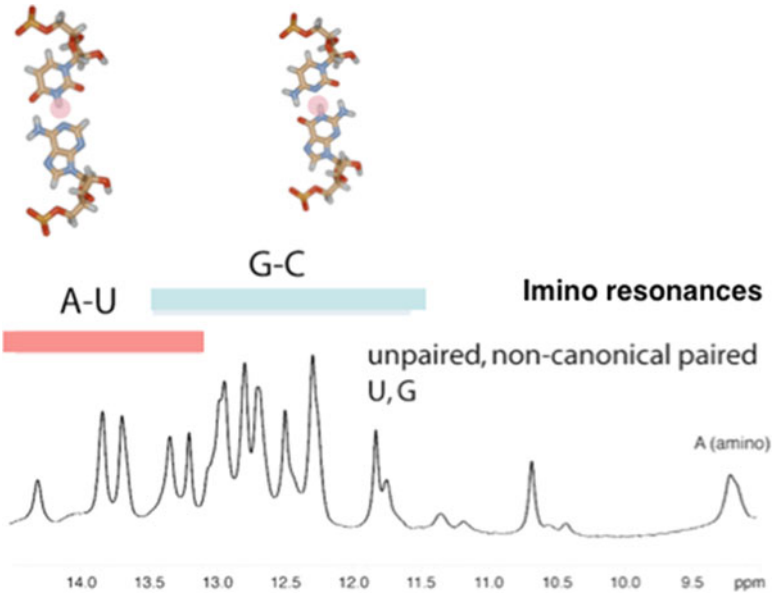


Fig. 7.3 Imino ^{1}H spectrum of a folded RNA. The canonical A-U and G-C pairs are shown, with the imino ^{1}H highlighted. The A-U resonances appear downfield from G-C resonances, but the spectral separation is not complete. Imino resonances from non-paired and non-canonical base pairs are often observed further upfield, as indicated. The spectrum is that of $tRNA_3^{Lys}$ in 10 mM $MgCl_2$ 100 mM NaCl, 10 mM Na phosphate, pH 6.5 at 25°C

single quantum coherence (HSQC) spectroscopy, the imino ^{1}H resonances from G and U are resolved by different ^{15}N chemical shift, despite overlap in their corresponding ^{1}H chemical shift (Fig. 7.4). Amino resonances are often broadened due to both solvent and chemical exchange by rotation around the C-N bond (Fig. 7.4) [26–31]; however C-amino resonances are usually observed as distinct upfield and downfield resonances upon base-pair formation. Adenosine and guanine amino resonances are usually not observed in Watson-Crick base pairing, but their observation can be diagnostic of non-canonical interactions in RNA. The use of HSQC experiments resolves the distinct ^{15}N chemical shifts of the amino resonances.

To assign imino ^{1}H resonances, standard NOESY experiments are performed. Sophisticated solvent suppression techniques are required to avoid suppression of imino resonances due to magnetization saturation transfer from solvent. We usually use shaped pulses, such as shifted laminar pulses [39], to achieve selective excitation of exchangeable and non-exchangeable resonances with a null at the water resonance. NOEs link spins through dipolar relaxation mechanisms that depend on both relative motions of the two spins, and $1/r^6$, where r is the distance separating the two spins. The strength of NOE interaction thus depends very strongly on distances, and NOEs are not observed beyond internuclear distances of 6 Å.

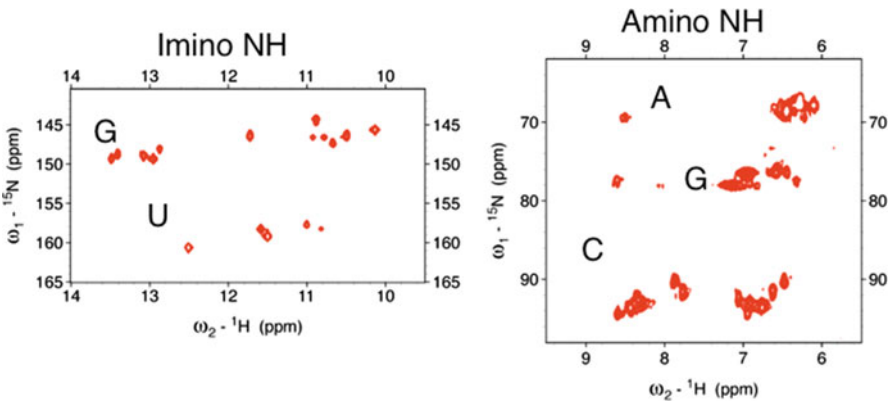


Fig. 7.4 Heteronuclear ^{15}N NMR facilitates analysis of RNAs that are uniformly labeled with stable isotopes. (*left*) ^{1}H- ^{15}N HSQC of the imino region of a folded RNA, showing the distinct ^{15}N chemical shifts for U and G imino nitrogens. (*right*) ^{1}H-^{15}N HSQC for amino resonances in a folded RNA

NOE interactions map secondary structures in RNA. Within an A-form RNA helix, imino protons from adjacent base pairs are within 4 Å, providing a convenient pathway for assignment of secondary structure. NOEs are observed that connect adjacent base-pairing imino protons in the region between 10 and 15 ppm. A helix is identified by a series of sequential NOEs. These NOE connectivities are facilitated by indirect magnetization transfer (spin diffusion) through the amino protons that are proximal to the imino protons of G-C and A-U base pairs. We therefore use relatively long NOE mixing times to ensure efficient NOE transfer amongst imino protons. For a 25 kDa RNA, we use mixing times of >100 ms to observe inter base-pair NOEs.

7.4 tRNA as an Example

Using the combination of heteronuclear NMR and NOESY spectra, we assigned the imino proton NMR spectrum of human $tRNA_3^{Lys}$. Heteronuclear NMR on a ^{15}N-labeled $tRNA_3^{Lys}$ sample demonstrated formation of a folded $tRNA_3^{Lys}$ structure in the presence of 10 mM Mg^{2+} (not shown), with the number of peaks consistent with formation of the standard tRNA fold. The imino proton NMR spectrum of folded $tRNA_3^{Lys}$ in 10 mM Mg^{2+} was assigned using standard homonuclear NOESY experiments, and confirmed by heteronuclear NMR presented above. Figure 7.5 shows a NOESY experiment obtained on $tRNA_3^{Lys}$ at long mixing time to allow observation of the inter base-pair NOES. The strongest NOE is observed between the two imino protons of the G6-U67 base pair, which are close in space (2.4 Å); this NOE is observed in short mixing time NOESYs (50 ms), and provides a reference point for assignment of the tRNA.

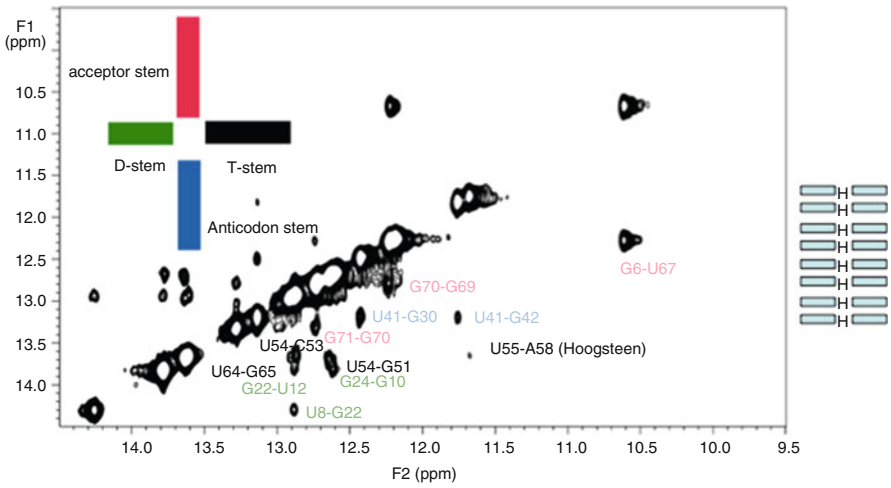


Fig. 7.5 NOESY spectrum of human $tRNA_3^{Lys}$ showing imino-imino ^1H NOES. The assignments of the different tRNA helical stems are shown. Data were acquired with a mixing time of 250 ms at 25°C.

Resonances for the acceptor- , T- and anticodon-stems of tRNA were assigned using the NOE connectivity walk. These helical segments are standard A-form geometry. Reference points included the G6-U67 pair in the acceptor stem, as well as the reversed Hoogsteen U54-A58 pair in the T-loop. The Hoogsteen Pair was identified by a strong U54(H3)-A58(H8) NOE, which is characteristic of this pairing geometry. Interestingly, several imino proton resonances from terminal base pairs in these stems were not observed due to solvent exchange in the NOESY experiment. This exchange behavior is typical for terminal base pairs.

The D-stem and loop represent the key area of tertiary interactions in folded tRNA [37]. To assign the resonances in the D-stem, we took advantage of novel chemical shifts and NOEs observed for exchangeable proton resonances in this region. The U8-A14 base pair has an unusual downfield chemical shift (14.28 ppm) and is the starting point to assign the imino resonances in this stem. The remaining imino resonances in the D-stem were readily assigned through sequential NOEs (Fig. 7.5), and include the C13-G22, U12-A23, C11-G24 and G10-C25 pairs.

7.5 Larger RNAs

The advent of transverse relaxation optimized spectroscopy (TROSY) has been critical in pushing NMR approaches to larger molecular weight systems. Standard NMR experiments suffer from decreasing T2 values as the rotational correlation time increases with molecular weight. T2The TROSY experiment makes use of the longstanding observation that distinct components of a J-coupled multiplet

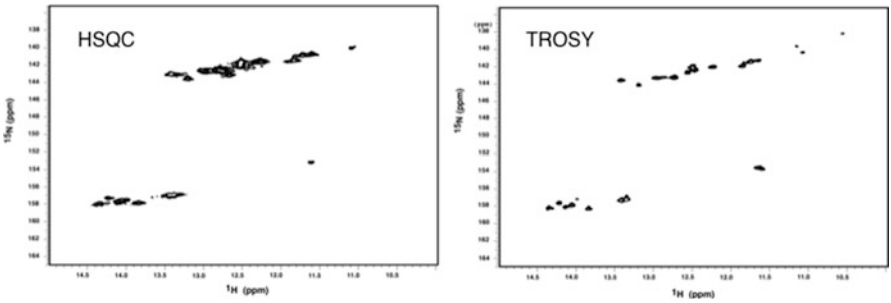


Fig. 7.6 Comparison of a standard imino ^{1}H-^{15}N HSQC (*left*) and ^{1}H-^{15}N TROSY (*right*) for the 50 kDa HIV initiation complex. Data were acquired on a 1:1 complex of ^{15}N-labeled tRNA: unlabeled HIV RNA at 18.8 T

experience different contributions from dipolar and chemical shift anisotropy (CSA) relaxation mechanisms. Wuthrich and co-workers realized that by selecting the cross peak component for which the dipolar and CSA contributions to T2 cancel, a narrow resonance is selected with increased apparent T2. Since CSA contribution to transverse relaxation rates increases as the B_0^2, whereas the dipolar contribution is independent of field strength, the two contributions can be matched by varying B_0. For coupled ^{15}N and ^{1}N, the minimum linewidths are achieved at between 900 MHz and 1000MHZ ^{1}H frequency. This is a considerable driving force for the development of stronger magnets.

TROSY allows investigation of large >25 kDa RNAs, and has been essential for the investigation of the large RNA complexes involved in HIV reverse transcription initiation. A comparison of a standard ^{1}H-^{15}N HSQC with a ^{1}H-^{15}N TROSY experiment shows the considerable improvement in linewidth achieved by the TROSY at 18.8 T (800 MHz ^{1}H frequency) (Fig. 7.6). Similar improvements can be achieved by using ^{1}H-^{13}C TROSY for the aromatic resonances in RNA (not shown). The aliphatic ribose ^{1}H resonances have smaller CSA contributions, and TROSY is less effective at narrowing their linewidths. TROSY elements integrated into more complex pulse sequences can allow for effective multidimensional NMR that avoids magnetization or coherence loss during delay times.

Detection of scalar couplings through hydrogen bonds allows direct definition of base pairing schemes in RNAs. The partial covalent nature of a N-H—N hydrogen bond allows scalar coupling between hydrogen bonded ^{15}N nuclei. This is ideal in nucleic acid base pairs. Using a TROSY detection scheme, scalar couplings are directly detected in the HNN-COSY experiment. Using this experimental approach, A-U and G-C base pairs are defined. Non-canonical pairing interactions are also observed. For example, in the HCV IRES domain IIId domain, a reversed Hoogsteen U-A pair is formed, involving the N3 position of U and the N7 position of A. The unique ^{15}N chemical shifts of N7 versus N1 resonances unambiguously assign this interaction (Fig. 7.7).

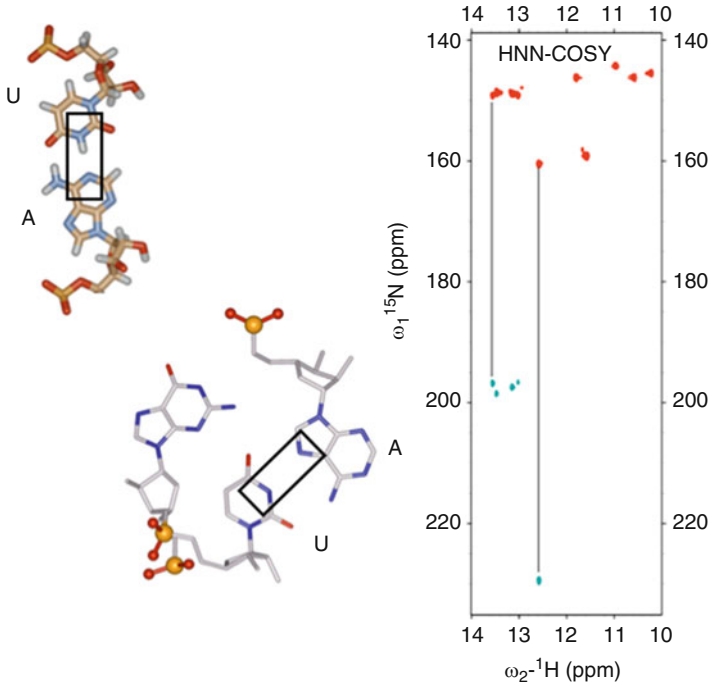


Fig. 7.7 HNN COSY experiment detecting through-bond scalar couplings in RNAs. In this experiment, correlations were observed from U N3 nitrogens to either A N1 nitrogens (standard A-U pair) or A N7 nitrogen (Hoogsteen pairing) (Data were acquired on HCV IRES domain IIId at 18.8 T. G-C correlations are not shown)

In this review, we will not discuss in detail the full structure determinations of RNAs by NMR. Structure determination requires assignment of exchangeable and non-exchangeable resonances. Various through-bond methods allow correlations of base exchangeable and non-exchangeable resonances, and base to ribose resonances. Ribose ^{1}H chemical shift overlap is overcome by ^{13}C labeling, which resolves the H1′, H2′/H3′, H4′ and H5′/5″ resonances (Fig. 7.8). Through bond correlation methods, including HCCH-COSY and –TOCSY methods allow ribose assignments. Resonances on adjacent nucleotides are linked on smaller (<60nt) RNAs by through-bond transfer through the ^{31}P nucleus. Once assignments are obtained, structural restraints are collected through standard NOESY and scalar coupling experiments. In addition, residual dipolar couplings can be measured in partially aligned media that provide addition long-range conformational restraints for RNAs. Three-dimensional structures of RNAs are calculated using restrained molecular dynamics protocols that include the experimentally derived NMR restraints. High-quality structures of RNAs up to 30 kDa have been obtained [6, 24].

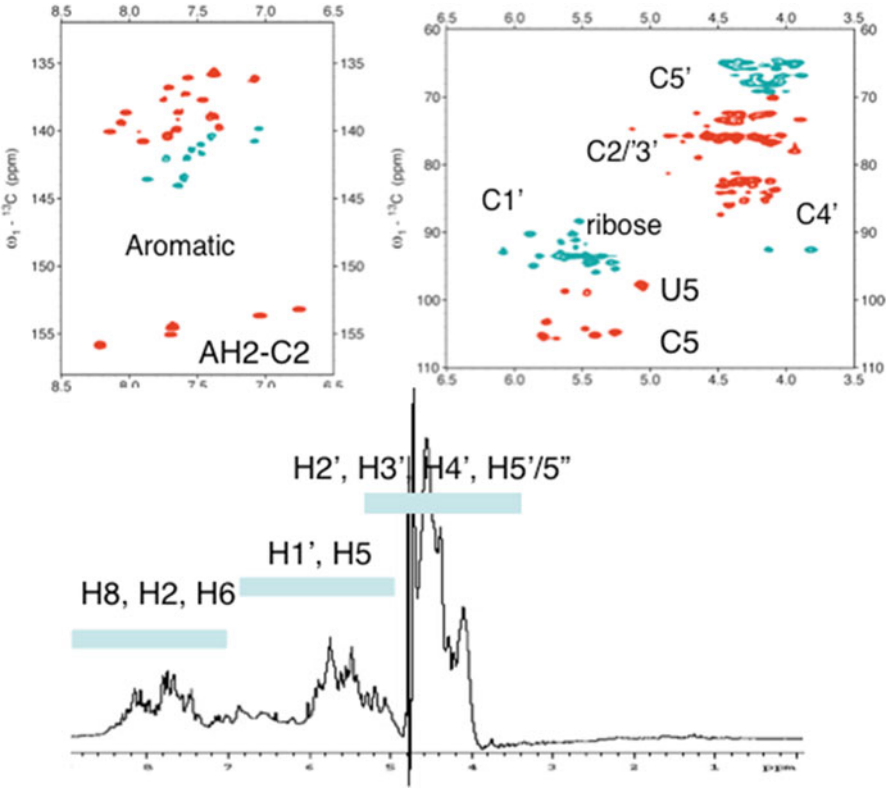


Fig. 7.8 Spectral overlap in non-exchangeable RNA resonances is relieved by ^{13}C labeling. (*bottom*) One-dimensional spectrum of an RNA showing the high spectral overlap of non-exchangeable resonances. (*top*) ^{1}H- ^{13}C correlations in RNA decrease spectral overlap. ^{1}H-^{13}C HSQC of the aromatic and ribose portions of the spectrum show the resolved ^{13}C chemical shifts for distinct classes of resonances (Data were acquired on a 15 kDa RNA in D_2O at 18.8 T)

The power of detailed structure determination of RNAs is demonstrated by our structure determination of the A-rich loop from the HIV genomic RNA [35]. This small, 14 nt RNA, was studied using homonuclear and heteronuclear NMR. Assignments were obtained for the exchangeable and non-exchangeable resonances, and the conformations calculated by restrained molecular dynamics. This structure was determined before the use of residual dipolar couplings in solution NMR. The conformation of the A-rich loop reveals a 3′-stacked conformation for the adenosines, with the hairpin stabilized by a U-turn interaction (Fig. 7.9). The conformation of this loop resembles that of a tRNA anticodon loop. By pre-organizing the RNA structure, the stacked conformation facilitates base pairing with complementary sequences; loop-loop kissing interactions are common in RNA. The stacked structure of the HIV A-rich loop may facilitate the proposed interaction with the human $tRNA_3^{Lys}$ anticodon loop.

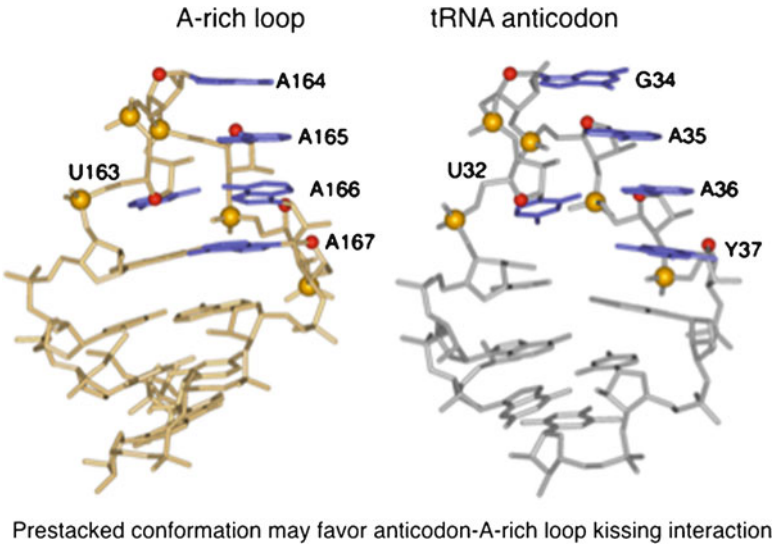


Fig. 7.9 Structure of the HIV A-rich loop RNA determined by NMR and comparison with the conformation of the $tRNA^{Phe}$ anticodon stem loop. The RNA adopts a U-turn conformation with stacked nucleotides, similar to the anticodon

7.6 HIV Initiation Complex Probed by NMR

We have used NMR of exchangeable imino ^{1}H resonances to determine the secondary structure of the HIV initiation complex. Despite the large molecule weight of the 1:1 $tRNA_3^{Lys}$-HIV_{69} RNA complex (MW 47,000), the imino ^{1}H NMR spectrum was well resolved and interpretable. Large changes in the imino spectrum of tRNA (and HIV 69mer) are observed upon 1:1 complex formation (Fig. 7.10), indicating structural rearrangements in the tRNA. Resonances from the tRNA acceptor stem (e.g. G6), D stem, and tRNA tertiary structure (amino resonances, U8, U55) [36] are lost or shifted in the RNA-RNA complex. New resonances from imino protons within the tRNA are observed. These chemical shift changes suggest large changes in the tRNA structure upon complex formation.

Stable-isotope labeling of one component in an RNA complex allows selective NMR observation of that component. We have used this approach for the initiation complex. We prepared 100% ^{13}C, ^{15}N-labeled $tRNA_3^{Lys}$ and formed a 1:1 complex with unlabeled HIV genomic RNA (69mer). Within the 1:1 RNA-RNA complex, drastic changes in the spectrum for $tRNA_3^{Lys}$ are observed (Fig. 7.11). The D-stem, with its set of tertiary interactions involving G10-C25, C11-G24, U12-A23, and U8, undergoes large chemical shift changes in the complex. Likewise, shifts are observed in the T- and acceptor stem resonances G6-U67, C2-G71, U55 in the T loop (Fig. 7.11), which support the proposed hybridization of 18 nts at the 3′

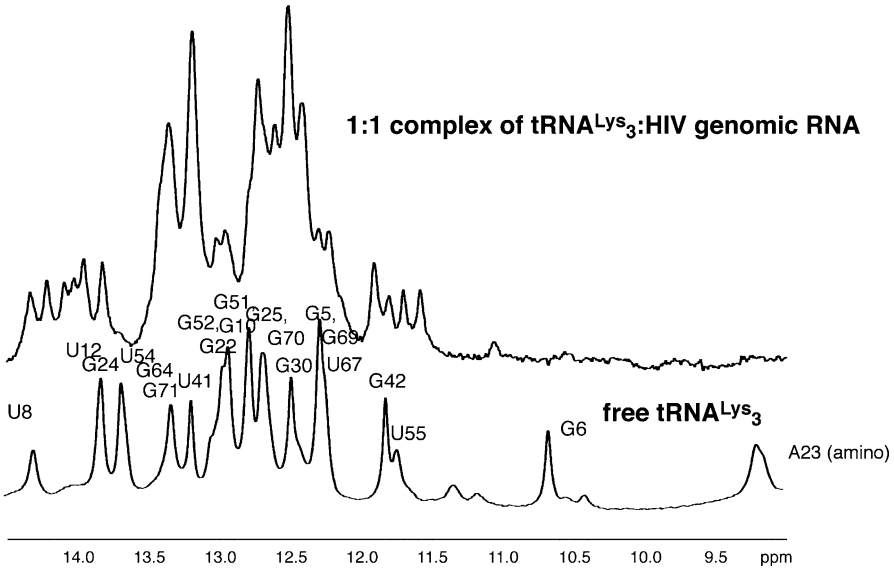


Fig. 7.10 Imino ^{1}H NMR spectra of folded human $tRNA_3^{Lys}$ transcript (*bottom*) compared to that of 1:1 complex of human $tRNA_3^{Lys}$ with 69 nt HIV1 genomic RNA (*top*). Spectra were acquired at 800 MHz in 10 mM $MgCl_2$, 100 mM NaCl, 10 mM Na phosphate, pH 6.5 at 25°C. Assignments of resonances in folded $tRNA_3^{Lys}$ were reported previously

end of $tRNA_3^{Lys}$ (located in the T- and acceptor stems) with the complementary PBS sequence in HIV RNA. These chemical shift mapping experiments suggest changes in the D, T and acceptor stems of the tRNA upon binding to HIV genomic RNA.

The following general strategy was used to determine the secondary structure of the tRNA-HIV complex using NMR. First, imino proton resonances were assigned to tRNA or HIV RNA by preparing two individual samples with one component uniformly labeled with ^{13}C, ^{15}N nucleotides, and the partner unlabeled (Fig. 7.12). Unlabeled complex samples were used for homonuclear NOESY experiments to measure imino-imino NOEs. Combined with the HSQC data obtained with labeled samples, these data allowed assignment of helical elements. Finally, intramolecular (tRNA-tRNA or HIV-HIV) helical elements were assigned using HNN COSY approaches [7]. We explore each step in detail below.

We first assigned the imino ^{1}H resonances for base pairs in the 18 bp PBS helix, whose formation was suggested by the chemical shift changes in tRNA upon complex formation. We performed homonuclear NOESYs at several mixing times (50, 100, 200 ms) to observe the characteristic imino-imino proton NOEs in A-form RNA helices (discussed above). These NOESY experiments confirm that 18 bp helix is formed in the tRNA-HIV complex (Fig. 7.13). NOEs were observed connecting 16 continuous base pairs from C_H196-G_T59 until G_H181-C_T74. Our assignment of

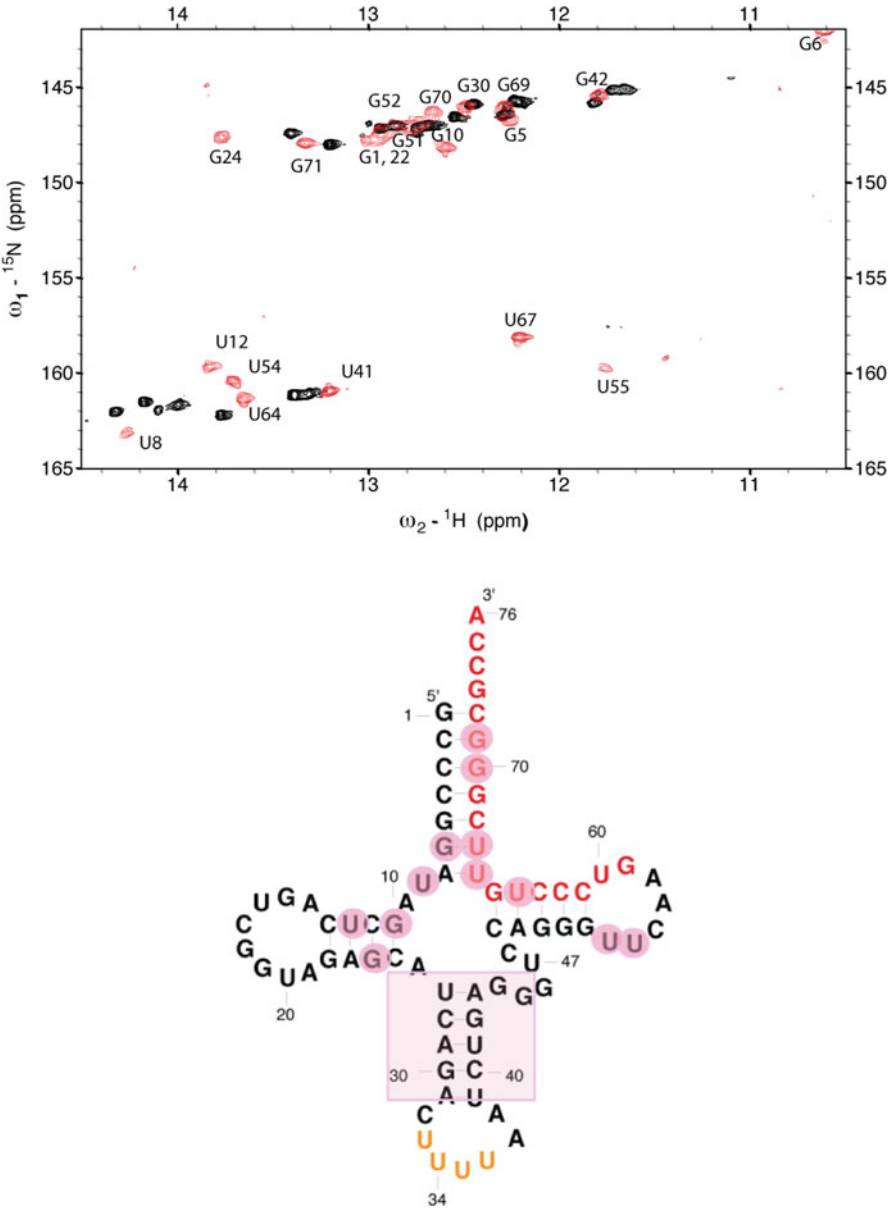


Fig. 7.11 ^{15}N-^{1}H TROSY spectrum of uniformly ^{13}C, ^{15}N labeled human $tRNA_3^{Lys}$ either alone (*red*) or in 1:1 complex with unlabeled HIV-1 69nt RNA (*black*). The comparison of the two spectra highlights the large changes in tRNA structure that occurs upon complex formation. Major perturbations of the folded tRNA spectrum upon complex formation with HIV RNA occur in the acceptor, T and D-stems, whereas resonances for the anticodon stem are relatively unperturbed. The spectra were acquired at 800 MHz in 10 mM $MgCl_2$ 50 mM NaCl, 10 mM Na phosphate, pH 6.5 at 25°C using a BEST-TROSY pulse sequence as described in the text for optimized signal-to-noise

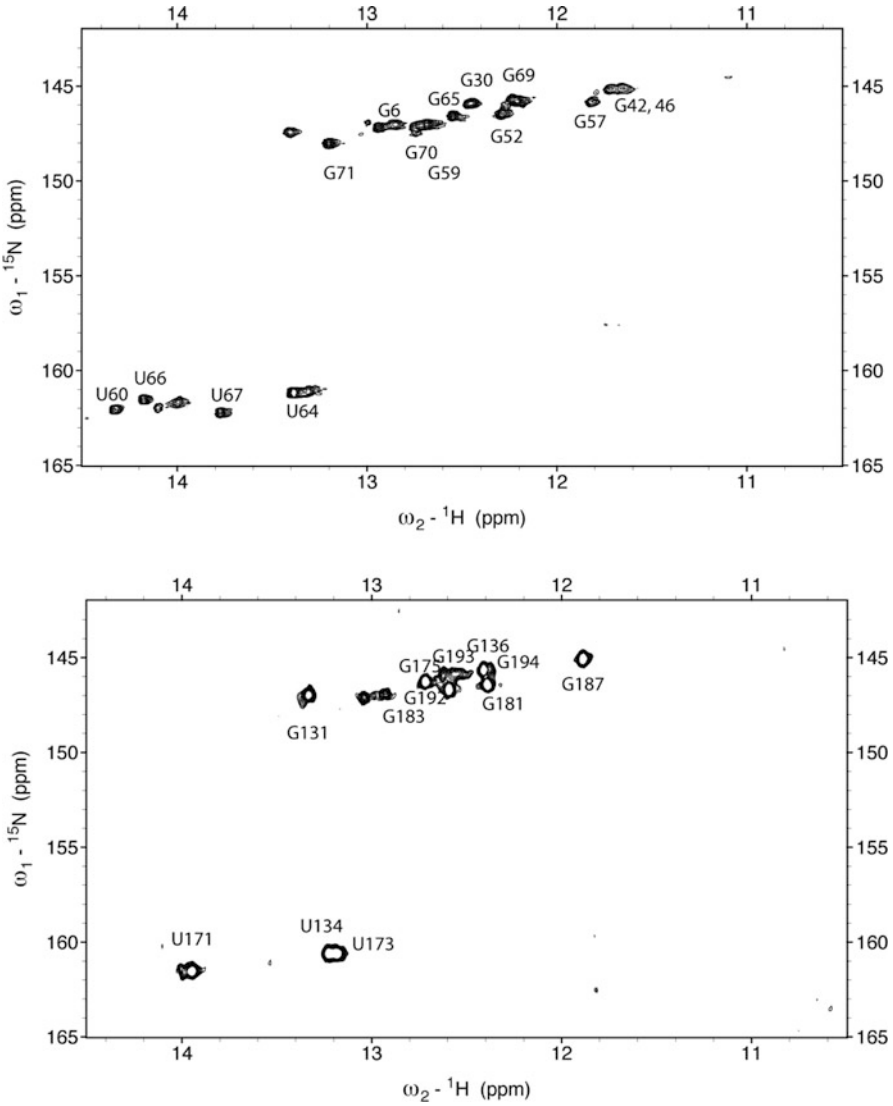


Fig. 7.12 ^{1}H-^{15}N TROSY experiments for the HIV initiation complex at 1:1 $tRNA_3^{Lys}$-HIV RNA stoichiometry. (*Left*) Spectrum for initiation complex formed with uniformly labeled ^{13}C, ^{15}N-$tRNA_3^{Lys}$ and unlabeled HIV-1 69nt RNA. (*Right*). Spectrum for initiation complex formed with uniformly labeled ^{13}C, ^{15}N-HIV-1 69nt RNA and unlabeled $tRNA_3^{Lys}$. Spectra were acquired at 800 MHz ^{1}H frequency at 25°C. Spectral assignments of the imino resonances as discussed in the text are indicated

this helix was confirmed by synthesis of an oligonucleotide containing the 18 base pair helix in isolation, capped at one end by a UUCG tetraloop. Assignment of the base pairing imino ^{1}H spectrum of this RNA agreed with those of the same helix in the HIV-tRNA complex.

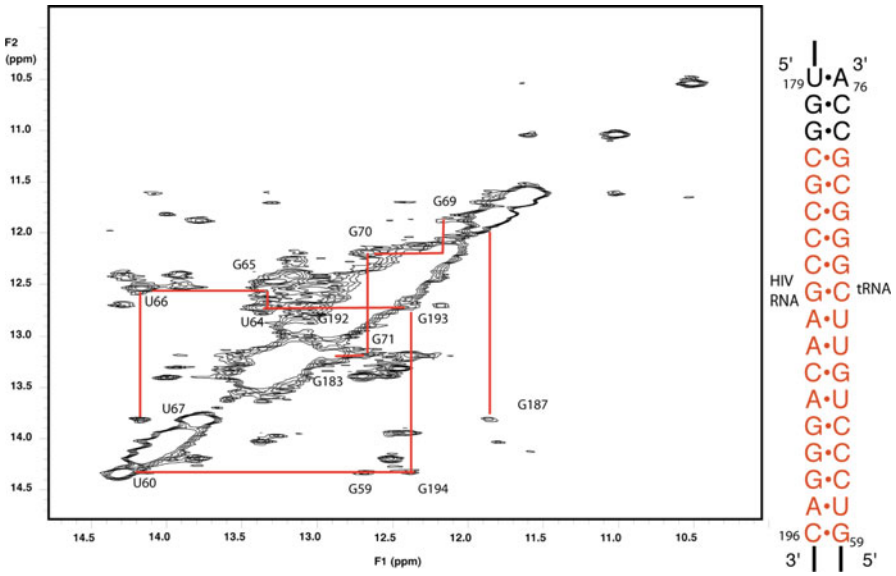


Fig. 7.13 NOESY spectrum of unlabeled 1:1 complex of human $tRNA_3^{Lys}$ and HIV1 69nt RNA. The region of imino ^1H-^1H NOES is shown; data were acquired at 800 MHz with 100 ms mixing time at 25°C. NOE connectivities are observed that allow assignment of 16 of the 18 base pairs formed between the tRNA and HIV RNA

The remaining secondary structural elements of the initiation complex were assigned using heteronuclear NMR on complexes with one labeled component. First, assignment of the 18 bp helix was confirmed by the sorting of imino protons from either tRNA or HIV RNA, as determined by the ^{15}N-TROSY experiments on complexes with either uniformly ^{15}N-labeled tRNA or HIV RNA. Intramolecular base pairing within an individual RNA component was identified by direct measurement of ^{15}N-^{15}N scalar couplings across base pairs using HNN-COSY experiments [7]. We confirmed formation of an intramolecular RNA-RNA helix within HIV RNA in the initiation complex. The helical connectivity extends from base pairs C132-G175 until U140-A168. The existence of this helix is supported by imino ^1H NOESY crosspeaks, and the observed correlation peaks in the HNN NOESY experiment on a sample with ^{13}C,^{15}N-labeled HIV RNA and unlabeled $tRNA_3^{Lys}$ (Fig. 7.13a). Using the opposite labeling strategy with tRNA, HNN COSY experiments confirmed the presence of extensive intra-tRNA base pairing within the initiation complex (Fig. 7.13b). The anticodon stem (U27-A43 to G30-C40) and D-stem (G10-C25 to C13-G22) are still formed in the complex. The 5′-end of the acceptor stem now pairs with a region formerly in the T-loop. A helix interrupted by a single-nucleotide bulge is formed from nucleotides G1 to A9 pairing with nucleotides G45 to U54. This pairing is confirmed by both HNN-COSY crosspeaks and imino-imino proton NOEs.

7.7 Discussion and Conclusions

The combined biochemical and NMR approach outlined here as allowed the definition of the secondary structure of the HIV initiation complex (Fig. 7.14). We first delineated a tractable system for NMR. The initiation complex between HIV RNA and $tRNA_3^{Lys}$ was formed biochemically by hybridization of the two RNAs using heat annealing. A 69 nt fragment of the HIV 1 Mal isolate primer binding site is well-behaved biophysically and allows efficient hybridization of either naturally modified or unmodified transcript human $tRNA_3^{Lys}$ with 1:1 stoichiometry.

We applied both homonuclear and heteronuclear NMR to define the secondary structure of the 50 kDa HIV initiation complex (Fig. 7.14). Imino proton resonances were assigned and intermolecular hydrogen bonding and NOEs used to map helical base pairs. The secondary structure of the initiation complex determined by NMR shows multiple features that may modulate its function (Fig. 7.8). Nucleotides

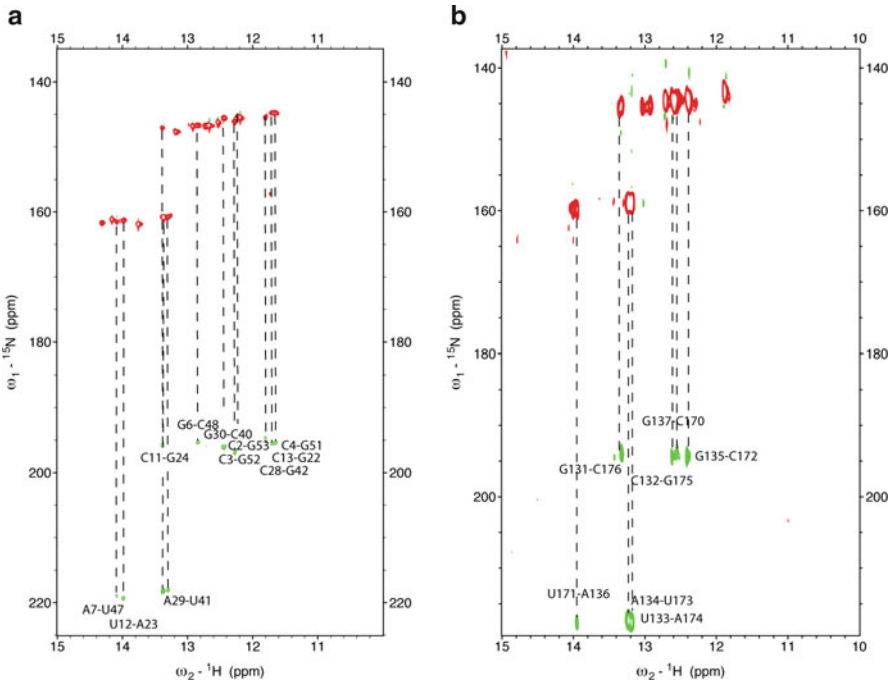


Fig. 7.14 H-NN COSY experiments showing through-space ^{15}N-^{15}N scalar couplings across base-pairing hydrogen bonds. (*Left*) H-NN COSY spectrum fo the initiation complex formed with uniformly labeled ^{13}C, ^{15}N-$tRNA_3^{Lys}$ and unlabeled HIV-1 69nt RNA. (*Right*) H-NN COSY spectrum for initiation complex formed with uniformly labeled ^{13}C, ^{15}N-HIV-1 69nt RNA and unlabeled $tRNA_3^{Lys}$. Both experiments only show correlations for intramolecular base pairing for the ^{15}N-labeled component of the complex (Data were acquired at 25°C at 800 MHz using a T1-relaxation optimized (BEST-HNN COSY) version that uses band selective pulses centered on the imino proton resonances as described in the materials and methods)

179–196 of HIV genomic RNA pair with nts 69–76 at the 3′ end of tRNA. This stable intermolecular 18 base pair primer helix binds in the active site cleft of reverse transcriptase, positioning HIV nucleotide G178 as the template for the first round of DNA synthesis. Downstream from the 18 bp primer helix, there is a 3 bp stretch of single-stranded nucleotides followed by an intramolecular helix in HIV RNA [14] with base pairs from nts C132-G175 to G137-C170. $tRNA_3^{Lys}$ undergoes significant remodeling of its secondary structure upon initiation complex formation. The 5′ end of the tRNA and T-stem region pair with each other in the initiation complex; nts G1 to A9 form an imperfect duplex with nucleotide 45–54. The anticodon and D-stems remained base paired as in the tRNA secondary structure.

The secondary structure of the initiation complex confirms prior predictions made by a variety of biochemical and modeling experiments [18]. The 18 base pair helix between the tRNA and HIV is a central feature of the secondary structure, as demonstrated by many biochemical and viral replication experiments. Formation of this thermodynamically stable intermolecular helix drives structural rearrangements in both the HIV genomic RNA and $tRNA_3^{Lys}$.

The tRNA rearrangement that occurs upon formation of the initiation complex explains the choice of $tRNA_3^{Lys}$ as a primer in HIV. The stable pairing between the T and acceptor stems drives refolding of the tRNA sequence upon initiation complex formation. The refolded tRNA structure maintains the U-rich anticodon stem loop to make potential contacts with the A-rich loop within the HIV RNA. Although we do not observe imino proton resonances consistent with anticodon-A-loop interaction, chemical probing and recent NMR experiments [3], suggest that such a pairing occurs. Rapid solvent exchange often broadens imino resonances from weakly paired helical regions.

The intramolecular helix formed by HIV genomic RNA just downstream from the 18 bp PBS helix represents a block to reverse transcriptase. This helix is site of two reverse transcription pauses during minus-strand strong-stop DNA synthesis at positions +3 and +6. (see Fig. 7.14); rates of reverse transcription are seven-fold slower for the first 15 nts compared to subsequent nucleotides, with limited pausing also after the potential A-rich sequence-anticodon loop interactions. The RNA interactions observed in our structure present a plausible explanation for the slow elongation rates during the early phase of reverse transcription, and may explain the limited extension observed in the preformed initiation complex within the HIV virion.

Recent single-molecule fluorescence experiments on HIV RT bound to the initiation complex have shown how the RNA structural features in the initiation complex control protein-RNA dynamics [22]. The initiation complex is recognized by HIV reverse transcriptase, with the 18 bp helix bound within the cleft of HIV RT [13, 38]. This would place the downstream HIV helix adjacent to the active site. Using fluorescence resonance energy transfer (FRET), Legrice, Zhuang and co-workers found dynamic switching between correct, active initiation complex orientation within the binding cleft of RT, and an inactive orientation in which the +1 position is flipped 180° and located near the RNAase H domain. The presence

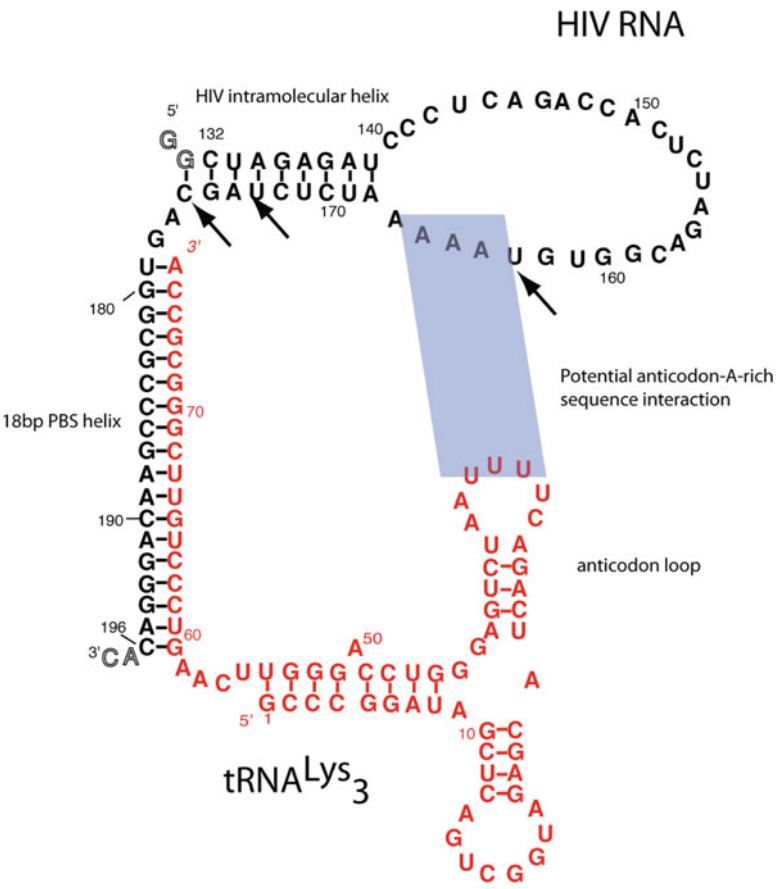


Fig. 7.15 Secondary structure of the HIV initiation complex determined here by NMR spectroscopy. HIV 1 genomic RNA (Mal-1 isolate) is in *black*, $tRNA_3^{Lys}$ in *red*. The key regions of the structure, including 18 bp PBS helix, HIV genomic RNA intramolecular helix, and the conservation of the tRNA anticodon stem-loop are highlighted. Potential U-rich anticodon loop interaction with A-rich sequence in the HIV RNA is *shaded*. *Arrows* indicate pausing sites observed in kinetic investigations of the HIV-1 initiation complex with reverse transcriptase. These strong pauses are at positions +3, +6 and +16 (color figure onlinle)

of the intramolecular HIV helix stabilizes RT in the incorrect orientation, explaining the poor kinetics of elongation through this helix; once RT traverses the +6 position, the active orientation of RT is greatly favored Fig. 7.15.

The results presented here highlight the power of modern NMR experiments to map global RNA folds, and demonstrate how RNA structure may control the activity of reverse transcription. More detailed structural data on the RNA initiation complex using NMR and crystallography are required to drive further mechanistic understanding and development of novel therapeutics targeting initiation.

Acknowledgements The authors thank Dr. Insil Kim, Joren Retel and Dr. Corey Liu for assistance in sample preparation and NMR spectroscopy. Supported by GM69314. The Stanford Magnetic Resonance Laboratory is partially supported by the Stanford University School of Medicine.

References

1. Bajji AC, Sundaram M, Myszka DG, Davis DR (2002) An RNA complex of the HIV-1 A-loop and tRNA(Lys,3) is stabilized by nucleoside modifications. J Am Chem Soc 124:14302–14303
2. Baudin F, Marquet R, Isel C, Darlix JL, Ehresmann B, Ehresmann C (1993) Functional sites in the 5′ region of human immunodeficiency virus type 1 RNA form defined structural domains. J Mol Biol 229:382–397
3. Bilbille Y, Vendeix FA, Guenther R, Malkiewicz A, Ariza X, Vilarrasa J, Agris PF (2009) The structure of the human tRNALys3 anticodon bound to the HIV genome is stabilized by modified nucleosides and adjacent mismatch base pairs. Nucleic Acids Res 37:3342–3353
4. D'Souza V, Dey A, Habib D, Summers MF (2004) NMR structure of the 101-nucleotide core encapsidation signal of the Moloney murine leukemia virus. J Mol Biol 337:427–442
5. D'Souza V, Summers MF (2004) Structural basis for packaging the dimeric genome of Moloney murine leukaemia virus. Nature 431:586–590
6. Davis JH, Tonelli M, Scott LG, Jaeger L, Williamson JR, Butcher SE (2005) RNA helical packing in solution: NMR structure of a 30 kDa GAAA tetraloop-receptor complex. J Mol Biol 351:371–382
7. Dingley AJ, Grzesiek S (1998) Direct observation of hydrogen bonds in nucleic acid base Pairs by Internucleotide 2JNN couplings. J Am Chem Soc 120:8293–8297
8. Dominguez C, Schubert M, Duss O, Ravindranathan S, Allain FH (2011) Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 58:1–61
9. Goldschmidt V, Ehresmann C, Ehresmann B, Marquet R (2003) Does the HIV-1 primer activation signal interact with tRNA3(Lys) during the initiation of reverse transcription? Nucleic Acids Res 31:850–859
10. Goldschmidt V, Paillart JC, Rigourd M, Ehresmann B, Aubertin AM, Ehresmann C, Marquet R (2004) Structural variability of the initiation complex of HIV-1 reverse transcription. J Biol Chem 279:35923–35931
11. Goldschmidt V, Rigourd M, Ehresmann C, Le Grice SF, Ehresmann B, Marquet R (2002) Direct and indirect contributions of RNA secondary structure elements to the initiation of HIV-1 reverse transcription. J Biol Chem 277:43233–43242
12. Hart JM, Kennedy SD, Mathews DH, Turner DH (2008) NMR-assisted prediction of RNA secondary structure: identification of a probable pseudoknot in the coding region of an R2 retrotransposon. J Am Chem Soc 130:10233–10239
13. Huang H, Chopra R, Verdine GL, Harrison SC (1998) Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. Science 282:1669–1675
14. Isel C, Ehresmann C, Keith G, Ehresmann B, Marquet R (1995) Initiation of reverse transcription of HIV-1: Secondary structure of the HIV-1 RNA/$tRNA^{Lys}$ (template/primer) complex. J Mol Biol 247:236–250
15. Isel C, Keith G, Ehresmann B, Ehresmann C, Marquet R (1998) Mutational analyses of the $tRNA_3^{Lys}$/HIV-1 RNA (primer/template) complex. Nucl Acids Res 26:1198–1204
16. Isel C, Lanchy JM, Le Grice SF, Ehresmann C, Ehresmann B, Marquet R (1996) Specific initiation and switch to elongation of human immunodeficiency virus type 1 reverse transcription require the post-transcriptional modifications of primer tRNA3Lys. EMBO J 15:917–924
17. Isel C, Marquet R, Keith G, Ehresmann C, Ehresmann B (1993) Modified nucleotides of $tRNA_3^{Lys}$ modulate primer/template loop-loop interaction in the initiation complex of HIV-1 reverse transcription. J Biol Chem 268:25269–25272

18. Isel C, Westhof E, Massire C, Le Grice SF, Ehresmann B, Ehresmann C, Marquet R (1999) Structural basis for the specificity of the initiation of HIV-1 reverse transcription. EMBO J 18:1038–1048
19. Kang SM, Wakefield JK, Morrow CD (1996) Mutations in both the U5 region and the primer-binding site influence the selection of the tRNA used for the initiation of HIV-1 reverse transcription. Virology 222:401–414
20. Leroy JL, Kochoyan M, Huynh-Dinh T, Gueron M (1988) Characterization of base-pair opening in deoxynucleotide duplexes using catalyzed exchange of the imino proton. J Mol Biol 200:223–238
21. Liang C, Rong L, Gotte M, Li X, Quan Y, Kleiman L, Wainberg MA (1998) Mechanistic studies of early pausing events during initiation of HIV-1 reverse transcription. J Biol Chem 273:21309–21315
22. Liu S, Harada BT, Miller JT, Le Grice SF, Zhuang X (2010) Initiation complex dynamics direct the transitions between distinct phases of early HIV reverse transcription. Nat Struct Mol Biol 17:1453–1460
23. Lu K, Miyazaki Y, Summers MF (2010) Isotope labeling strategies for NMR studies of RNA. J Biomol NMR 46:113–125
24. Lukavsky PJ, Kim I, Otto GA, Puglisi JD (2003) Structure of HCV IRES domain II determined by NMR. Nat Struct Biol 10:1033–1038
25. Lukavsky PJ, Puglisi JD (2005) Structure determination of large biological RNAs. Methods Enzymol 394:399–416
26. McConnell B (1978) Exchange mechanisms for hydrogen bonding protons of cytidylic and guanylic acids. Biochemistry 17:3168–3176
27. McConnell B (1984) The amino 1H resonances of oligonucleotide helices: d(CGCG). J Biomol Struct Dyn 1:1407–1421
28. McConnell B (1986) General acid–base catalysis in nucleobase amino proton exchange: cytidine. J Biomol Struct Dyn 4:419–436
29. McConnell B, Hoo DL (1982) 1H-NMR lifetimes of cytosine interactions with the DNA melting probe, methylmercury. Chem Biol Interact 39:351–362
30. McConnell B, Politowski D (1984) Buffer catalysis of amino proton exchange in compounds of adenosine, cytidine and their endocyclic N-methylated derivatives. Biophys Chem 20:135–148
31. McConnell B, Rice DJ, Uchima FD (1983) Exceptional characteristics of amino proton exchange in guanosine compounds. Biochemistry 22:3033–3037
32. Miyazaki Y, Irobalieva RN, Tolbert BS, Smalls-Mantey A, Iyalla K, Loeliger K, D'Souza V, Khant H, Schmid MF, Garcia EL et al (2010) Structure of a conserved retroviral RNA packaging element by NMR spectroscopy and cryo-electron tomography. J Mol Biol 404:751–772
33. Moazed D, Stern S, Noller HF (1986) Rapid chemical probing of conformation in 16 S ribosomal RNA and 30 S ribosomal subunits using primer extension. J Mol Biol 187:399–416
34. Mougel M, Houzet L, Darlix JL (2009) When is it time for reverse transcription to start and go? Retrovirology 6:24
35. Puglisi EV, Puglisi JD (1998) HIV-1 A-rich RNA loop mimics the tRNA anticodon structure. Nat Struct Biol 5:1033–1036
36. Puglisi EV, Puglisi JD (2007) Probing the conformation of human tRNA(3)(Lys) in solution by NMR. FEBS Lett 581:5307–5314
37. Puglisi JD, Putz J, Florentz C, Giege R (1993) Influence of tRNA tertiary structure and stability on aminoacylation by yeast aspartyl-tRNA synthetase. Nucleic Acids Res 21:41–49
38. Sarafianos SG, Marchand B, Das K, Himmel DM, Parniak MA, Hughes SH, Arnold E (2009) Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. J Mol Biol 385:693–713
39. Smallcombe S (1993) Solvent suppression with symmetrically-shifted pulses. J Am Chem Soc 115:4776–4785
40. Tzakos AG, Grace CR, Lukavsky PJ, Riek R (2006) NMR techniques for very large proteins and rnas in solution. Annu Rev Biophys Biomol Struct 35:319–342

41. Wakefield JK, Kang SM, Morrow CD (1996) Construction of a type 1 human immunodeficiency virus that maintains a primer binding site complementary to tRNA(His). J Virol 70:966–975
42. Wakefield JK, Morrow CD (1996) Mutations within the primer binding site of the human immunodeficiency virus type 1 define sequence requirements essential for reverse transcription. Virology 220:290–298
43. Wakefield JK, Rhim H, Morrow CD (1994) Minimal sequence requirements of a functional human immunodeficiency virus type 1 primer binding site. J Virol 68:1605–1614
44. Wakefield JK, Wolf AG, Morrow CD (1995) Human immunodeficiency virus type 1 can use different tRNAs as primers for reverse transcription but selectively maintains a primer binding site complementary to tRNA(3Lys). J Virol 69:6021–6029
45. Yusupova G, Lanchy JM, Yusupov M, Keith G, Le Grice SF, Ehresmann C, Ehresmann B, Marquet R (1996) Primer selection by HIV-1 reverse transcriptase on RNA-tRNA(3Lys) and DNA-tRNA(3Lys) hybrids. J Mol Biol 261:315–321

Chapter 8
Approaches to Protein-Ligand Structure Determination by NMR Spectroscopy: Applications in Drug Binding to the Cardiac Regulatory Protein Troponin C

Ian M. Robertson, Sandra E. Pineda-Sanabria, and Brian D. Sykes

Abstract NMR spectroscopy is an effective tool employed by medicinal chemists in the drug discovery pipeline. NMR spectroscopy is convenient because it can provide structural information relatively easily. For example, chemical shift mapping has been a tool employed for many years to identify ligand binding sites and to determine the stoichiometry and affinity of ligand binding. However, the determination of a high resolution solution structure of a target-drug complex can be much more laborious and time consuming. This is especially inconvenient in comparison with the crystal soak difference Fourier methods used for X-ray structures. Over the years, we have sought methods which are more rapid when the general overall structure of the target is known. We discuss the application of some of these methods herein; including a number of *in silico* methods developed to augment traditional NOE based NMR methods.

8.1 Introduction

A key element in rational drug design is the rapid determination of protein-ligand structures. X-ray crystallography and NMR Spectroscopy are the most commonly used techniques to determine high-resolution structures of protein-ligand complexes. While NMR spectroscopy has the advantage that different drugs can easily be titrated into the protein in solution, the determination of the full high resolution solution structure involves the assignment of thousands of nuclear Overhauser effect contacts (NOEs) and other data, which tends to be a time-consuming task. If the structure of a target protein is known and the NMR spectra of the ^{13}C,^{15}N-labeled protein are assigned, the measurement of intermolecular

I.M. Robertson • S.E. Pineda-Sanabria • B.D. Sykes (✉)
Department of Biochemistry, University of Alberta, Edmonton, AB T6G 2H7, Canada
e-mail: ian.robertson@kcl.ac.uk; pinedasa@ualberta.ca; brian.sykes@ualberta.ca.

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_8,

NOEs between the ligand and the protein may be enough to faithfully localize the orientation of the ligand in the complex. This procedure can drastically reduce the overall time for determining the orientation of the ligand on its target, while still providing invaluable structural details about the important pharmacophores of a molecule. There are situations, however, when this method is still too laborious (*e.g.* for high-throughput applications) or not possible (*e.g.* for a low affinity ligand, resulting in weak of NOEs).

NMR spectroscopy is unique from other screening and structural methods in that it can determine binding constants as well as identify the ligand binding site on a protein by monitoring residue specific changes in chemical shifts upon ligand binding. This approach is called chemical shift mapping and it is routinely used for rapidly identifying the binding site of a ligand on a protein and is a critical step in SAR-by-NMR [1]. Chemical shift mapping is most typically done by following the amide nuclei of an ^{15}N-labeled target molecule, and then highlighting the residues that are significantly shifted on the van der Waals surface of the target protein. These results can also be used as ambiguous distance restraints by docking programs, such as HADDOCK [2] or the mapped surface can be used for a binding-site centered docking with an automated docking program such as AutoDock [3]. Since chemical shift perturbations are a response to a change in the local chemical environment induced by ligand binding, it is possible to predict the location of the ligand rings from the magnitude and sign of the chemical shift changes if we assume these changes are caused predominantly by proximity to aromatic groups of a ligand. Finally, the use of a protein with a paramagnetic center can provide distance restraints between the ligand and the protein. A number of groups have investigated the efficacy of combining a variety of these methods to accurately and rapidly resolve protein-ligand structures [4–7]. In order to examine the usefulness of these techniques, we shall study their applications in regards to drug binding into the cardiac muscle contractile regulatory protein, troponin C (cTnC).

cTnC is a dumbbell shaped Ca^{2+}-binding protein that triggers contraction upon Ca^{2+}-binding its regulatory N-lobe (cNTnC). Ca^{2+} enters the cytosol of a muscle cell and binds to cNTnC, which induces a slight opening of cNTnC. Following Ca^{2+} association with cNTnC, the "switch region" of the protein troponin I (cTnI) binds to cNTnC, stabilizing the fully open conformation of cNTnC. The "inhibitory region" of cTnI is subsequently released from actin, which results in the contraction-dependent actin-myosin interaction. During relaxation, Ca^{2+} dissociates from cNTnC as it leaves the cytosol, cTnI translocates from cNTnC back to actin, which blocks the actin-myosin interface to inhibit contraction. The C-domain of cTnC (cCTnC) has two high-affinity metal binding sites that are occupied throughout the contraction-relaxation cycle. The binding partner of cCTnC is a different segment of cTnI, the "anchoring region". The primary role of the cCTnC-cTnI interaction is to tether cTnC to the thin filament, but recent insights into the structure and function of insect flight muscle troponin suggest this interaction may also play a role in regulation of contraction [8, 9]. For recent reviews into the molecular mechanism of contraction, see [10–12]. Given that both the C- and

N-terminal domains of cTnC interact with cTnI to regulate contraction, the design of small molecules that alter either interaction hold therapeutic promise [13–15].

8.2 Intermolecular NOEs

The identification of NOEs between a protein and a ligand is a crucial step in solving the structure of a complex by NMR spectroscopy. Knowledge of the high-resolution structure of a protein-ligand complex can provide insights into the mechanism of a ligand as well as help in the design of new pharmaceuticals. The standard scheme of determining a protein-ligand structure by NMR spectroscopy begins by assigning the resonances of a ^{13}C,^{15}N- or ^{15}N-labeled protein. Following the resonance assignment, the structure of the protein is solved by assigning intramolecular NOEs to obtain distance restraints. Once the tertiary structure of the target is known, the unlabeled ligand is assigned when in complex with the protein using a variety of isotope filtered experiments [16–18]. Finally, NOE contacts between the unlabeled ligand and labeled target are measured to obtain distance restraints between the ligand and protein [19–21]. Frequently, the protein structure is not significantly altered upon binding to the ligand making the re-determination of the protein structure superfluous. In these cases, it may be sufficient to use only intermolecular NOEs to provide distance restraints between the labeled protein and unlabeled ligand. Unfortunately, the target molecule does often undergo structural changes when binding a ligand. If there has been a structure of the target molecule in complex with a related ligand, it may be possible to use intermolecular NOEs to localize the ligand onto the complex structure, with the homologous ligand removed. We have recently applied this technique to study a complex with cNTnC and the small molecule, W7 [22]. The structure of cNTnC was obtained from the NMR solution structure of cNTnC-cTnI [23], with cTnI removed prior to the docking of W7. The complex we will focus on in this report is cCTnC bound to the polyphenol, resveratrol (Fig. 8.1a).

Resveratrol is an antioxidant present at high concentrations in red wine and in the skins of grapes [24] and it has been studied intensely due to its benefits to cardiovascular health [25]. The low incidence of cardiovascular disease among Mediterranean populations despite a diet rich in fat, has prompted some researchers to suggest that the high consumption of red wine offers protection against heart disease, and has been dubbed, the “French paradox” [26, 27]. We have found that resveratrol targets cCTnC, and this interaction may impart some of its protective properties [28]. The structure of cCTnC bound to the green tea polyphenol, epigallocatechin gallate (EGCg) [29], the anchoring region of cTnI ($cTnI_{34–71}$) [30, 31], and the cardiotonic drug EMD 57033 [32] have been solved by either NMR spectroscopy or X-ray crystallography. Even though the ligands of cCTnC are chemically dissimilar, they all induce an analogous conformational change

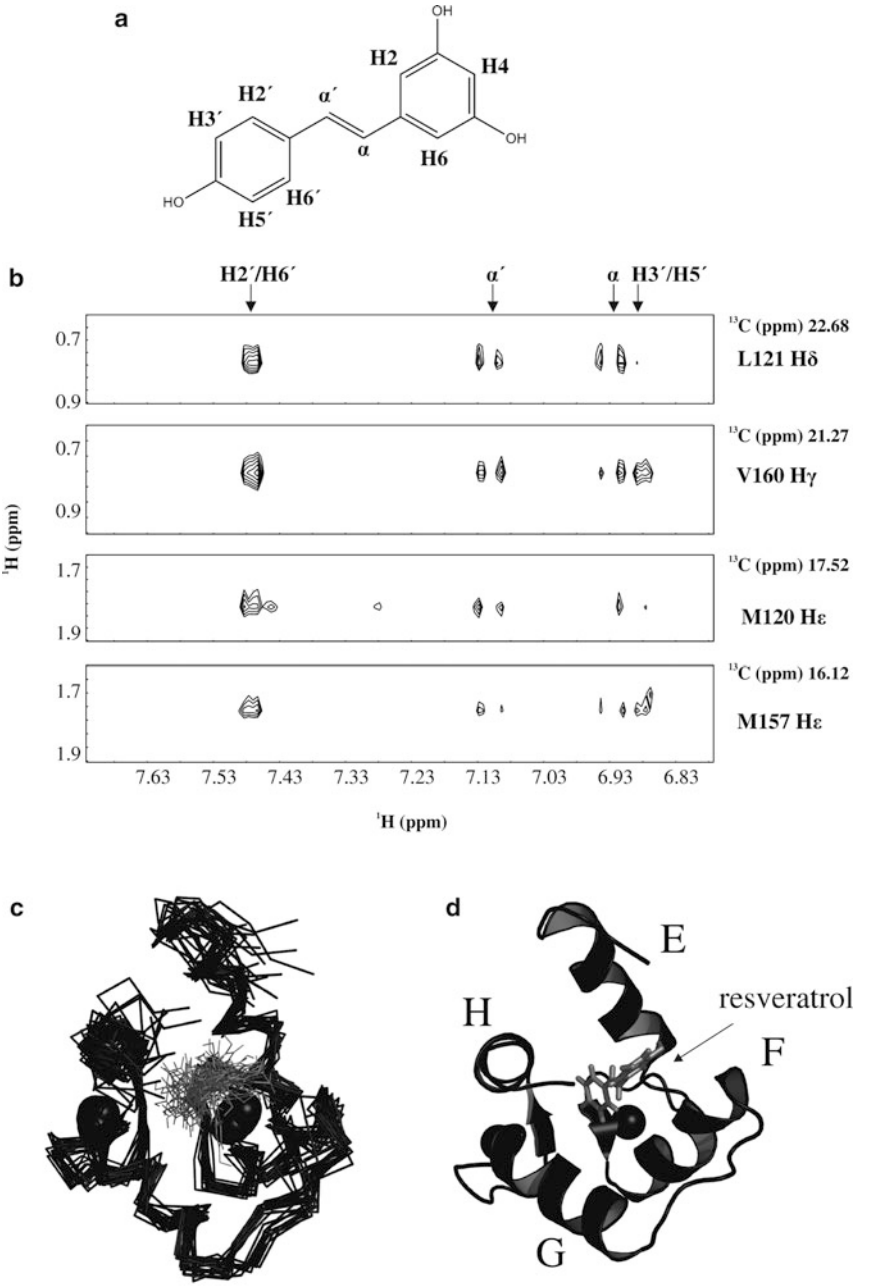


Fig. 8.1 Structure of cCTnC in complex with resveratrol as determined by intermolecular NOEs. (**a**) chemical structure of resveratrol. (**b**) several strips from the ^{13}C-edited-filtered NOESY NMR experiment that identifies NOEs between an unlabeled ligand and a ^{13}C-labeled protein. (**c**) Ensemble of the 20 lowest energy structures of cCTnC-resveratrol. (**d**) the lowest energy structure from the ensemble. cCTnC is shown in *black* with the helices labeled and resveratrol is colored in *grey*. *Black spheres* represent Ca^{2+}

of cCTnC [29]. Instead of solving the structure of cCTnC again, we measured intermolecular NOEs between resveratrol and cCTnC and docked resveratrol onto the cCTnC-EGCg structure (Fig. 8.1b-d). In order to dock resveratrol, we repeated the simulated annealing procedure in Xplor-NIH [33, 34] using the intramolecular NOEs for cCTnC from the cCTnC-EGCg structure, and replaced the intermolecular NOEs from EGCg to cCTnC with restraints between resveratrol and cCTnC. We propose that this procedure is better than simply performing a rigid docking since it allows for some flexibility of cCTnC. We chose the cCTnC-EGCg structure because both EGCg and resveratrol are polyphenols and thus may induce similar structural perturbations of cCTnC.

Monitoring strong protein-ligand complexes can be a fruitful source of structural information. However, problems can arise when the unlabeled ligand is too large so that it contains an abundance of overlapping signals, or when the intermolecular NOEs are weak, making it difficult to distinguish noise from NOEs. Therefore, it is often necessary to make use of other methods to assist in the determination of protein-ligand structures. We will assess some of these approaches, by comparing the predictions from these methods with the cCTnC-resveratrol structure as determined with the NOE data.

8.3 Chemical Shift Mapping

Chemical shift mapping is a frequently used application of NMR spectroscopy which allows the estimation of a binding site of a ligand on a target molecule. Typically the perturbations of ^{15}N-amide chemical shifts are used, which can provide information on the binding location of the compound on the protein as well as its binding constant. While chemical shift mapping is a rapid method to estimate the binding surface on a protein, it has several pitfalls, including a lack of resolution, and is hampered by any allosteric effects induced by a ligand binding. This is because chemical shifts are perturbed in response to a change in local chemical environment, which may be due to proximity to a ligand or because of a conformational change induced by the ligand binding.

The use of amide chemical shift mapping with the interaction between cCTnC and resveratrol is shown in Fig. 8.2a. Although the perturbations seem to point towards the central groove as the binding site of resveratrol, the largest shifts are from residues residing in the F-G linker, which are not near to resveratrol in the structure (Fig. 8.1c, d). The comparison of the residues perturbed by resveratrol, with those induced by other ligands of cCTnC: EGCg, EMD 57033, and $cTnI_{34\text{-}71}$ reveals that all these residues induce the same pattern of amide chemical shift change (Fig. 8.2c). These ligands have unique structural features, and while they all bind in the core of the protein the orientations they adopt are quite varied. The similar amide chemical shift perturbations are most likely from a similar conformational perturbation since all ligand bound forms of cCTnC in an open state (the F-G and E-H helix pairs are moved apart to accommodate the ligands).

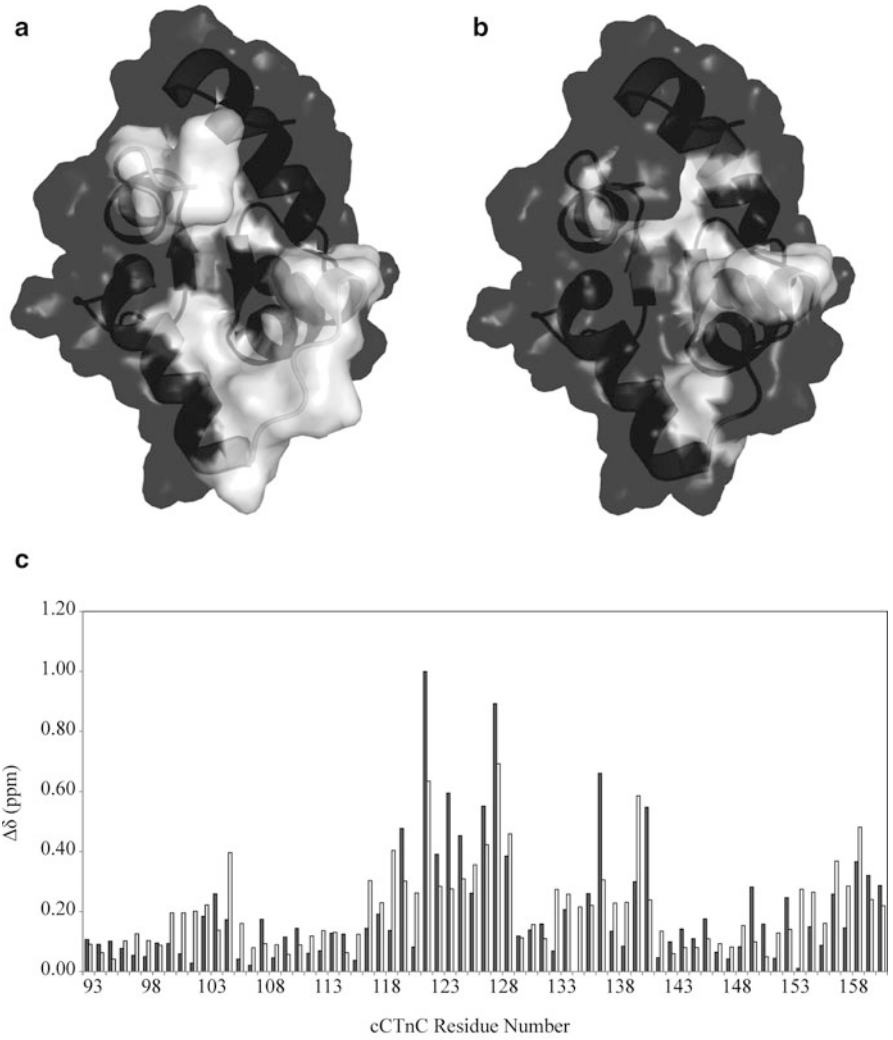


Fig. 8.2 Chemical shift mapping of resveratrol induced shifts on cCTnC from the cCTnC-EGCg structure (2kdh.pdb). (**a**) ^{15}N-amide chemical shifts that were perturbed larger than the mean chemical shift perturbation. (**b**) residues that were perturbed in the methyl region of the ^{1}H,^{13}C-HSQC spectrum. cCTnC is shown as a transparent surface (*black*), with the perturbed resonances highlighted in *white*. (**c**) bar diagram that compares the total amide chemical shift ($\Delta\delta$) pattern induced by resveratrol (*black bars*) and the average chemical shift perturbation induced by EMD 57033, $cTnI_{34\text{-}71}$, and EGCg (*white bars*). $\Delta\delta = ((\Delta\delta^{1}\mathrm{H})^2 + 1/25(\Delta\delta^{15}\mathrm{N})^2)^{1/2}$

Although the amide shifts may they tell us very little about the specific binding site of resveratrol, they do give us some idea of the conformation of cCTnC.

Another way to map the binding surface of a ligand on a target molecule is to look at the ^{13}C-methyl chemical shift perturbations. Since most protein-ligand structures

are stabilized via hydrophobic interactions, following methyl perturbations when a ligand binds to the protein may be more informative that using ^{15}N-amide chemical shifts. In Fig. 8.2b we show mapped residues that underwent perturbations in the terminal methyls. Most of the perturbed residues, are on the E-, F-, and H-helices, as well as on the β-sheet of cCTnC. These results are more consistent with the structure we determined with intermolecular NOEs. In fact, many of the methyls that were perturbed during the titration were also involved in making NOEs with resveratrol. Chemical shift mapping of the methyls clearly does a better job than amide nuclei of identifying the binding site of resveratrol, but the predictions are still qualitative.

8.4 J-Surface Mapping

To supplement the qualitative chemical shift mapping, we have mapped the J-surface of resveratrol on cCTnC. The ring current effect from aromatic ligands can induce chemical shift perturbations of nuclei proximal to the ligand. McCoy and Wyss described the application of this phenomenon to identify the binding surface of a ligand to the hepatitis C virus NS3 protease [35]. Caveats of using this method to identify the ligand binding site are that the ligand must contain an aromatic group and that the ring current effect from the ligand is the primary source of chemical shift perturbation. In order to localize the binding site of resveratrol on cCTnC we performed J-surface mapping with the program Jsurf using the amide or methyl chemical shift perturbation data (Fig. 8.3). Jsurf approximates the origin of chemical shift perturbations as a single point-dipole at the center of the aromatic ring of a ligand [36]. The centre of the aromatic ring is then predicted based on the magnitude and sign of the proton perturbation. For each perturbed residue, the aromatic ring inducing the chemical shift change can be at any number of places and for each additional residue considered this location is narrowed down until finally a surface is identified where the ligand rings is likely to reside.

We used the amide protons chemical shift perturbations to see if Jsurf could accurately predict the binding of resveratrol (Fig. 8.3a). The predicted J-surface runs through the F and G helices, which is not consistent with the NOEs and for resveratrol to lie here would violate many van der Waals forces. While it is likely that some of the amide resonances perturbed in the presence of resveratrol are due to ring current effects, evidently many are caused not by direct contact with the ring, but rather by an allosteric perturbation induced by resveratrol binding. We repeated the J-surface analysis using the methyl protons (Fig. 8.3b). The result of using the methyl protons was a binding site localized to the hydrophobic pocket in a very similar location as calculated by the intermolecular NOEs. As an aside, Jsurf requires an input file that has atom numbers from the protein structure file (i.e. PDB file format) with the corresponding shift for that atom. In contrast to amide protons, there are three protons associated with each methyl. So the input file we used for Jsurf had three protons for each methyl perturbation included in the calculation.

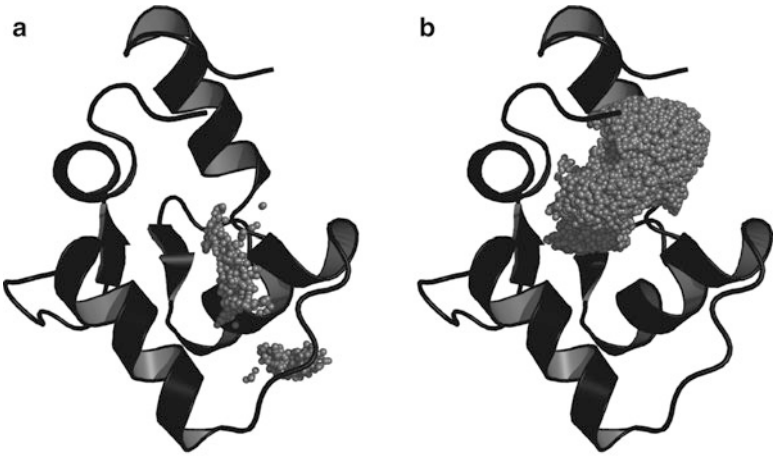


Fig. 8.3 The J-surface (*grey spheres*) from the amide (**a**) and methyl (**b**) chemical shift perturbations mapped on the structure of cCTnC (*black cartoon*) from the cCTnC-EGCg structure (2kdh.pdb)

8.5 Automated Docking

Although Jsurf does a reasonable job of predicting the binding site of resveratrol on cCTnC using the methyl proton perturbations, it does not predict the actual binding pose of resveratrol. In order to do this, we used AutoDock4.2 which predicts the conformations of a ligand bound to a macromolecular target of known structure as well as the free energy of binding. AutoDock4.2 allows flexibility for both the ligand and for a limited number of residues in the target protein indicated by the user, it also allows the user to select a region of the target molecule to perform the conformation searching (the grid volume) [37, 38]. AutoDock 4.2 with the AutoDockTools graphical interface was used to predict 150 binding conformations using a population size of 500 and 5,000,000 energy evaluations. The docking was guided by limiting the grid volume to encompass residues of cCTnC that the chemical shift and J-surface mappings predicted to be involved in binding resveratrol. Residues which showed chemical shifts greater than one standard deviation from the mean in the $^{1}H,^{13}C$ HSQC NMR experiment were defined as flexible residues.

A total of 15 clusters were predicted for resveratrol using an r.m.s.d. tolerance of 2.0 Å for each cluster. All 150 structures predict that resveratrol binds in the hydrophobic groove of cCTnC (Fig. 8.4a). This is not surprising, considering the grid volume chosen from the experimental data encompassed primarily this region of cCTnC. In AutoDock, identifying the lowest-energy cluster and the most populated cluster are two ways of choosing the best conformer [39, 40]. The ten lowest energy structures are shown in Fig. 8.4b, and include two different

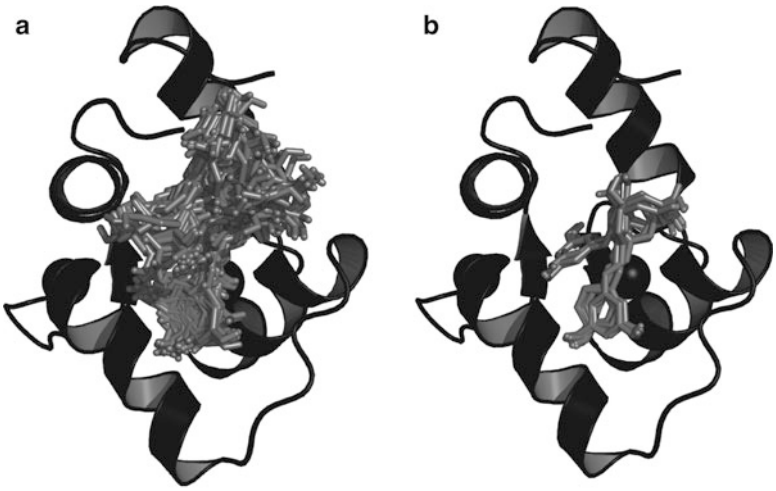


Fig. 8.4 Overlay of the poses of resveratrol (*grey sticks*) on cCTnC (*black cartoon*) from the cCTnC-EGCg structure (2kdh.pdb). (**a**) all 150 poses and (**b**) the ten lowest energy poses generated using AutoDock

conformations of resveratrol as having the lowest energy. The lowest energy conformer has resveratrol oriented parallel to the β-sheet of cCTnC, whereas the next lowest energy cluster predicts resveratrol binds in the hydrophobic pocket of cCTnC pointed towards the β-sheet. The buried structure of resveratrol more closely resembles the methyl-based Jsurf prediction and NOE data; however, the NOE data has the diphenol ring oriented away from the protein, not deep within the hydrophobic pocket as was calculated by AutoDock.

8.6 Restraints Derived from Paramagnetic Restraints

The prediction of the binding site of resveratrol on cCTnC by chemical shift mapping, J-surface mapping, and docking had some success; but, in the end it would be difficult to validate the results without less ambiguous data, such as NOEs. The use of paramagnetic relaxation enhancement (PRE) has been used extensively to aid in protein structure determination [41], in protein-protein structure determination [42], and in protein-ligand structure determination [43]. PRE is a phenomenon which arises when unpaired electrons from a paramagnetic center, such as a lanthanide ion, increase the relaxation of nuclei in a distance dependant manner. Pseudocontact shifts (PCS) induced by paramagnetic lanthanide ions with anisotropic magnetic susceptibility tensors can also tell the user information about the orientation of the ligand with respect to the metal. The Otting group has recently

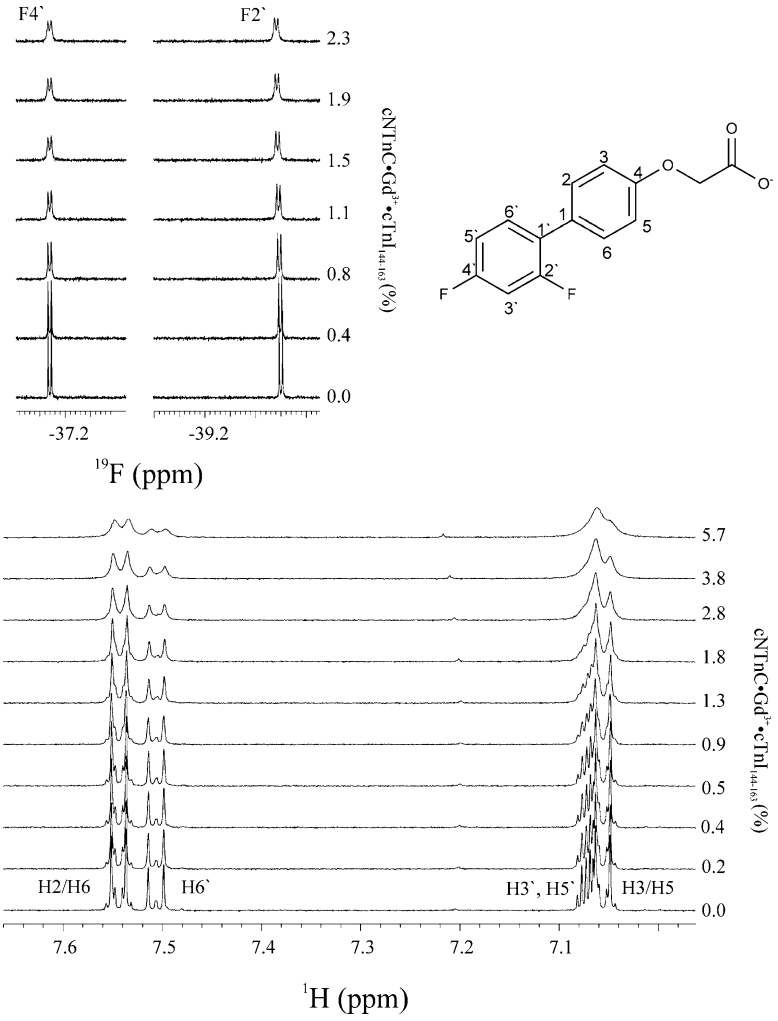


Fig. 8.5 Paramagnetic relaxation enhancement of signals of dfbp-o by Gd^{3+} bound to cNTnC-cTnI. The stacked 1D ^{19}F spectrum of dfbp-o is shown in the *upper left*, and the stacked 1D ^{1}H spectrum of dfbp-o is shown at the *bottom*. Both stacked spectra are shown as a function of percentage dfbp-o bound to cNTnC•Gd^{3+}•cTnI. The numbering of dfbp-o is shown on its chemical structure as a reference

employed PCS to determine the structure of a protein-drug complex [44]. On the other hand, the lanthanide gadolinium (Gd^{3+}) has an isotropic paramagnetic environment, so it does not yield information about the orientation between the nuclei and metal, but does provide distance information.

It has been shown that the trivalent lanthanides can substitute for calcium in troponin C [45, 46]. The C-terminal domain of troponin C binds two metal

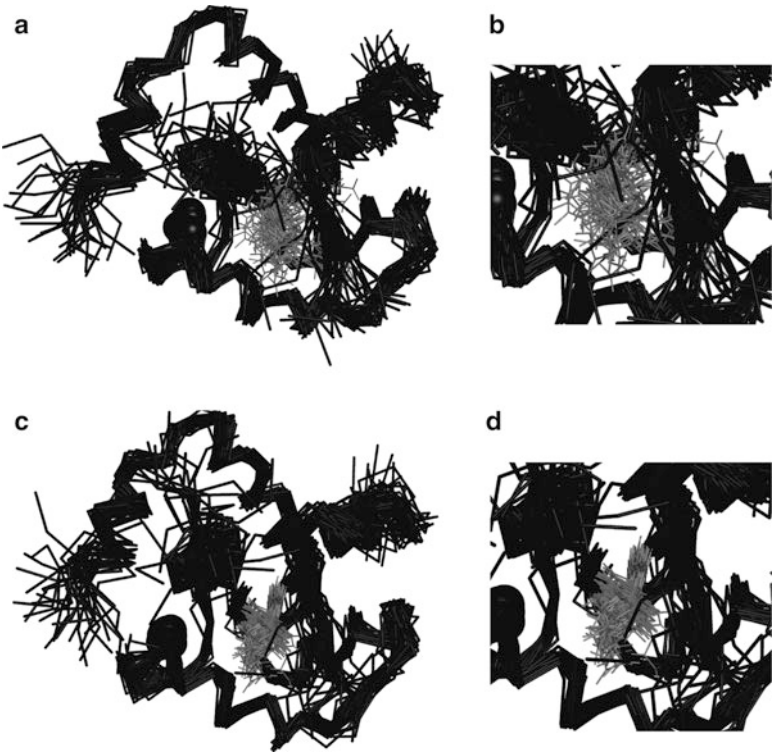


Fig. 8.6 Ribbon diagram of cNTnC bound to cTnI (both in *black*) and dfbp-o (*grey*) in stick representation. (**a**) Refinement of the complex using only intermolecular NOEs. (**b**) close-up view of the binding site, highlighting the disorder of dfbp-o. (**c**) Refinement of the complex using intermolecular NOEs and PRE distance restraints. (**d**) close-up view of the binding site in (**c**)

ions, and therefore it would be difficult to obtain structural restraints between resveratrol and both of these ions. Fortuitously, the cNTnC contains only one Ca^{2+} binding site. In a recent publication [47], we have used PRE to obtain distance restraints between Gd^{3+} and the cardiotonic agent, 2′,4′-difluoro(1,1′-biphenyl)-4-yloxy acetic acid (dfbp-o). In this type of experiment, free dfbp-o was titrated with Gd^{3+}-bound cNTnC-cTnI, and at each titration point we measured the transverse and longitudinal relaxation rates of ^{19}F and ^{1}H of dfbp-o. As the concentration of cNTnC-cTnI-Gd^{3+} increased, the relaxation rates of dfbp-o increased (see Fig. 8.5 for signal broadening), and this increases in relaxation rates were used to calculate this distance between dfbp-o and Gd^{3+} as previously described [48]. We calculated the structure of the tertiary complex, cNTnC-cTnI-dfbp-o using only NOE restraints, and NOE and PRE restraints (Fig. 8.6). The addition of the PRE restraints radically improved the resolution of dfbp-o in the structure, and illustrates the usefulness of this technique to aid in protein-ligand structure refinement.

8.7 Conclusion

The calculation of protein-ligand structures is an important step in the development of novel pharmaceuticals. The high resolution description of the binding site and pose of a ligand is a crucial step in the rationale design of novel drugs. In this review we have surveyed a few approaches available to scientists to help in the resolving of protein-ligand structures. Intermolecular NOEs are a crucial step in solving protein-ligand structures; however, they are not always attainable. Reasons for this include but are not limited to: time-constraints, no ^{13}C-labeled protein available, or a weak binding constant. There are other techniques that can validate structures solved with few NOEs, or substitute for NOEs when none exist. While ^{15}N-amide nuclei are the most used nucleus for estimating the binding site of a ligand, conformational changes in the protein upon ligand binding can be misleading. We show that the use of ^{13}C-methyl chemical shift mapping and J-surface prediction using methyl perturbations more accurately identify the binding site of a ligand. Identification of the binding pose of the ligand in the absence of NOE data can be done with AutoDock; however, it is difficult to choose the cluster that most closely resembles the real structure. Other experimentally derived restraints, such as PREs can significantly improve protein-ligand structures when either NOEs are not sufficient to define the binding pose of the ligand or when NOEs are not available.

Acknowledgements We would like to thank Dr. G. Moyna for making the source code for Jsurf available, and for helpful instructions on its use, Drs. M. Oleszczuk and M. Li for insightful discussions on drug- troponin C interactions. The authors would also like to thank David Corson for protein expression and purification; Robert Boyko and Nick Shaw for spectrometer maintenance; and Robert Boyko for in-house software development.

References

1. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Discovering high-affinity ligands for proteins: SAR by NMR. Science 274:1531–1534
2. Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737
3. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30:2785–2791
4. Bertini I, Fragai M, Giachetti A, Luchinat C, Maletta M, Parigi G, Yeo KJ (2005) Combining in silico tools and NMR data to validate protein-ligand structural models: application to matrix metalloproteinases. J Med Chem 48:7544–7559
5. Cioffi M, Hunter CA, Packer MJ, Pandya MJ, Williamson MP (2009) Use of quantitative H-1 NMR chemical shift changes for ligand docking into barnase. J Biomol NMR 43:11–19
6. Krishnamoorthy J, Yu VCK, Mok YK (2010) Auto-FACE: an NMR based binding site mapping program for fast chemical exchange protein-ligand systems. PLoS One 5:e8943
7. Cioffi M, Hunter CA, Packer MJ, Spitaleri A (2008) Determination of protein-ligand binding modes using complexation-induced changes in H-1 NMR chemical shift. J Med Chem 51:2512–2517

8. Bullard B, Agianian B, Krzic U, Linke WA, Leonard KR (2004) Independent regulation of insect flight muscle by two isoforms of troponin C. Biophys J 86:215a–216a
9. De Nicola G, Burkart C, Qiu F, Agianian B, Labeit S, Martin S, Bullard B, Pastore A (2007) The structure of Lethocerus troponin C: insights into the mechanism of stretch activation in muscles. Structure 15:813–824
10. Gomes AV, Potter JD, Szczesna-Cordary D (2002) The role of troponins in muscle contraction. IUBMB Life 54:323–333
11. Li MX, Wang X, Sykes BD (2004) Structural based insights into the role of troponin in cardiac muscle pathophysiology. J Muscle Res Cell Motil 25:559–579
12. Parmacek MS, Solaro RJ (2004) Biology of the troponin complex in cardiac myocytes. Prog Cardiovasc Dis 47:159–176
13. Kass DA, Solaro RJ (2006) Mechanisms and use of calcium-sensitizing agents in the failing heart. Circulation 113:305–315
14. Li MX, Robertson IM, Sykes BD (2008) Interaction of cardiac troponin with cardiotonic drugs: a structural perspective. Biochem Biophys Res Commun 369:88–99
15. Sorsa T, Pollesello P, Solaro RJ (2004) The contractile apparatus as a target for drugs against heart failure: interaction of levosimendan, a calcium sensitiser, with cardiac troponin c. Mol Cell Biochem 266:87–107
16. Gemmecker G, Olejniczak ET, Fesik SW (1992) An improved method for selectively observing protons attached to C-12 in the presence of H-1-C-13spin Pairs. J Magn Reson 96:199–204
17. Ikura M, Bax A (1992) Isotope-filtered 2d Nmr of a protein peptide complex – study of a skeletal-muscle myosin light chain kinase fragment bound to calmodulin. J Am Chem Soc 114:2433–2440
18. Ogura K, Terasawa H, Inagaki F (1996) An improved double-tuned and isotope-filtered pulse scheme based on a pulsed field gradient and a wide-band inversion shaped pulse. J Biomol NMR 8:492–498
19. Lee W, Revington MJ, Arrowsmith C, Kay LE (1994) A pulsed-field gradient isotope-filtered 3d C-13 Hmqc-Noesy experiment for extracting intermolecular Noe contacts in molecular-complexes. FEBS Lett 350:87–90
20. Robertson IM, Spyracopoulos L, Sykes BD (2009) The evaluation of isotope editing and filtering for protein-ligand interaction elucidation by Nmr. Biophys Challeng Emerg Threats:101–119
21. Stuart AC, Borzilleri KA, Withka JM, Palmer AG (1999) Compensating for variations in H-1-C-13 scalar coupling constants in isotope-filtered NMR experiments. J Am Chem Soc 121:5346–5347
22. Hoffman RMB, Sykes BD (2009) Structure of the inhibitor W7 bound to the regulatory domain of cardiac troponin C. Biochemistry 48:5541–5552
23. Li MX, Spyracopoulos L, Sykes BD (1999) Binding of cardiac troponin-I147-163 induces a structural opening in human cardiac troponin-C. Biochemistry 38:8289–8298
24. Jang MS, Cai EN, Udeani GO, Slowing KV, Thomas CF, Beecher CWW, Fong HHS, Farnsworth NR, Kinghorn AD, Mehta RG, Moon RC, Pezzuto JM (1997) Cancer chemopreventive activity of resveratrol, a natural product derived from grapes. Science 275:218–220
25. Corder R, Douthwaite JA, Lees DM, Khan NQ, dos Santos ACV, Wood EG, Carrier MJ (2001) Endothelin-1 synthesis reduced by red wine – red wines confer extra benefit when it comes to preventing coronary heart disease. Nature 414:863–864
26. Kopp P (1998) Resveratrol, a phytoestrogen found in red wine. A possible explanation for the conundrum of the 'French paradox'? Eur J Endocrinol 138:619–620
27. Goldberg DM, Hahn SE, Parkes JG (1995) Beyond alcohol – beverage consumption and cardiovascular mortality. Clin Chim Acta 237:155–187
28. Pineda-Sanabria SE, Robertson IM, Sykes BD (2011) Structure of trans-resveratrol in complex with the cardiac regulatory protein troponin C. Biochemistry 50:1309–1320
29. Robertson IM, Li MX, Sykes BD (2009) Solution structure of human cardiac troponin C in complex with the green tea polyphenol, (−)-Epigallocatechin 3-Gallate. J Biol Chem 284:23012–23023

30. Takeda S, Yamashita A, Maeda K, Maeda Y (2003) Structure of the core domain of human cardiac troponin in the Ca(2+)-saturated form. Nature 424:35–41
31. Gasmi-Seabrook GM, Howarth JW, Finley N, Abusamhadneh E, Gaponenko V, Brito RM, Solaro RJ, Rosevear PR (1999) Solution structures of the C-terminal domain of cardiac troponin C free and bound to the N-terminal domain of cardiac troponin I. Biochemistry 38:8313–8322
32. Wang X, Li MX, Spyracopoulos L, Beier N, Chandra M, Solaro RJ, Sykes BD (2001) Structure of the C-domain of human cardiac troponin C in complex with the Ca2+ sensitizing drug EMD 57033. J Biol Chem 276:25456–25466
33. Schwieters CD, Kuszewski JJ, Clore GM (2006) Using Xplor-NIH for NMR molecular structure determination. Prog Nucl Magn Reson Spectrosc 48:47–62
34. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65–73
35. McCoy MA, Wyss DF (2002) Spatial localization of ligand binding sites from electron current density surfaces calculated from NMR chemical shift perturbations. J Am Chem Soc 124:11758–11763
36. Moyna G, Zauhar RJ, Williams HJ, Nachman RJ, Scott AI (1998) Comparison of ring current methods for use in molecular modeling refinement of NMR derived three-dimensional structures. J Chem Inf Comput Sci 38:702–709
37. Morris GM, Goodsell DS, Huey R, Olson AJ (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J Comput Aided Mol Des 10:293–304
38. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) A semiempirical free energy force field with charge-based desolvation. J Comput Chem 28:1145–1152
39. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19:1639–1662
40. Rosenfeld RJ, Goodsell DS, Musah RA, Morris GM, Goodin DB, Olson AJ (2003) Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. J Comput Aided Mol Des 17:525–536
41. Bertini I, Donaire A, Luchinat C, Rosato A (1997) Paramagnetic relaxation as a tool for solution structure determination: clostridium pasteurianum ferredoxin as an example. Protein Struct Func Genet 29:348–358
42. Clore GM, Iwahara J (2009) Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. Chem Rev 109:4108–4139
43. Pintacuda G, John M, Su XC, Otting G (2007) NMR structure determination of protein-ligand complexes by lanthanide labeling. Acc Chem Res 40:206–212
44. John M, Pintacuda G, Park AY, Dixon NE, Otting G (2006) Structure determination of protein-ligand complexes by transferred paramagnetic shifts. J Am Chem Soc 128:12910–12916
45. Gay GL, Lindhout DA, Sykes BD (2004) Using lanthanide ions to align troponin complexes in solution: order of lanthanide occupancy in cardiac troponin C. Protein Sci 13:640–651
46. Wang CLA, Leavis PC, Dehorrocks W, Gergely J (1981) Binding of lanthanide ions to troponin-C. Biochemistry 20:2439–2444
47. Robertson IM, Sun YB, Li MX, Sykes BD (2010) A structural and functional perspective into the mechanism of Ca^{2+}-sensitizers that target the cardiac troponin complex. J Mol Cell Cardiol 49:1031–1041
48. Marsden BJ, Hodges RS, Sykes BD (1988) H-1-Nmr studies of synthetic peptide analogs of calcium-binding site-Iii of rabbit skeletal troponin-C – effect on the lanthanum affinity of the interchange of aspartic-acid and asparagine residues at the metal-ion coordinating positions. Biochemistry 27:4198–4206

Chapter 9
How Do Nascent Proteins Emerge from the Ribosome?

Ada Yonath

Abstract The ribosome is the universal cellular "factory" for producing proteins by translating the genetic code. These highly efficient polymerases posses spectacular architecture and inherent mobility, allowing their smooth performance in decoding the genetic information as well as the formation of peptide bonds, elongation of the newly born proteins and their protection (for review see [45] ChemBioChem 10:63–72, 2009).

The site for peptide bond formation (PTC) is located within a highly conserved pseudosymmetrical region [1, 7] that connects all of the remote ribosomal features involved in its functions, and seems to be a remnant of an ancient RNA machine for chemical bonding [2, 3, 8, 11, 16]. The elaborate structure of this region and the motions of its components position the aminoacylated and peptidyl tRNAs in stereochemistry required for formation of peptide bonds, for substrate-mediated catalysis, and for substrate translocation, the activities enabling the nascent chain elongation.

Adjacent to the PTC is an elongated tunnel (of a 100 Å in length and up to 2 Å width), a universal feature of the ribosome, along which the newly born proteins progress until they emerge out of the ribosome (Fig. 9.1). Although the existence of an internal ribosomal tunnel was proposed in the 1960s, based on biochemical experiments [23, 33] it was widely assumed that nascent proteins are not degraded while growing on the ribosome's surface because they adopt helix conformation. In fact, studies aimed at supporting the above assumption were carried out [32] and doubts in the mere existence of the ribosomal tunnel was publicly expressed [27]

A. Yonath (✉)
Department of Structural Biology, Weizmann Institute, Helen & Milton A. Kimmelman Building, Rehovot 76100, Israel
e-mail: ada.yonath@weizmann.ac.il

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_9,

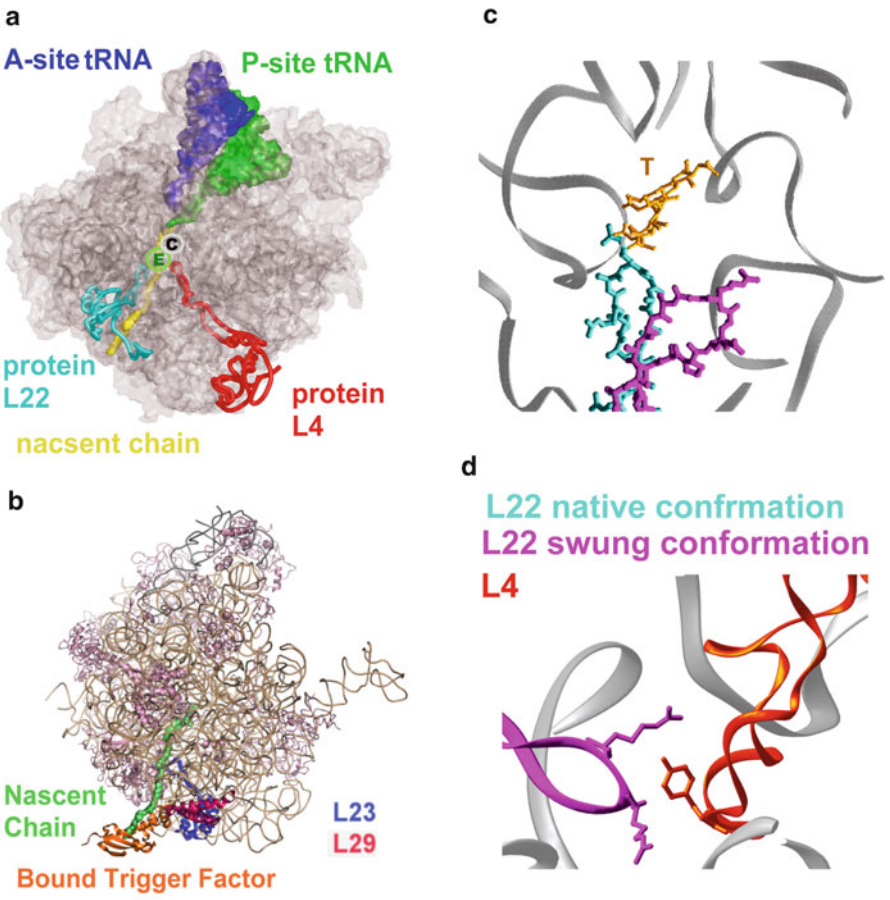


Fig. 9.1 The ribosomal tunnel: Front (**a**) and side (**b**) views of the entire large ribosomal subunits (in different shades of *gray*), with A- and P- site tRNAs (in *blue* and *green*, respectively). A modeled polyalanine in the tunnel, as well as bound trigger factor (in *orange*) are shown. Ribosomal components proteins that line various regions of the tunnel and are involved in nascent protein progression and/or its arrest, are shown in colors. C denotes a crevice where co-translational initial folding may occur [4], E shows the erythromycin pocket and T is the macrolide troleandomycin. The motions of the hairpin tip of protein L22, and the proximity of L22 swung conformation to protein L4 are shown in (**c**) and (**d**) (color figure online)

even after its initial visualization by three dimensional image reconstructions in eukaryotic and prokaryotic ribosomes, at 60 and 25 Å resolution, respectively [26, 43], in the 1980s, until it was rediscovered and verified by cryo electron microscopy [17, 36].

The existence of a tunnel was finally verified when it was observed in the high resolution crystal structures of the large ribosomal subunit [5, 19]. Simultaneously, the tunnel was also shown to provide the binding pocket of the main clinically

relevant antibiotics family that targets the ribosomes (Fig. 9.1a), which hamper the progression of nascent proteins through it [34]. Nonetheless, despite its uneven shape, including a wide crevice and a narrow constriction, originally tunnel involvement in the fate of the nascent chains was hard to conceive. Therefore, it was suggested, by at least one leading group, to be a passive conduit for the growing nascent chain with a Teflon-like character, with no obvious chemical properties allowing its interactions with progressing nascent chains [5, 31].

Nevertheless, further studies indicated tunnel active participation in nascent chain progression and its compaction, and diverse functional roles of the ribosomal tunnel are currently emerging and evidence for nascent proteins-tunnel interactions are being accumulated (e.g. for review see [24]), some of which are extensively involved in translation arrest and cellular signaling. Thus, indications for possible tunnel active participation in nascent chain initial compaction, leading to semi-folded chain segments were accumulated [e.g. [14, 21, 39–41]], and crystallographic studies identified a crevice where co-translational initial folding may occur [4] (Fig. 9.1a). Additionally, distinct conformations including helical segments of the nascent polypeptide chains were visualized within several regions of the ribosomal exit tunnel that have been implicated in nascent chain–ribosome interaction [10]. It was also shown biochemically and by simulation studies that initial protein folding can occur in the tunnel without steric overlap with the tunnel walls and a "folding zone" for early folding was identified within the last 20 Å of the exit tunnel [38]. Regardless if the state of early folding, when emerged out of the ribosome, nascent chains may interact with chaperones that assist their folding and/or prevent their aggregation and misfolding. In eubacteria the first chaperone that encounters the nascent proteins (Fig. 9.1b), called trigger factor, forms a shelter composed of hydrophobic and hydrophilic regions that provide environment that can compete with the aggregation tendency of the still unfolded chains [6, 35, 22]. Interestingly, free trigger factor seems to rescue proteins from misfolding and accelerate protein folding [25].

The tunnel walls are lined predominantly by ribosomal RNA. Ribosomal proteins L22 and L23 are two non-RNA tunnel wall components that are likely to control the tunnel gating and/or trafficking (Fig. 9.1). While protein L23 resides at the tunnel opening and in eubacteria it appears to have sufficient mobility for possible controlling the emergence of the newly born proteins [6], protein L22 is involved in elongation arrest. Self elongation arrest and its mutual impact on cellular processes gained increased interest in recent years. Suggestion concerning the mechanisms inducing arrest, simple tunnel jamming at its narrowest constriction cannot be the only mechanism for the arrest, since it would have interfered with the progression of all proteins carrying bulky amino acids. Alternatively, the arrest segment may assume specific folds within the tunnel, which can prevent nascent protein progression along it. It is conceivable that such semi-folded segments could also inhibit peptide bond formation or hinder tRNA translocation. However, more likely is the notion that arrest occurs as a consequence of conformational alterations in the tunnel walls, caused by the semi folded segment, as suggested for the SecM and TnaC systems [15, 18, 28–30]. Diverse self stalling sequences that sense cellular

physiology that stall ribosomes during their own translation, hence control protein biosynthesis have been identified in prokaryotes and eukaryotes [20]. Although the various stalling segments show little sequence similarity, they seem to be involved in extensive communication between the ribosomes and the cells [42] through distinct mechanisms and induce translation arrest that regulate downstream gene expression [37].

Among the striking diversity of arrest-inducing peptides found in many organisms, the most studied example is the SecM (secretion monitor) protein. This protein contains an amino acid segment (FXXXXWIXXXXGIRAGP), which was found to be required and sufficient to cause translation arrest during elongation [28]. This segment seems to be capable of interacting with ribosomal interior components, thereby interfering with its own translation elongation, and providing a regulatory mechanism for the expression of the genetic information. It was proposed that the interactions of the SecM protein trigger significant conformational alterations of the tunnel walls, in which the main actor is protein L22. This protein has an elongated structure, with its hairpin tip residing at the narrowest constriction of the exit tunnel (Fig. 9.1). This protein spans more of half of the tunnel environs, so that its C- and N-terminals reside in the vicinity of the tunnel's opening, hence can sense cellular signals. Its other end, namely the tip of its hair pin is capable of swinging across the tunnel as revealed in the crystal structure of the large ribosomal subunit in complex with the macrolide antibiotic troleandomycin [9].

By flipping across the tunnel, L22 hairpin tip can hamper nascent protein progression, thus acting as a tunnel gate. In support of this proposed mechanism is the finding that arrest caused by the SecM arrest sequence is bypassed by mutations in L22 hairpin tip region as well as in the 23 S rRNA nucleotides that interact with L22 in its swung conformation [28]. The significant influence of L22 conformation on the tunnel wall shape is evident by the finding that although protein L22 is located in the vicinity but does not belong to the macrolide binding pocket [34], it appears to be involved erythromycin resistance. Thus, minute sequence alterations in the tip of the hairpin of protein L22 or in protein L4, which resides in proximity to the swung L22 (Fig. 9.1d), were shown to trigger significant displacements of the RNA components of the tunnel walls (Wekselman, private communication) so that the tunnel can host erythromycin while allowing the progression of nascent chains [44, 12].

SecM-elongation arrest occurs when the transient sequence reaches the PTC and the N-terminal region of the protein has already emerged outside the ribosome, and can interact with the SecM protein export machinery [28]. Noteworthy, although this arrest sequence was identified in the SecM system, it can function as translation arresting element even when inserted into unrelated sequences. Furthermore, additional studies show involvement of specific protein sequences in the regulation of various cell processes. Examples are the control of SecA and SecM translation by protein secretion [29]. Strikingly, recent studies showed that in eukaryotes the regulation of membrane protein biogenesis, the nascent chains together with the ribosome may act as cellular sensors while progressing through the tunnel [13].

Thus, the currently available observations imply direct interactions between specific residues of the leader peptide with distinct locations in the ribosomal tunnel in prokaryotes and eukaryotes.

Collectively, the above findings explain why overproduction of signal sequence-defective SecM in *E. coli* is extremely toxic and suggests that specific genomic sequences have evolved in order to avoid sequences that interact nonproductively with ribosomal components [29]. They also indicate that protein L22 appears to have dual functions: act as a cellular sensor as well as a progression barrier and that the nascent protein exit tunnel seems to be responsible for cell-ribosome signaling mechanisms alongside intra-ribosomal regulation processes, governing the fate of nascent proteins expression.

Acknowledgements Thanks are due to all members of the ribosome groups at the Weizmann Institute and the Max Planck Society for their experimental efforts and illuminating discussion. Support was provided by the US National Inst. of Health (GM34360), the German Ministry for Science and Technology (BMBF 05-641EA), GIF 853–2004, Human Frontier Science Program (HFSP) RGP0076/2003 and the Kimmelman Center for Macromolecular Assemblies. AY holds the Martin and Helen Kimmel Professorial Chair. X-ray diffraction data were collected the EMBL and MPG beam lines at DESY; F1/CHESS, Cornell University, SSRL/Stanford University, ESRF/EMBL, Grenoble, BL26/PF/KEK, Japan and 19ID&23ID/APS/Argonne National Laboratory.

References

1. Agmon I, Bashan A, Zarivach R, Yonath A (2005) Symmetry at the active site of the ribosome: structural and functional implications. Biol Chem 386:833–844
2. Agmon I (2009) The dimeric proto-ribosome: structural details and possible implications on the origin of life. Int J Mol Sci 10:2921–2934
3. Agmon I, Bashan A, Yonath A (2006) On ribosome conservation and evolution Isr. J Ecol Evol 52:359–379
4. Amit M, Berisio R, Baram D, Harms J, Bashan A, Yonath A (2005) A crevice adjoining the ribosome tunnel: hints for cotranslational folding. FEBS Lett 579:3207–3213
5. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289:905–920
6. Baram D, Pyetan E, Sittner A, Auerbach-Nevo T, Bashan A, Yonath A (2005) Structure of trigger factor binding domain in biologically homologous complex with eubacterial ribosome reveals its chaperone action. Proc Natl Acad Sci USA 102:12017–12022
7. Bashan A, Agmon I, Zarivach R, Schluenzen F, Harms J, Berisio R, Bartels H, Franceschi F, Auerbach T, Hansen HA, Kossoy E, Kessler M, Yonath A (2003) Structural basis of the ribosomal machinery for peptide bond formation, translocation, and nascent chain progression. Mol Cell 11:91–102
8. Belousoff MJ, Davidovich C, Zimmerman E, Caspi Y, Wekselman I, Rozenszajn L, Shapira T, Sade-Falk O, Taha L, Bashan A, Weiss MS, Yonath A (2010) Ancient machinery embedded in the contemporary ribosome. Biochem Soc Trans 38:422–427
9. Berisio R, Schluenzen F, Harms J, Bashan A, Auerbach T, Baram D, Yonath A. (2003) Structural insight into the role of the ribosomal tunnel in cellular regulation. Nat Struct Biol 10:366–370

10. Bhushan S, Gartmann M, Halic M, Armachel J, Jaraschl A, Mielke T, Berninghausen O, Wilson DN, Beckmann R (2010) Helical segments of the nascent polypeptide chains visualized within distinct regions of the ribosomal exit tunnel. Nat Struct Mol Biol 17:313–317
11. Bokov K, Steinberg SV (2009) A hierarchical model for evolution of 23 S ribosomal RNA. Nature 457:977–980
12. Bommakanti AS, Lindahl L, Zengel JM (2008) Mutation from guanine to adenine in 25 S rRNA at the position equivalent to E. coli A2058 does not confer erythromycin sensitivity in Sacchromyces cerevisae. RNA 14:460–464
13. Chiba S, Lamsa A, Pogliano K (2009) A ribosome-nascent chain sensor of membrane protein biogenesis in Bacillus subtilis. EMBO J 28:3461–3475
14. Crowley KS, Reinhart GD, Johnson AE (1993) The signal sequence moves through a ribosomal tunnel into a noncytoplasmic aqueous environment at the ER membrane early in translocation. Cell 73:1101–1105
15. Cruz-Vera LR, Gong M, Yanofsky C (2006) Changes produced by bound tryptophan in the ribosome peptidyl transferase center in response to TnaC, a nascent leader peptide. Proc Natl Acad Sci USA 103:3598–3603
16. Davidovich C, Belousoff M, Bashan A, Yonath A (2009) The evolving ribosome: from non-coded peptide bond formation to sophisticated translation machinery. Res Microbiol 160: 487–492
17. Frank J, Zhu J, Penczek P, Li Y, Srivastava S, Verschoor A, Radermacher M, Grassucci R, Lata RK, Agrawal RK (1995) A model of protein synthesis based on cryo-electron microscopy of the E. coli ribosome. Nature 376:441–444
18. Gong F, Yanofsky C (2002) Instruction of translating ribosome by nascent Peptide. Science 297:1864–1867
19. Harms J, Schluenzen F, Zarivach R, Bashan A, Gat S, Agmon I, Bartels H, Franceschi F, Yonath A (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. Cell 107:679–688
20. Ito K, Chiba S, Pogliano K (2010) Divergent stalling sequences sense and control cellular physiology. Biochem Biophys Res Commun 39:1–6
21. Johnson AE, Jensen RE (2004) Barreling through the membrane. Nat Struct Mol Biol 11: 113–114
22. Kaiser CM, Chang HC, Agashe VR, Lakshmipathy SK, Etchells SA, Hayer-Hartl M, Hartl FU, Barral JM (2006) Real-time observation of trigger factor function on translating ribosomes. Nature 444:455–460
23. Malkin L, Rich A (1967) Partial resistance of nascent polypeptide chains to proteolytic digestion due to ribosomal shielding. J Mol Biol 26:329–346
24. Mankin AS (2006) Nascent peptide in the “birth canal” of the ribosome. Trends Biochem Sci 31:3–11
25. Martinez-Hackert E, Hendrickson WA (2009) Promiscuous substrate recognition in folding and assembly activities of the trigger factor chaperone. Cell 138:923–934
26. Milligan RA, Unwin PN (1986) Location of exit channel for nascent protein in 80 S ribosome. Nature 319:693–695
27. Moore PB (1988) The ribosome returns. Nature 331:223–227
28. Nakatogawa H, Ito K (2002) The ribosomal exit tunnel functions as a discriminating gate. Cell 108:629–636
29. Nakatogawa H, Murakami A, Ito K (2004) Control of SecA and SecM translation by protein secretion. Curr Opin Microbiol 7:145–150
30. Nakatogawa H, Ito K (2004) Intraribosomal regulation of expression and fate of proteins. Chembiochem 5:48–51
31. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA (2000) The structural basis of ribosome activity in peptide bond synthesis. Science 289:920–930
32. Ryabova LA, Selivanova OM, Baranov VI, Vasiliev VD, Spirin AS (1988) Does the channel for nascent peptide exist inside the ribosome? Immune electron microscopy study. FEBS Lett 226:255–260

33. Sabatini DD, Blobel G (1970) Controlled proteolysis of nascent polypeptides in rat liver cell fractions. II. Location of the polypeptides in rough microsomes. J Cell Biol 45:146–157
34. Schluenzen F, Zarivach R, Harms J, Bashan A, Tocilj A, Albrecht R, Yonath A, Franceschi F (2001) Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. Nature 413:814–821
35. Schluenzen F, Wilson DN, Tian P, Harms JM, McInnes SJ, Hansen HA, Albrecht R, Buerger J, Wilbanks SM, Fucini P (2005) The binding mode of the trigger factor on the ribosome: implications for protein folding and SRP interaction. Structure (Camb) 13:1685–1694
36. Stark H, Mueller F, Orlova EV, Schatz M, Dube P, Erdemir T, Zemlin F, Brimacombe R, van Heel M (1995) The 70S Escherichia coli ribosome at 23A resolution: fitting the ribosomal RNA, Structure 3:815–821
37. Tanner DR, Cariello DA, Woolstenhulme CJ, Broadbent MA, Buskirk AR (2009) Genetic identification of nascent peptides that induce ribosome stalling. J Biol Chem 284:34809–34818
38. Tu LW, Deutsch C (2010) A folding zone in the ribosomal exit tunnel for Kv1.3 Helix formation. J Mol Biol 396:1346–1360
39. Walter P, Johnson AE (1994) Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. Annu Rev Cell Biol 10:87–119
40. Woolhead CA, Johnson AE, Bernstein HD (2006) Translation arrest requires two-way communication between a nascent polypeptide and the ribosome. Mol Cell 22:587–598
41. Woolhead CA, McCormick PJ, Johnson AE (2004) Nascent membrane and secretory proteins differ in FRET-detected folding far inside the ribosome and in their exposure to ribosomal proteins. Cell 116:725–736
42. Yap MN, Bernstein HD (2009) The plasticity of a translation arrest motif yields insights into nascent polypeptide recognition inside the ribosome tunnel. Mol Cell 34:201–211
43. Yonath A, Leonard KR, Wittmann HG (1987) A tunnel in the large ribosomal subunit revealed by three-dimensional image reconstruction. Science 236:813–816
44. Zaman S, Fitzpatrick M, Lindahl L, Zengel J (2007) Novel mutations in ribosomal proteins L4 and L22 that confer erythromycin resistance in E. coli. Mol Microbiol 66:1039–1050
45. Zimmerman E, Yonath A (2009) Biological implications of the ribosome's stunning stereochemistry. Chembiochem 10:63–72

Chapter 10
Course Abstracts

Nanobodies as Crystallization Chaperones for Mouse Prion Protein (Poster)

Romany N.N. Abskharon • Alexandre Wohlkonig • Sameh Soror • Els Pardon • Han Remaut • Hassan El Hassan • Jan Steyaert (✉)
VIB Department of Molecular and Cellular Interactions, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium
e-mail: Jan.Steyaert@vub.ac.be

Gabriele Giachin • Alessandro Didonna • Giuseppe Legname
Scuola Internazionale Superiore di Studi Avanzati – International School for Advanced Studies (SISSA-ISAS), Institute for Neurodegenerative Diseases, Trieste, Italy

Abstract Prions are fatal neurodegenerative transmissible agents causing many diseases (e.g. Creutzfeldt-Jakob disease in human, spongiform encephalopathy in bovine and scrapie in sheep). The structural investigation of prions is challenging due to the protein intrinsic disorder (mostly located in the N-terminal region). We used PrP specific nanobodies derived from Camelid antibodies to determine the structure of full-length mouse PrP^C (23–230) by x-ray crystallography. Initial attempts to co-crystallize full-length mouse PrP^C in complex with Nb484 did not produce crystals. However, limited *in-situ* proteolysis produced plate-like crystals which diffracted to 2.7 Å resolution. Here we report the characterization at molecular level of the interaction of mouse PrP^C and a nanobody (Nb484).

Our goal is to use nanobodies as a molecular tool to obtain a better understanding of the mechanism of the amyloidogenic disease formation. We also crystallized Nb484 alone and collected a complete dataset at high resolution (1.2 Å). Our experiments suggest that the nanobodies can be used as a molecular tool by helping the crystallization and by inhibiting the PrP^C to PrP^{Sc} transition.

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8_10,

Role of Non-covalent Intermolecular Interactions in Docking Guanine-Ligands (Oral Presentation)

Elisabeth D. Balitskaya • Timothy V. Pyrkov • Roman G. Efremov (✉)
Russian Academy of Sciences, M.M. Shemyakin & Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Ul. Miklukho-Maklaya, 16/10, 117997 GSP, Moscow V-437, Russia
e-mail: r-efremov@yandex.ru

Elisabeth D. Balitskaya
Department of Bioengineering, M.V. Lomonosov Moscow State University, Biological Faculty, Leninskie Gori, 1/73, 119991 GSP-1, Moscow, Russia

Abstract Guanin-containing ligands play essential role in vital metabolism pathways. There are a lot of antiviral drugs that contain guanin as a basic structure. Due to the property of viruses to adapt to various antibiotics there is a necessity to search for new and to upgrade already existing drugs. The initial step in this process is often to dock drugs and their prototypes (ligands) to the active site of target proteins (receptors).

Using program PLATINUM (http://model.nmr.ru/platinum/) we research structural and functional features of a molecular recognition of guanin-containing ligands by receptors. We demonstrate that guanine binding hydrogen-bonding motifs play essential role in this process along with hydrofobic and stacking interactions. It was found from an analisys of guanine-containing protein structures from the Protein Data Bank.

Besides we have designed a set of filters including empirical score which consists of terms describing separate hydrogen bonds (with atoms N2 and N6 of guanine), stacking and hydrophobic interactions. These criteria provide more precise identification of true orientation of a ligand in the protein active site in comparison with scoring functions implemented in program package GOLD (Jones et al., J Mol Biol 267, 727–748, 1997).

Solid-State NMR Studies on Beta-2-Microglobulin Crystals and Fibrils (Poster)

Emeline Barbet-Massin • Jozef Lewandowski • Lyndon Emsley • Guido Pintacuda (✉)
Centre de RMN à Très Hauts Champs, Université de Lyon (CNRS/ENS Lyon/UCB Lyon 1), 69100 Villeurbanne, France
e-mail: guido.pintacuda@ens-lyon.fr

Stefano Ricagno • Martino Bolognesi
Biotechnology Laboratory, IRCCS Fondazione Policlinico San Matteo, Pavia, Italy

Stefano Ricagno • Sofia Giorgetti • Vittorio Bellotti
Department of Biochemistry, University of Pavia, Pavia, Italy

Stefano Ricagno
Department of Biomolecular Sciences and Biotechnology, University of Milano, Milan, Italy

Abstract Elucidating the fine structure of amyloid fibrils as well as understanding their processes of nucleation and growth remains a difficult yet essential challenge, directly linked to our current poor insight into protein misfolding and aggregation diseases. Here we consider beta-2-microglobulin (β2m), the MHC-1 light chain component responsible for dialysis-related amyloidosis, which can give rise to amyloid fbrils *in vitro* under various experimental conditions, including low and neutral pH. We have used solid-state NMR to probe the structural features of fibrils formed by full-length β2m (99 residues) at pH 2.5 and pH 7.4. A close comparison of 2D ^{13}C-^{13}C and ^{15}N-^{13}C correlation experiments performed on β2m, both in the crystalline and fibrillar states, suggests that in spite of structural changes affecting the protein loops linking the protein β-strands, the protein chain retains a substantial share of its native secondary structure in the fibril assembly. Moreover, variations in the chemical shifts of the key Pro32 residue suggest the involvement of a *cis-trans* isomerization in the process of β2m fibril formation. Lastly, the analogy of the spectra recorded on β2m fibrils grown at different pH values hints at a conserved architecture of the amyloid species thus obtained.

Thermal Stabilization of DMPC/DHPC Bicelles by Addition of Cholesterol Sulfate (Poster)

Rebecca Shapiro • Amanda Brindley • Rachel Martin (✉)
Chemical Biology Program, Department of Chemistry, Natural Sciences Institute, University of California, Irvine, USA
e-mail: rwmartin@uci.edu

Abstract Bicelles made from a mixture of short- and long-chain phospholipids are an important orienting medium in NMR studies of biomolecules. In the present study, doping DMPC/DHPC bicelles with cholesterol sulfate is found to increase the temperature range over which stable alignment occurs. Cholesterol sulfate, a minor component of mammalian membranes, appears to combine the advanges of adding cholesterol to phospholipid bicelles to those of adding charged amphiphiles: it lowers the gel-to-liquid crystal phase transition temperature of the hydrocarbon chains and introduces repulsive interactions that prevent adjacent bicelles from adhering and precipitating. Therefore this bicelle composition allows NMR data for RDC and CSA protein structure constraints to be acquired at or below room temperature, an obvious advantage for solid-state and solution studies of heat-sensitive proteins. Furthermore, cholesterol sulfate is found to extend the alignment-temperature range

without requiring samples with very low bicelle concentrations, which are easily disrupted by introducing solutes. This makes the DMPC/DHPC/cholesterol sulfate particularly promising for membrane protein studies, where concentrated bicelle samples are necessary to achieve sufficient protein concentration. Extensions to relevant experiments and important biological samples will be discussed. Supported by NSF CHE-0847375.

Monitoring Intermediary Metabolism by NMR Isotopomer Analysis (Oral Presentation)

Rui A. Carvalho (✉)
Department of Life Sciences & Center for Neurosciences, Faculty of Sciences and Technology, University of Coimbra, Coimbra 3001-401, Portugal
e-mail: isotopomero@gmail.com

Abstract Several methodologies involving the use of stable isotopes and the detection of their incorporation in metabolic intermediates by NMR will be presented. These include substrate competition analysis for monitoring metabolic remodeling, administration of anaplerotic and oxidative substrates for evaluation of pyruvate cycling and Krebs cycle turnover, and use of deuterated water for evaluating both glucose homeostasis and de novo lipogenesis. Some examples of simultaneous multiple tracer administration will be given. ^{13}C and ^{2}H isotopes are adequate for these analyses since their natural abundance levels, being very low, allow their use in tracer amounts, a crucial feature when applying these methodologies in humans. Some recent results obtained using each methodology in animal models will also be presented and particular emphasis will be given to the advantages/disadvantages of NMR isotopomer analysis in comparison to other methodologies of isotopomer analysis.

Real Time NMR Study of Beta-2-Microglobulin Folding and Characterisation of Intermediate States (Poster and Oral Presentation)

Thomas Cutuil (✉) • **Bernhard Brutscher** • **Isabel Ayala**
Laboratoire de RMN, Institut de Biologie Structurale (CEA, CNRS, UJF), Grenoble, France
e-mail: thomas.cutuil@ibs.fr

Alessandra Corazza
Department of Biomedical Science and Technology, University of Udine, Udine, Italy

Vincent Forge
CEA; DSV; iRTSV, Laboratoire de Chimie et Biologie des M√©taux, Grenoble, France

Abstract HumanÔÄ†beta-2-microglobulin is a small globular 11.7 kDa protein that can form amyloid fibrils in vitro and in vivo that are responsible for amyloidosis disease in patients undergoing dialysis due to renal failure. beta-2-microglobulin is commonly used as a model protein to study protein folding and misfolding, as well as the initial steps of amyloid fibril formation.

It is also widely said that a folding intermediate of beta-2-microglobulin is structurally close to species on the pathway of amyloid fibril formation, and may be involved in the in-vivo formation of these fibrils.

Recently, real-time NMR experiments have revealed a folding mechanism involving two intermediate states, one (I1) being very similar to the final folded state, the other one (I2) giving no detectable signature in the NMR spectra.

Through real-time SOFAST-NMR folding experiments, the evolution of the population of the different species has been followed, giving access to the kinetics and mechanism involved. The exploration of the folding energy landscape, using temperature and concentration dependence, showed that this NMR-invisible I2-intermediate, is an oligomer or a mixture of oligomers, while the I1-intermediate kinetics and thermodynamics are consistent with a previously described cis-trans proline transition.

Atomic Structure of a Nanobody Trapped Key Intermediate of β2m Amyloidogenesis (Poster and Oral Presentation)

Katarzyna Domanska • Vasundara Srinivasan • Saskia Vanderhaegen • Lode Wyns • Vittorio Bellotti • Jan Steyaert (✉)
Department of Molecular and Cellular Interactions, VIB and Department of Structural Biology Brussels, VUB, Pleinlaan 2, B-1050, Brussels, Belgium
e-mail: Jan.Steyaert@vub.ac.be

Abstract Atomic-level structural investigation of the key conformational intermediates of amyloidogenesis remains a challenge. This is due to the nature of the process, which may be described as a dynamic equilibrium between diverse structural species. These intermediates have dissimilar sizes and occur in very uneven amounts and timeframes. Here we demonstrate the utility of heavy chain only antibodies derived from camel (1, 2) for the structural investigation of pre-fibrillar intermediates of β2m amyloidosis. The antigen-binding site of these antibodies consists of a single domain, referred to as VHH or nanobody (2). For these studies, we have selected five single domain antibodies that block the fibrillation of a proteolytic amyloidogenic fragment of β2m (ΔN6β2m). The crystal

structure of ΔN6β2m in complex with one of these nanobodies (Nb24) allows an insight into atomic level of fibrillation mechanism of β2m. The structure of N6β2m variant provides the explanation why this truncated form is less stable – and unlike wild-type protein – has a higher tendency to self-associate and form amyloid fibrils even at physiological pH.

References

1. Hamers-Casterman C et al (1993) Naturally occurring antibodies devoid of light chains. Nature 363(6428):446–448
2. Muyldermans S, Cambillau C, Wyns L (2001) Recognition of antigens by single-domain antibody fragments: the superfluous luxury of paired domains. Trends Biochem Sci 26(4): 230–235

Quest for Foldable Peptides (Poster and Oral Presentation)

Panagiota S. Georgoulia • Nicholas M. Glykos (✉)
Department of Molecular Biology and Genetics, Democritus University of Thrace, University Campus, Dragana, Alexandroupolis 68100, Greece
e-mail: glykos@mbg.duth.gr

Abstract Probing the lower size-limit for peptides with protein-like characteristics has, apart from its biological significance, implications in drug design in the sense that it may allow the design of peptides with a specific three-dimensional structure. The primary aim of this project is to identify putatively foldable tetra- and penta-peptides that can form stable structures in aqueous solutions. A second aim of this project is to systematically evaluate the ability of current generation forcefields to correctly predict the structure and dynamics of such short peptides by comparing the computationally-derived predictions with experimental data that will be obtained using NMR and/or X-ray crystallography. We are in the process of analysing the results obtained from molecular dynamics simulations on a selected set of 1440 tetrapeptides and 7200 pentapeptides. The simulations performed thus far amount to a total of 180 microseconds of simulation time in explicit solvent and with full treatment of the electrostatics. Putatively foldable peptides are identified through the application of a target function which includes terms based on the frame-to-frame RMSD matrix, the atomic fluctuations, the interatomic vector distances etc. Preliminary results obtained from the analysis of the tetrapeptides' simulation set will be presented.

Production and Mechanistic Characterization of Macrolide Antibiotics (Oral Presentation)

Colin J.B. Harvey (✉) • Chaitan Khosla
Department of Chemistry, Stanford University, Stanford, USA
e-mail: cjharvey@stanford.edu

Colin J.B. Harvey • Joseph Puglisi
Department of Structural Biology, and Stanford Magnetic Resonance Laboratory (SMRL), Stanford University, Stanford, USA

Abstract Macrolide antibiotics, typified by Erythromycin, have been a clinically relevant class of compounds for more than 50 years. Medicinal chemistry efforts have lead to the development of several semi-synthetic derivatives with improved properties. Thus far, a majority of these efforts have involved synthetic modification of products isolated from cultures of their native hosts. This approach is severely limited by the highly functionalized nature of these molecules, often making regiospecific modification difficult. Here the production of a series of novel derivatives of Erythromycin produced by precursor-directed biosynthesis is presented. This is an approach by which simple synthetic precursors are fed to engineered biosynthetic systems resulting in the introduction of chemical diversity prior to construction of the complete natural product, effectively removing the issue of regiospecificity.

Presented here are a series of compounds that are not only produced by this novel approach, but contain bioorthogonal functionality for use as handles for further modification as well as possessing bioactivity superior to that of the wild-type compounds.

With these proof-of-principal results in place, a more in-depth study of exactly how these compounds interact with their target, the bacterial ribosome, through the use of cell-free translation and single molecule fluorescent spectroscopy is proposed.

Melatonin Protects Glial Cells From β-Amyloid Neurotoxicity (Poster)

Maksim Ionov (✉) • Barbara Klajnert • Maria Bryszewska
Department of General Biophysics, University of Lodz, Lodz, Poland
e-mail: maksion@biol.uni.lodz.pl

Victoria Burchell • Andrey Abramov
Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK

Introduction. The main component of senile plaques in Alzheimer's disease (AD), aggregated amyloid beta peptide (βA) **(1)**, is neurotoxic and implicated in AD pathology **(2,3)**. Toxicity of the full-length peptides βA 1–40 and βA 1–42 is directly connected to aggregation, since non aggregated peptides are non toxic **(4)**. Melatonin is a hormone secreted from the pineal gland, levels of which are decreased in aging, particularly in AD subjects. This hormone is known to possess neuroprotective properties against βA toxicity in vivo, but the mechanism of protection remains controversial **(5)**. In cultures of mixed neurons and astrocytes, we find that melatonin is protective against neuronal and astrocytic death induced by aggregated full length βA 1–40 and the fragments βA 25–40 and βA 1–28. Melatonin had no effect on the process of fibrillation of βA and did not alter βA-induced calcium signalling in astrocytes. Melatonin significantly reduced the rate of βA-induced reactive oxygen species production, protecting astrocytes against the consequent mitochondrial depolarisation. Thus, scavenging of reactive oxygen species by melatonin appear to be the primary effect in protection of neurons and astrocytes against βA toxicity.

Experimental. Synthetic βA128 [DAEFRHDSGYEVHHQKLVFFAEDVGSNK]; βA25-40 [SNKGAIIGLMVGGVV] and βA140 [DAEFRHDSGYEVHHQKLVF-FAEDVGSNKGAIIGLMV GGVV] were purchased from JPT Peptide Technologies GmbH (Berlin, Germany).

Cell culture: Mixed cultures of hippocampal neurones and glial cells were prepared as described previously **(6)** with modifications, from Sprague-Dawley rat pups 2–4 days post-partum (UCL breeding colony).

Measurements of ψm and ROS: For measurement of ψm, cells were loaded with Rh123 (1 μg/ml; Molecular Probes). For measurement of ROS production dihydroethidium (2 μM) was present in the solution during the experiment.

Toxicity Experiments: For toxicity assays we loaded cells simultaneously with 20 M propidium iodide (PI), and 4.5 M Hoechst 33342 (Molecular Probes, Eugene, OR). Using phase contrast optics, a bright field image allowed identification of cells. A total number of 600–800 neurones or glial cells were counted in 20–25 fields of each coverslip.

Formation of amyloid fibrils-ThT assay: The process of aggregation was monitored using the dye ThT, whose fluorescence is dependent on the formation of amyloid aggregates. The excitation and emission wavelengths were 450 and 490 nm, respectively. Spectrofluorimetric data were analyzed in order to calculate the kinetic constants.

Results. ThT fluorescence is sensitive to the presence of amyloid fibrils, and has therefore been used as an indicator of fibril formation. Amyloid fibrils were formed in vitro and the process was monitored over time. Melatonin (50 μM) not prevent fibril formation for βA 1–40, βA 25–40 and βA 1–28.

Melatonin effectively protects the cells against mitochondrial depolarisation induced by aggregated peptides, suggesting that βA induced mitochondrial depolarisation is ROS dependent. Melatonin successfully reduced the rate of ROS production of all studied forms of βA (n = 49 for βA 1–40; n = 53 for βA 25–40; n = 37 for βA 1–28).

The aggregated forms of the βA were toxic to neurons and astrocytes, significantly increasing the percentage death for cells. Melatonin significantly protected both types of cells against aggregated peptides (Figure 1A-B).

βA-induced excessive ROS production led to profound mitochondrial depolarisation and PTP opening **(6)**. Importantly, the presence of melatonin completely normalised mitochondrial membrane potential, suggesting a critical role for oxidative stress in PTP opening and mitochondrial depolarisation.

Conclusion. In conclusion, we can suggest that melatonin protects neurons and astrocytes by reduction of oxidative stress.

References

1. McNaull BBA, Todd S, McGuinness B, Passmore AP (2010) Mini-Rev Gerontol 56:3–14
2. Hardy J, Selkoe DJ (2002) Progr Probl Road Therap Sci 297:353–356
3. Masters CL, Simms G, Weinman NA, Multhaup G, McDonald BL, K. Beyreuther (1985) Proc Nat Acad Sci USA 82:4245–4249
4. Pike CJ, Burdick D, Walencewicz AJ, Glabe CG, Cotman CW (1993) J. Neurosci 13:1676–1687
5. Pappolla MA, Sos M, Omar RA, Bick RJ, Hickson-Bick DL, Reiter RJ, Efthimiopoulos S, Robakis NK (1997) J. Neurosci 17:1683–1690
6. Abramov AY, Canevari L, Duchen MR (2003) J. Neurosci 23: 5088–5095

Spectrophotmetric Studies of Small Ligand-Polynucleotide Binding: Benzo[B]Thieno[2,3-C]Quinolone and Pentamidine Derivatives (Poster)

Ivana Jarak (✉) • **Marijana Hranjec** • **Ivo Piantanida** • **Grace Karminski-Zamola**
Institution: Department of Organic Chemistry, Faculty of Chemical Engineering and Technology, University of Zagreb, Marulićev trg 20, P. O. Box 177, HR-10000 Zagreb, Croatia
e-mail: jarak.ivana@gmail.com

Abstract The specific and noncovalent interactions of small organic molecules with DNA and RNA have been of great interest for the development of new therapeutics for various diseases, because they provide molecular basis for structure-activity relationship, better understanding of their bioactivity mechanisms as well as rational design of sequence specific polynucleotide-binding molecules. Although many of the important DNA-binding anticancer drugs were discovered in phenotypic, cell-based screens, in vitro experiments have been developed that enable a precise

determination of how a compound interacts with DNA (A large percentage of chemotherapeutic anticancer drugs interact with DNA directly by intercalation or groove binding, or prevent the proper relaxation of DNA through the inhibition of topoisomerases). Searching for compounds with potential antitumor activity we synthesized the series of planar benzo[b]thieno[2,3-c]quinolones as well as a series of groove-binding penatamidine derivatives. Here, we present detailed binding studies of prepared molecules with double-stranded polynucleotides using spectrophotometric methods (UV, fluorescence and CD spectrometry) and structure-activity relationship. Results will be discussed on the poster.

Kinetics of Folding and Aggregation of the Protein Transthyretin and Their Contribution for Understanding Amydoidogenesis (Poster)

Catarina S.H. Jesus • Rui M.M. Brito
Chemistry Department, Faculty of Science and Technology, University of Coimbra, Coimbra, Portugal

Catarina S.H. Jesus • Zaida L. De Almeida • Daniela C. Vaz • Tiago Q. Faria • Rui M.M. Brito (✉)
Center for Neuroscience and Cell Biology, University of Coimbra, Coimbra, Portugal
e-mail: rbrito@ci.uc.pt

Maria João M. Saraiva
Instituto de Ciências Biomédicas de Abel Salazar, and Institute for Molecular and Cellular Biology, University of Porto, Porto, Portugal

Abstract In amyloidosis, normally innocuous soluble proteins polymerize to form insoluble fibrils, which have been associated with a range of pathological states including spongiform encephalopathies, Alzheimer's disease, Parkinson's disease and familial amyloidotic polyneuropathies (FAP), among many others. In certain forms of FAP, the amyloid fibrils are mostly constituted by variants of transthyretin (TTR), a homotetrameric plasma protein implicated in the transport of thyroxine and retinol [1]. According to the model proposed for amyloid fibril formation by TTR, at nearly physiological conditions, the tetramer dissociates to a non-native monomer which, in turn, depending on its conformational stability, undergoes partial unfolding leading to aggregate prone structures that associate and eventually assemble as amyloid fibrils [2]. Studies on folding and aggregation kinetics of TTR are aimed at the identification of the unfolding/ refolding and aggregation mechanisms of TTR variants and also the detection of possible intermediate species implicated in the aggregation process leading to amyloid formation.

In this work, we compare the refolding kinetics of WT-TTR and one of its most common amyloidogenic variants V30M-TTR. Refolding kinetics was followed by

intrinsic tryptophan fluorescence at different urea concentrations, at 25ºC and pH 7. Our results demonstrate that WT- and V30M-TTR refold with a very high field (>95%) to the native tetrameric structure, and that at relatively low protein concentration, a mechanism involving one intermediate is suitable to represent the refolding kinetics of these TTR variants. The results suggest that the amyloidogenic variant refolds at least two orders of magnitude slower than WT-TTR. Thus, despite the small differences between the crystal structures of both proteins, the refolding process from unfolded monomers to the corresponding native tetramer is kinetically much more favorable for WT- than for V30M-TTR. Since the V30M-TTR takes longer to refold to the native tetramer state, the formation of transient non-native monomeric species with high tendency for aggregation may be favored.

TTR aggregation was induced in vitro by the addition of NaCl to the acid unfolded form at pH 2 and 25ºC, which generates a molten globule that aggregates into small soluble oligomers that eventually assemble into amyloid fibrils. The kinetics of the unfolded → molten globule → soluble oligomers pathway was monitored by far- and near-UV CD and fluorescence measurements. The results obtained with WT-TTR show that upon salt addition the unfolded protein changes its conformation to the molten globule state. Soluble aggregates are then accumulated and only minor amounts of the monomeric forms are detected hours after salt addition. Additionally, NMR studies with WT-TTR samples labelled with ^{13}C-Ala show that a small fraction of the 12 Ala residues per TTR monomer remain in unstructured and solvent accessible regions after amyloid protofilament formation in agreement with the structural model previously proposed [3].

References

1. Brito RMM, Damas AM, Saraiva MJ (2003) Curr Med Chem Immun Endoc Metab Agents 3:349–360
2. Quintas A, Vaz DC, Cardoso I, Saraiva MJM, Brito RMM (2001) J Biol Chem 276:27207–27213
3. Correia BE, Loureiro-Ferreira N, Rodrigues JR, Brito RMM (2006) Protein Sci 15: 28–32

Dynamics of a Skeletal Troponin C – Troponin I Chimera Probed by Comparison of Experimental and Simulated NMR Relaxation Parameters (Poster)

Olivier Julien • Pascal Mercier • Brian D. Sykes (✉)
Department of Biochemistry, University of Alberta, Edmonton, AB, Canada
e-mail: brian.sykes@ualberta.ca

Claire Allen • Tharin M.A. Blumenschein
School of Chemistry, University of East Anglia, Norwich, UK

Olivier Fisette • Patrick Lagüe
Département de biochimie et de microbiologie, Université Laval, Québec, QC, Canada

Carlos H.I. Ramos
Instituto de Química, Universidade Estadual de Campinas, Campinas, SP, Brazil

Abstract The activation of skeletal and cardiac muscle is triggered by the release of calcium from the sarcoplasmic reticulum. The calcium sensor is the troponin complex that is formed by three subunits: the calcium-binding protein troponin C (TnC), the inhibitory protein troponin I (TnI) and the tropomyosin-associated protein troponin T (TnT). When calcium binds to TnC, the resulting conformational change allows TnC to bind TnI, leading to the removal of the C-terminal region of TnI from actin. Consequential movement of the tropomyosin allows the binding of the myosin head to actin resulting in a power stroke. Regions of these proteins are highly flexible and the importance of these intrinsically disordered sections has been recently recognized and rationalized (Hoffman et al. *J. Mol. Biol.* 2006 361:625–633).

Structural studies of the muscle system have been very successful in determining the structural organization of most of the molecular components involved in force generation at the atomic level. Although mainly α-helical, the structure and dynamics of TnI remains controversial, particularly in its C-terminal region. Different structures have been presented for this region: a single α-helix observed by x-ray crystallography, a "mobile domain" containing a small β-sheet derived from NMR restraints, and a mainly unstructured region according to NMR relaxation data. To investigate this, we have constructed a skeletal TnC-TnI chimera that contains the N-domain of TnC (1–91), a short linker (GGAGG), and the C-terminal region of TnI (98–182). Our objective is to determine which of the proposed structures best fit the experimental ^{15}N relaxation data for this chimera. The comparison between experimental and NMR relaxation parameters calculated from molecular dynamic simulations will be presented to assess the validity of the models.

Structure and Dynamics of Delta Subunit of RNA Polymerase from *Bacillus subtilis* (Poster and Oral Presentation)

Pavel Kadeřávek (✉) • Veronika Motáčková • Petr Padrta • Lukáš Žídek • Vladimír Sklenář
National Centre for Biomolecular Research, Masaryk University, Brno, Czech Republic
e-mail: kada@ncbr.chemi.muni.cz

Carl Diehl • Mikael Akke
Center for Molecular Protein Science, Biophysical Chemistry, Lund University, Lund, Sweden

Hana Šanderová • Libor Krasný
Institute of Microbiology and Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic

Abstract RNA polymerase is responsible for the DNA transcription in a cell. The RNA polymerase of Gram positive bacteria consists of seven subunits. Current project focuses on the delta subunit which increases the transcriptional specificity and efficiency of the RNA synthesis [1, 2].

A delta subunit of RNA polymerase is a two domain protein. Only the N-terminal part has a well defined structure, which has been recently solved in our laboratory using multi-dimensional NMR spcectroscopy [3]. The motions of the protein backbone within the structured part were investigated using ^{1}H-^{15}N spectroscopy in order to reveal the functionally relevant parts of the molecule.

The study of protein dynamics was based on the analysis of relaxation rates of the backbone ^{1}H-^{15}N spin pairs. To investigate motions on the nano-to-picosecond timescale the standard set of relaxation rates (R_1, R_2, NOE) was measured. The experiments were performed at two magnetic fields (500 MHz, 600 MHz) and the acquired data were interpreted using the Model-Free approach [4, 5]. In addition, motions on the micro-to-milisecond timescale were studied using relaxation dispersion experiments [6, 7].

The results show a correlation between the residues undergoing slow exchange and the conserved residues predicted to form an interaction surface with other subunits of the molecular complex. On the contrary, the most flexible residues on the pico-to-nanosecond timescale were located in a non-interacting part of the protein.

References

1. Dobinson KF, Spiegelman GB (1987) Biochemistry 26:8206–8213
2. Juang YL, Helmann JD (1994) J Mol Biol 239:1–14
3. Motáčková V, Šanderová H, Žídek L, Nováček J, Padrta P, Švenková A, Korelusová J, Jonák J, Krásný L, Sklenář V (2010) Proteins 78:1807–1810
4. Lipari G, Szabo A (1982) J Am Chem Soc 104:4546–4559
5. Lipari G, Szabo A (1982) J Am Chem Soc 104:4559–4570
6. Palmer AG, Kroenke CD, Loria JP (2001) Meth Enzymol 339:204–238
7. Long D, Liu ML, Yang DW (2008) J Am Chem Soc 130:2432–2433

A Comparative Analysis of the Equilibrium Dynamics of a Designed Protein Inferred from NMR, X-Ray, and Computations (Poster)

Lin Liu • Ivet Bahar (✉)
Department of Computational Biology, University of Pittsburgh, Pittsburgh, USA
e-mail: bahar@ccbb.pitt.edu

Lin Liu • Leonardus M.I. Koharudin • Angela M. Gronenborn (✉)
Department of Structural Biology, School of Medicine, University of Pittsburgh, Biomedical Science Tower 3, Pittsburgh, PA 15213, USA
e-mail: amg100@pitt.edu

Abstract A detailed analysis of high-resolution structural data and computationally predicted dynamics was carried out for a designed sugar-binding protein. The mean-square deviations in the positions of residues derived from nuclear magnetic resonance (NMR) models and those inferred from X-ray crystallographic B-factors

for two different crystal forms were compared with the predictions based on the Gaussian Network Model (GNM) and the results from molecular dynamics (MD) simulations. GNM systematically yielded a higher correlation than MD, with experimental data, suggesting that the lack of atomistic details in the coarse-grained GNM is more than compensated for by the mathematically exact evaluation of fluctuations using the native contacts topology. Evidence is provided that particular loop motions are curtailed by intermolecular contacts in the crystal environment causing a discrepancy between theory and experiments. Interestingly, the information conveyed by X-ray crystallography becomes more consistent with NMR models and computational predictions when ensembles of X-ray models are considered. Less precise (broadly distributed) ensembles indeed appear to describe the accessible conformational space under native state conditions better than B-factors. Our results highlight the importance of using multiple conformations obtained by alternative experimental methods, and analyzing results from both coarse-grained models and atomic simulations, for accurate assessment of motions accessible to proteins under native state conditions.

Investigating Domain-Swapped Proteins by NMR (Oral Presentation)

Lin Liu • Ivet Bahar (✉)
Department of Computational Biology, University of Pittsburgh, Pittsburgh, USA
e-mail: bahar@ccbb.pitt.edu

Lin Liu • In-Ja L. Byeon • Angela M. Gronenborn (✉)
Department of Structural Biology, School of Medicine, University of Pittsburgh, Biomedical Science Tower 3, Pittsburgh, PA 15213, USA
e-mail: amg100@pitt.edu

Abstract For most proteins under physiological conditions, the native functional state is a single, stable structure. However, certain circumstances may push protein folding into distinctly different structures. The most common alternative structures comprise different multimeric assemblies of identical polypeptide chains, since multimers are endowed with structural and functional advantages. Among thousands of homo-oligomeric protein structures, there is a small, but growing subset 'domain-swapped proteins' [1], in which either the exchanged subunit or the remain portion in the oligomer is identical to the one in the corresponding monomer, except for the region that links the exchanging domains. Although no unifying molecular mechanism of domain swapping has been figured out, it appears that domain swapping is closely associated with the unfolding/folding process of proteins. For some proteins, distinct intermediates may exist, while for others,

complete un/folding may occur. Studying the dynamics of specific domain-swapped systems by NMR will allow us to gain a better understanding of the mechanism of domain swapping.

Reference

1. Bennett MJ, Choe S, Eisenberg D (1994) Refined structure of dimeric diphtheria toxin at 2.0 Å resolution. Protein Sci 3:1444–1463

Structural Studies of the Sigma Factor Regulator HasS and Its Interaction with the Signaling Domain of the Specific Heme Transporter HasR (Poster)

Idir Malki • Gisele C. Amorim • Ada Prochnicka-Chalufour • Muriel Delepierre (✉) • Nadia Izadi-Pruneyre

Département de Biologie Structurale et de Chimie, Institut Pasteur, Unité de Résonance Magnétique Nucléaire des Biomolécules, CNRS URA 2185, 75724, Paris Cedex 15, France
e-mail: muriel.delepierre@pasteur.fr

Cécile Wandersman

Département de Microbiologie, Institut Pasteur, Unité des Membranes Bactériennes, CNRS URA 2172, 75724, Paris Cedex 15, France

Abstract To satisfy their need for iron, several Gram-negative bacteria use a heme uptake system, Has (Heme acquisition system). This system involves an extracellular protein, called hemophore HasA. The function of HasA is to acquire free or hemoprotein-bound heme and to deliver it to HasR, a specific outer membrane receptor. HasR belongs to the TonB-dependent outer membrane receptors class. To internalize their substrate, all the receptors of this family need the energy driven by an inner membrane complex composed of TonB-ExbB-ExbD. Like some other TonB-dependent receptors, HasR has a signaling activity. Indeed, it possesses a periplasmic domain (about 100 residues) involved in a signaling cascade that regulates expression of genes required for heme transport. This signaling domain of HasR interacts with a sigma factor regulator HasS that controls the activity of a sigma factor protein HasI in the cytoplasm. HasS is an inner membrane protein with a large periplasmic. It has been shown that the signal activating the transcription regulation via HasS is the concomitant presence of heme and HasA on the HasR receptor.

The aim of this project is to study the structure of the periplamic domain of HasS and the signaling domain of HasR and to identify the nature of the signal transmitted from the extracellular part of HasR to HasS. Various periplasmic fragments of HasS were built and expression tests are in progress. Furthermore, we overexpressed and purified the signaling domain of HasR, which is not seen in the crystal structure of this receptor. NMR experiments show that the protein is folded and stable during several weeks. Thus its structure determination by NMR should be possible.

Cell Free Expression Systems Optimized for Disulfide-Rich Proteins (Poster)

Wael Mohamed (✉) • **Veronica Tamu Dufe** • **Khadija Wahni** • **Joris Messens**
VIB Department of Molecular and Cellular Interaction, Vrije Universiteit Brussel, Brussels, Belgium

Brussels Center for Redox Biology, Vrije Universiteit Brussel, Brussels, Belgium

Structural Biology Brussels, Vrije Universiteit Brussel, Brussels, Belgium
e-mail: Wael.mohamed@vub.ac.be

Hirotaka Takahashi • **Yaeta Endo**
Cell-Free Science and Technology Research Center, Ehime University, Matsuyama, Japan

Abstract Many disulfide-rich proteins immensely relevant for immunology and oncology await structure determination and biochemical characterization due to lacking of correctly folded and functional protein. The open system of cell free protein expression using wheat germ extract or *E. coli* is ideal for the expression of these proteins and also for the addition of the correct mix of thiol-disulfide oxidoreductases such as the bacterial disulfide bond (Dsb) proteins, the eukaryotic protein disulfide isomerase (PDI) and the eukaryotic quiescin sulfhydryl oxidase (QSOX), which are important cellular enzymes for the correct oxidative folding of polypeptides substrates by oxidation, reduction and/or isomerisation. Using *E. coli* expression system, disulfide rich proteins like i.e. Found In Inflammatory Zone (Fizz1), Vascular Endothelial Growth Factor (VEGF), Placental Growth Factor (PLGF), Scorpion toxin and the full length inner membrane protein *E. coli* Disulfide bond D (*E. coli* DsbD) are always found in inclusion bodies or their expression levels are extremely low. By cell-free expression system using wheat germ extract and *E. coli* extract, we were able to express and purify these disulfide rich proteins in a soluble form. The addition of Human QSOX to the open system increased the level of expression of mouse VEGF, which suggests its important role as a chaperon that may help protein folding. The activity of the purified proteins as well as biochemical characterization are still in progress. In conclusion, cell-free protein expression system is a promising system for the production of difficult to express disulfide rich proteins

Structural Aspects of Peptide-Receptor Recognition (Poster)

I. Ozerov • **T.V. Pyrkov** • **R.G. Efremov** (✉)
Biological Faculty, Department of Bioengineering, M.V. Lomonosov Moscow State University, Leninskie Gori, 1/73 119991 GSP-1, Moscow, Russia
e-mail: r-efremov@yandex.ru

Abstract There are many low-molecular peptides with known physiological effects. Besides that, more and more novel artificial peptides are discovered. Their biological activity is investigated for the purpose of making new drugs.

For better understanding of the peptide's role in nature it is necessary to inquire into interactions of such ligands with their receptors and to determine intermolecular contacts specific for the certain functional groups. Such study would be useful for the design of new peptides with predefined pharmacological and kinetic properties.

During this research the interactions of peptide ligands with their receptors have been analysed. 3D structures of ligand-receptor complex were taken from PDBBind database. The database consists of more than 3000 receptor-ligand complexes and includes more than 150 structures with peptides posessing two or more peptide bonds. A number of modern quantitative approaches such as the concept of MHP (Molecular Hydrophobicity Potential) for the numerical estimation of nonpolar interactions were used. Also the other types of intermolecular contacts were analyzed including hydrogen bonds and stacking. The aim of this work was to determine structural features typical for the peptide-binding receptors responsible for specific recognition of peptides in the active site. The results of this research will be used for the efficient design of novel peptides with modified biological activity.

Solution Structure of a Highly Potent 2′ Modified siRNA Duplex (Poster)

Peter Podbevsek (✉) • **Janez Plavec**

Slovenian NMR Center, National Institute of Chemistry, Hajdrihova 19, SI-1001 Ljubljana, Slovenia

e-mail: peter.podbevsek@ki.si

Charles R. Allerson • Balkrishen Bhat

Department of Medicinal Chemistry, Isis Phamaceuticals, Inc., 1896 Rutherford Road, Carlsbad, CA 92008, USA

Janez Plavec

Faculty of Chemistry and Chemical Technology, University of Ljubljana, SI-1000 Ljubljana, Slovenia

EN-FIST Center of Excellence, Dunajska 156, SI-1000 Ljubljana, Slovenia

Balkrishen Bhat

Regulus Therapeutics, 1896 Rutherford Road, Carlsbad, CA 92008, USA

Abstract RNA interference (RNAi) is triggered by short RNA duplexes, which can be used for the silencing of virtually any gene. Naturally occurring siRNAs are produced by enzymatic cleavage of a longer double stranded RNA molecule into shorter RNA 21-nt duplexes with 2-nt 3′ overhangs and 5′ phosphates. Once in cells, siRNA associates with several proteins forming an RNA-induced silencing

complex (RISC). The RISC possesses nuclease activity and cleaves the target mRNA. This reduces the level of target mRNA in cells and effectively knocks down a specific gene. Unfortunately, synthetic siRNAs consisting of solely standard nucleotides exhibit short half-life in serum due to the activity of endo- and exonucleases. A fully modified duplex, which is comprised of alternating 2′-F and 2′-OMe nucleotides exhibits several desirable pharmacokinetic properties like higher thermal and plasma stability (1). Consequently, 2′-F/2′-OMe siRNA showed a more than 500-fold increase of in vitro potency versus unmodified siRNA. To gain some insight into the structural features, which result in the high potency of the modified siRNA, we determined the 3D structure of the fully modified siRNA duplex in solution.

The two 21-nt oligonucleotides (1-GGGUAAAUACAUUCUUCAUUU-21 and 22-AUGAAGAAUGUAUUUACCCUU-42) efficiently hybridize thus forming an A-type double helix with 3′ UU overhangs on both strands. The helical segment is completely complementary and exhibits 19 Watson-Crick base pairs. However, NMR data suggests that the stability of individual base pairs is not uniform through the whole length of the construct. The last three base pairs display somewhat different properties. Their imino protons are accessible to solvent exchange, which is an indication of base pair opening. Stabilization of these base pairs is achieved through favorable stacking interactions. Furthermore, U20-U21 overhang in the sense strand stacks efficiently on the terminal U19·A22 base pair. On the other hand, there is poor, if any, base-base stacking of the antisense strand U41 and U42 overhang on the terminal GC base pair. As a result, the structure of the U41-U42 overhang is poorly defined. However, this appears to have no effect on the terminal GC base pair. Differences in stability of helical segments can be attributed to the nucleotide sequence of the construct, which ends with two AU base pairs on one end opposed to three GC base pairs on the opposite end.

Our findings are in agreement with previous reports, which suggest that a difference in stability of duplex ends is required for the incorporation of the correct siRNA strand into RISC. The strand whose 5′ end is at the less stable end of the helix becomes the guide strand, while the other strand becomes the passenger strand. Our data show a clear difference in the relative stability of helix ends. The labile base pairs A18·U23 and U19·A22 suggest that A22-U42 strand will serve as a guide strands and will thus control the incorporation of the siRNA duplex into the RISC complex.

Reference

1. Allerson CR, Sioufi N, Jarres R, Prakash TP, Naik N, Berdeja A, Wanders L, Griffey RH, Swayze EE, Bhat B (2005) J Med Chem 48:901–904

Structural Studies of the Light Driven Enzyme NADPH: Protochlorophyllide Oxidoreductase (Oral Presentation)

Andrew Proudfoot (✉) • C Jeremy Craven • C Neil Hunter • Mike P. Williamson
Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, UK
e-mail: a.proudfoot@sheffield.ac.uk

Abstract The light driven enzyme Protochlorophyllide Oxidoreductase (POR) is responsible for catalysing the reduction of the C_{17}–C_{18} double bond of the D ring of protochlorophyllide (Pchlide), in the presence of NADPH, forming chlorophyllide (Chlide)[1]. The reduction of Pchlide involves a light-induced hydride transfer reaction from the pro-S face of nicotinamide adenine dinucleotide phosphate (NADPH), to the C_{17} position coupled to the addition of a proton to the C_{18} position forming Chlide[2]. The reaction catalysed by POR is a key step in chlorophyll biosynthesis and is essential in the development of chloroplasts[3].

Due to the large size of POR (37 kDa) work has been conducted to optimise the conditions used to conduct the NMR, allowing all of the 322 signals to be resolved with sufficient intensity to assign them. The enzyme is from a thermophilic organism, permitting spectra to be obtained at 50ºC. The improvement in linewidth with temperature is dramatic, indicating aggregation at lower temperature. Spectra from perdeuterated protein do not show the expected improvement in linewidth, probably because of impurities copurifying with the protein. Finally, chemical exchange broadening means that ^{15}N R_2 rates are ca. 40% faster at 800 MHz than they are at 600 MHz. The combined effect of these factors means that the best spectra are obtained at 600 MHz using non-deuterated protein.

References

1. Heyes DJ, Hunter CN (2005) Trends Biochem Sci 30:642–649
2. Townley HE et al (2001) Proteins-Struct Funct Genet 44:329–335
3. Griffiths WT et al (1996) Febs Lett 398:235–238

Inferential Structure Determination: Towards Faster Calculations (Poster)

Yannick G. Spill (✉) • Mansiaux Yohann • Nilges Michael
Department of Structural Bioinformatics, Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France
e-mail: yannick.spill@pasteur.fr

Extracting the essential constraints that allow protein structures to be determined. Protein Structure determination methods often use NOESY-derived distance restraints. This data can be both inconsistent and of varying quality depending on the position in the sequence of the involved residues. Here we use two tools, FIRST[1] and QUEEN[2], to narrow down a set of constraints to it's essential constituents. Using the ISD software package[3], we were able to show that convergence is maintained even after deleting 11 out of 12 restraints for some simulations, without substantial degradation of the main structural features. Suppressing these restraints also lowered the variance of the positions of the atoms and shows that these tools could help to make datasets more consistent.

Improvements of ISD's replica-exchange sampling scheme. We implemented an optimization scheme that automatically adjusts the parameters to achieve even acceptance ratii.

References

1. Chubynsky MV, Thorpe MF (2007) Algorithms for three-dimensional rigidity analysis and a first-order percolation transition. Phys Rev E, Am Phys Soc 76: 041135
3. Nabuurs SB, Spronk CAEM, Krieger E, Maassen H, Vriend G, Vuister GW (2003) Quantitative Evaluation of Experimental NMR Restraints. J Am Chem Soc, American Chemical Society, 125:12026–12034
3. Rieping W, Habeck M, Nilges M (2005) Inferential structure determination. Science 309: 303–306

Tracking Fast Structural Dynamics in Hemoglobin Using Time-Resolved Wide-Angle X-Ray Scattering (Poster)

A. Spilotros (✉) • M. Cammarata • M. Levantino • G. Schirò • M. Wulff • A. Cupane
Dipartimento di Scienze Fisiche ed Astronomiche, Università degli Studi di Palermo, via Archirafi 36, I-90123 Palermo, Italy
e-mail: alessandro.spilotros@fisica.unipa.it

Abstract It has been shown that Human Hemoglobin (Hb) can adopt several structures and they have different affinity for ligands (oxygen, carbon monoxide). For this reason it is necessary to have an accurate description of the structural dynamics, i.e. the time scale of the structural change [1]. We have used the method of time-resolved wide angle X-ray scattering (TR-WAXS) developed on beamline ID09B at the European Synchrotron and Radiation Facility (ESRF) in Grenoble to study the conformational transitions of hemoglobin in solution [2].

We investigated the tertiary and quaternary conformational changes of human hemoglobin under nearly physiological conditions triggered by laser-induced ligand photolysis. Although many time resolved experiments using different probes have

been done [3 and references therein], x-ray scattering has an enormous advantage with respect to optical probes being a direct structural probe.

We measured the Hb structural change following photolysis in the wide time range from few nanosecond to several tens of milliseconds: the change in shape of the signal immediately revealed the relevant time scales of tertiary and quaternary conformational transitions.

References

1. Eaton WA et al (2007) IUBMB Life 59:586–599
2. Cammarata M et al (2008) Nat Method 5:881–886
3. Balakrishnan G et al (2004) J Mol Biol 340:843–856

Molecular Modeling of Bisbenzylisoquinoline Alkaloids-DNA Complexes (Poster and Oral Presentation)

Ioana Stanculescu • Cezar Bendic (✉)
Department of Physical Chemistry, Faculty of Chemistry, University of Bucharest, 4-12 Regina Elisabeta Bd., 030018 Bucharest, Romania
e-mail: cbendic@gw-chimie.math.unibuc.ro

Abstract The interactions between double stranded deoxyribonucleic acid (**DNA**) and bisbenzylisoquinoline alkaloids, tetrandrine (**TET**) and its derivate N,N-dibenzyl tetrandrine (**diB-TET**), potential anticancer drugs, were studied to elucidate the binding mechanism: intercalation or minor/major groove binding. The objective of the work was to evaluate the contribution of the different structural factors that determine the main binding mechanism using molecular modeling. The models of the drug–nucleic acid complexes were built by manual docking followed by molecular mechanics optimization with implicit solvent effect using OPLS force field. In order to identify and analyze intermolecular interactions for the drug–**DNA** complexes, the SHB_interactions program, based on Mulliken overlap populations (**OP**s) as a quantitative quantum chemical criterion of the atom-atom intermolecular interaction, was used [1]. The source code for SHB_interactions program, written in C, instructions and some examples are available at: http://gwchimie.math.unibuc.ro/staff/cbendic/shb/SHB_interactions.html.

The most favorable interaction is obtained by intercalation and minor groove binding for **TET** and by major groove binding for **diB-TET**. As expected the contribution of the electrostatic term becomes more important for the dicationic derivative. The analysis of Mulliken overlap populations shows that **OP** due to H-bonds represents a small part of all atom-atom interactions. Although the H-Bonds contribution is small, they probably assure the binding selectivity of the compounds and their biological activity [1]. The C-H . . . O (N) bonds, weaker than

the classic hydrogen bonds, appears more frequently. Intermolecular interaction distances are characteristic to H bond and van der Waals interactions.

Reference

1. Cezar Bendic (2008) Overlap Population – a Quantum Chemical Criterion for Atom-Atom Intermolecular Interaction. In: Advances in quantum chemical bonding structures (ed). Transworld Research Network, 251–270, 2008

Relating Structure to Function for Natural Surfactant Proteins (Poster and Oral Presentation)

Steven Vance (✉) • Cameron Mckenzie • Alan Cooper
WestChem Department of Chemistry, Joseph Black Building, University of Glasgow, Glasgow G12 8QQ, Scotland, UK
e-mail: stevva@chem.gla.ac.uk

Brian Smith
Faculty of Biomedical and Life Sciences, Biochemistry & Cell Biology, Joseph Black Building, University of Glasgow, Glasgow G12 8QQ, Scotland, UK

Malcolm Kennedy
Faculty of Biomedical and Life Sciences, Ecology & Evolutionary Biology, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, Scotland, UK

Abstract Natural surfactants and biofoams are known to play important roles within nature. The source of this activity can, in the majority of examples, be attributed to the presence of small molecule surfactants such as lipids. Recently however, a new group of surfactant proteins have been identified. Two surfactant proteins; Ranaspumin-2 (Rsn2) from the foam nests of the Tungara frog and Latherin from horse sweat have been investigated. Each has been shown to possess unusual characteristics specifically evolved to fulfil their unique function. A solution structure has been determined for Rsn2 but does not explain the source of its activity. Therefore, it is proposed that the molecule must undergo a major conformational rearrangement on interaction with the air-water interface. This alternative conformation is being investigated. NMR structure determination for Latherin is ongoing.

Grafting of Extracellular Loops of Y Receptors onto a Beta-Barrel Scaffold (Poster)

Reto Walser (✉) • **Oliver Zerbe**
Organic Chemistry Institute, University of Zurich, Winterthurerstr. 190, Zurich ZH 8057, Switzerland
e-mail: retowalser@yahoo.com

Joerg Kleinschmidt
Department of Biology, University of Konstanz, D-78457 Constance, Germany

Abstract G-protein coupled receptors (GPCRs) belong to the class of seven-transmembrane proteins and play a pivotal role in signal transduction processes. The GPCRs share a common structural motif of an extracellular N-terminal segment, linked to a bundle of seven membrane-spanning helices that are connected through three loops on the intra- and extracellular sides, followed by an intracellular C-terminal segment [1,2].

Unfortunately, wild-type GPCRs are very difficult to produce recombinantely. In this work we aim at developing a small, preferably soluble protein, which possesses a stable core determining the global fold, and comprises surface-exposed loops that can be modified without compromising folding. In such a system the extracellular portions of a GPCR could be transferred onto that core resulting in a chimeric protein that could be considered as a "minireceptor". Such a protein contains all the parts of the receptor hypothesized to be important for ligand binding (hence the term "receptor"), but at the same time displays the favourable characteristics of small, soluble proteins (hence the term "mini"). Our efforts are concentrated on the construction and characterization of such a minireceptor for the so-called Y-receptor [3] family of GPCRs, that is targeted by peptides of the neuropeptide Y (NPY) family. The scaffold is derived from b-barrel proteins, and NMR techniques are used to assess to which extent the global fold is retained in the loop-modified mutants.

References

1. Palczewski (2000) Science 289, 739
2. Rasmussen (2007) Nature 450, 383
3. Grundemar (1997) Neuropeptide Y and drug development. Academic Press, San Diego

Index

J.D. Puglisi and M.V. Margaris (eds.), *Biophysics and Structure to Counter Threats and Challenges*, NATO Science for Peace and Security Series B: Physics and Biophysics, DOI 10.1007/978-94-007-4923-8,