Methods in Molecular Biology 2253

Springer Protocols

Luisa Di Paola
Alessandro Giuliani *Editors*

Allostery

Methods and Protocols

MOREMEDIA

Humana Press

METHODS IN MOLECULAR BIOLOGY

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

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

For over 35 years, biological scientists have come to rely on the research protocols and methodologies in the critically acclaimed *Methods in Molecular Biology* series. The series was the first to introduce the step-by-step protocols approach that has become the standard in all biomedical protocol publishing. Each protocol is provided in readily-reproducible step-by-step fashion, opening with an introductory overview, a list of the materials and reagents needed to complete the experiment, and followed by a detailed procedure that is supported with a helpful notes section offering tips and tricks of the trade as well as troubleshooting advice. These hallmark features were introduced by series editor Dr. John Walker and constitute the key ingredient in each and every volume of the *Methods in Molecular Biology* series. Tested and trusted, comprehensive and reliable, all protocols from the series are indexed in PubMed.

Allostery

Methods and Protocols

Edited by

Luisa Di Paola

Unit of Chemical-Physics Fundamentals in Chemical Engineering, Department of Engineering, Universitá Campus Bio-Medico di Rome, Rome, Italy

Alessandro Giuliani

Environment and Health Department, Instituto Superiore di Sanità, Rome, Italy

Humana Press

Editors
Luisa Di Paola
Unit of Chemical-Physics Fundamentals
in Chemical Engineering
Department of Engineering
Universitá Campus
Bio-Medico di Rome
Rome, Italy

Alessandro Giuliani
Environment and Health Department
Instituto Superiore di Sanità
Rome, Italy

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-1153-1 ISBN 978-1-0716-1154-8 (eBook)
https://doi.org/10.1007/978-1-0716-1154-8

Preface

Proteins are at the interface between simple and complex systems as aptly synthesized some years ago in an enlightening paper by Hans Frauenfelder and Peter Wolynes (Frauenfelder, Hans, and Peter G. Wolynes. "Biomolecules: where the physics of complexity and simplicity meet." *Physics Today;(United States)* 47.2 (1994)). This peculiar position makes the study of protein structure and dynamics as the most convenient vantage points were to look at all those features like self-organization, signal/noise discrimination, specificity of interaction, multiple equilibria that are not present in organic molecules (whose behavior can be satisfactorily faced by standard chemo-physical approaches) and that are too difficult to grasp and analyze in biological systems (whose behavior emerges from a nondecomposable mixture of top-down and bottom-up regulations). It is not without meaning that one of the most influential journals in protein science is *Biophysical Chemistry* that in the title embeds the three most central fields of natural science.

The investigation in protein science stems from very reliable data coming from the "simplicity" end (atomic coordinates of protein structures from X-ray crystallography, chemo-physical properties of amino acid residues) and goes into the "complexity" territories of the discrimination of relevant "signals" from thermal noise in the case of allostery or the conundrum of the structure–function relation in natively unfolded proteins.

This creates a perfect playground to explore the "mesoscopic realm" (Laughlin, R. B., Pines, D., Schmalian, J., Stojković, B. P., & Wolynes, P. (2000). The middle way. *Proceedings of the National Academy of Sciences*, *97*(1), 32-37.) where the still largely unknown organization principles ordering the middle-scale between atoms and galaxies are hidden.

This very ambitious goal is (more or less) latent in all book chapters that rebound around the two basic issues of "allostery" and "network" that are present in almost all the chapter titles. These two issues are each other connected by the fact that proteins are the most basic "machine-like" objects sharing with human-made machine the need to organize their architecture to a purpose that pertains to a different organization layer. It is convenient to talk of a "purpose" more than a "function" (that is present even in simpler systems) for the same reason that differentiates an isolated piston (whose function is to transform the energy coming from oil explosion into a rhythmic motion) from the entire car that needs to integrate the different functions of its parts into a coherent whole. At odds with a car, a protein accomplishes its goal (e.g., to transport oxygen from lungs to tissues along blood flow) without a driver, thus it must take care of self-organizing according to its microenvironment (e.g., lowering its affinity constant for oxygen in the peripheral tissues and increasing the affinity in the lungs). This implies the need of a sensor–control–effector circuit like any self-adjusting device; these three tasks correspond to concerted changes of the whole configuration that start from a sensor and end up at the effector that is exactly what "allostery" is for: sensing a relevant stimulus, transporting the information across the entire structure, and changing the configuration accordingly.

In order to do so, the protein must have a wiring architecture that allows to both discriminate relevant signals by thermal noise (in a situation where signal-to-noise ratio is near unity in energetic terms) and make the signal to reach the correct effector (e.g., the active site). This wiring architecture can profitably be interpreted in terms of a "network" whose nodes are amino acid residues and links the effective contacts between them, and the

investigation of the features of the network architecture can shed light on the dynamics and efficacy of allosteric control and open the way to a completely new avenue of therapeutic intervention. The general paradigm of proteins as "self-organizing" machines is the "red line," unifying all the different chapters of the book.

In Chapter 1, the authors put into an historical perspective the problem of allostery and demonstrate how this issue constitutes the basic issue to give a scientific foundation to biology freeing the life science to invoke "Maxwell demons" to get rid of the otherwise impossible to understand extreme specificity of cell metabolism.

Chapter 2 has a methodological/computational flavor focusing on the necessary link between allostery and network formalism.

Guang Hu in Chapter 3 presents a rigorous way to "dynamize" structural network by considering their links as springs, and this modeling choice allows to simulate the allosteric behavior of the protein molecules.

Chapter 4 goes in depth into the energetic features of allostery. The authors are able to establish a link between the presence of oscillating modes traversing the structure and the underlying network wiring, thus giving a physically motivated picture of allosteric process.

Adnan Sljoka, in Chapter 5, proposes a mechanical perspective for the elucidation of the "second secret of life" (a suggestive but very well-motivated definition of allostery) in terms of rigidity perturbation.

Chapter 6 (apparently) deals with another theme that is the nature of protein–protein interaction, and this issue asks for the consideration of cooperative effects that are at the very basis of any kind of signal transduction across protein structures.

The convergence between protein–protein interaction and allostery is clarified in Chapter 7. Both allostery and protein–protein interaction rely on the shared need of contact network rewiring that is at the very basis of any motion of the "protein machines."

Chapter 8 is a reprise of the methodological line of reasoning of Chapter 2, going in depth into the presence of "sub-networks" (domains) of global contact network and introducing the crucial issue of "assortativity" that generates that breaks the symmetry of contacts distribution creating "specialized spatial patches" in protein territory.

The "symmetry breaking" introduced by Lesieur and Vuillon in Chapter 8 is further analyzed in Chapter 9, describing a method that combines the information on the correlated protein motions resulting from atomistic MD simulations with a network analysis based on graph partitioning into mutually exclusive groups, named communities.

Chapter 10 introduces a software suite designed to analyze molecular dynamics and structural ensembles in a network perspective allowing to conjugate the three main dimensions of protein science: dynamical, structural, and chemical, allowing the different classes of intra- and intermolecular interactions to be represented, combined, or alone in the form of interaction graphs starting from molecular dynamics trajectories.

Chapter 11 faces the most paradigmatic case of interaction specificity in biomedicine: the antigen–antibody recognition in terms of allostery allowing the reader to grasp the fundamental role of this phenomenon in life sciences.

The "action at distance" by the transduction of signal across an organized network structure is both the theme of Chapter 12 and the fundamental "recipe" of living entities at the molecular level.

The central position of allosteric-like signal transduction is the focus of Chapter 13, dealing with the probably most famous (and studied) hub protein. P53 located at the crossroad of cell cycle regulation, genome integrity, and cancer development. Elena Papaleo in

this chapter asks for the need for a fresh approach to the study of this protein, encompassing a more integrated and systemic attitude.

Chapter 14 introduces the most direct applicative dimension of the network-based approaches to allostery, introducing a totally new approach to pharmacology: allosteric drug paradigm that promises to make pharmacological research to exit from a bottleneck lasting since at least two decades.

Chapter 15 constitutes a (largely unexpected) widening of the perspective telling us not only proteins but even RNA systems display allosteric and cooperative features. Besides the applicative value for exploiting gene expression regulation, these findings open the way to very stimulating hypotheses about an RNA world at the very origin of life.

We can safely state that the authors contributing to this book are representatives of the frontiers of integrated approaches to protein science and the picture emerging is in turn coherent and fully integrated, thus contributing to the recovery from the fragmentation of scientific thought that is probably the most dangerous menace facing science in these times.

Rome, Italy

Luisa Di Paola
Alessandro Giuliani

Contents

Contributors

Victor S. Batista • *Department of Chemistry, Yale University, New Haven, CT, USA; Energy Sciences Institute, Yale University, New Haven, CT, USA*

Claudio Bucolo • *Department of Biomedical and Biotechnological Sciences, Section of Pharmacology, School of Medicine, University of Catania, Catania, Italy; Center for Research in Ocular Pharmacology—CERFO, University of Catania, Catania, Italy*

Luisa Di Paola • *Unit of Chemical-Physics Fundamentals in Chemical Engineering, Department of Engineering, Università Campus Bio-Medico di Rome, Rome, Italy*

Almerinda Di Venere • *Department of Experimental Medicine and Surgery, University of Rome "Tor Vergata", Rome, Italy*

Anasuya Dighe • *IISc Mathematics Initiative (IMI), Indian Institute of Science, Bangalore, India*

Alessandro Finazzi Agrò • *Department of Experimental Medicine and Surgery, University of Rome "Tor Vergata", Rome, Italy*

Vasundhara Gadiyaram • *IISc Mathematics Initiative (IMI), Indian Institute of Science, Bangalore, India*

Sambit Ghosh • *Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India; Department of Chemical and Biological Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA*

Alessandro Giuliani • *Environment and Health Department, Instituto Superiore di Sanità, Rome, Italy*

Melanie Goldstein • *Department of Chemistry and Biochemistry, Montclair State University, Montclair, NJ, USA*

Nina M. Goodey • *Department of Chemistry and Biochemistry, Montclair State University, Montclair, NJ, USA*

Guang Hu • *Center for Systems Biology, School of Biology and Basic Medical Sciences, Soochow University, Suzhou, China*

Matteo Lambrughi • *Computational Biology Laboratory, Danish Cancer Society Research Center, Copenhagen, Denmark*

David M. Leitner • *Department of Chemistry and Chemical Physics Program, University of Nevada, Reno, NV, USA*

Claire Lesieur • *Institut Rhônalpin des systèmes complexes, IXXI-ENS-Lyon, Lyon, France; AMPERE, CNRS, Univ. Lyon, Lyon, France*

Buyong Ma • *Basic Science Program, Leidos Biomedical Research, Inc. Cancer and Inflammation Program, National Cancer Institute, Frederick, MD, USA*

Giampiero Mei • *Department of Experimental Medicine and Surgery, University of Rome "Tor Vergata", Rome, Italy*

Ruth Nussinov • *Basic Science Program, Leidos Biomedical Research, Inc. Cancer and Inflammation Program, National Cancer Institute, Frederick, MD, USA; Sackler Institute of Molecular Medicine, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel*

Elena Papaleo • *Computational Biology Laboratory, Danish Cancer Society Research Center, Copenhagen, Denmark*

Alla Peselis • *Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY, USA*
Chiara Bianca Maria Platania • *Department of Biomedical and Biotechnological Sciences, Section of Pharmacology, School of Medicine, University of Catania, Catania, Italy*
Korey M. Reid • *Department of Chemistry and Chemical Physics Program, University of Nevada, Reno, NV, USA*
Ivan Rivalta • *Dipartimento di Chimica Industriale "Toso Montanari", Università di Bologna, Bologna, Italy; Univ Lyon, Ens de Lyon, CNRS, Université Lyon 1, Laboratoire de Chimie UMR 5182, Lyon, France*
Alexander Serganov • *Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY, USA*
Adnan Sljoka • *Department of Informatics, School of Science and Technology, Kwansei Gakuin University, Sanda, Hyogo, Japan; CREST, Japan Science and Technology Agency (JST), Tokyo, Japan; RIKEN, Center for Advanced Intelligence Project, Tokyo, Japan*
Valentina Sora • *Computational Biology Laboratory, Danish Cancer Society Research Center, Copenhagen, Denmark*
Matteo Tiberti • *Computational Biology Laboratory, Danish Cancer Society Research Center, Copenhagen, Denmark*
Saraswathi Vishveshwara • *Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India*
Laurent Vuillon • *LAMA, Univ. Savoie Mont Blanc, CNRS, LAMA, Le Bourget du Lac, France*
Jun Zhao • *Cancer and Inflammation Program, National Cancer Institute, Frederick, MD, USA*

Chapter 1

Allostery: The Rebound of Proteins

Alessandro Finazzi Agrò and Giampiero Mei

Abstract

The discovery of hemoglobin allosteric properties is briefly summarized and contextualized in the frame of the main biochemical revelations that characterized the first half of the XX century. In particular, the historical background of DNA, RNA, and protein structure research is recalled and the new role that protein-protein interaction may have on allosteric regulation is discussed.

Key words Myoglobin, Hemoglobin, Allosteric effect, Oligomerization

The history of Biochemistry in the past century can be jokingly, but even not so much, due to pertinacity and pride of the respective supporters, described as a fight for supremacy among proteins, DNA and RNAs.

No doubt that the first half of twentieth century was dominated by the biochemistry of proteins, especially those endowed with enzymatic or other functionally well understandable functions. The first and most famous among this class was, and is, hemoglobin, due to two remarkable and useful properties: our connatural fascination (or aversion) for blood and the abundance and relative easiness of its purification. In fact, hemoglobin was the first protein to be purified and crystallized more than 150 years ago [1]. From thence and for the following hundred years, proteins were the main character on the biochemistry stage. Not by chance, their collective name, proteins, comes indeed from the Greek word *πρωτειον* (proteion, primary) [2]. Another important ability of proteins was afterwards discovered: their role as organic catalysts soon denominated to enzymes (again a Greek-derived word: *ενζυμον*, enzumon = in leaven) [3]. As a matter of fact, already for thousands of years the ability of yeast to convert organic matter into human's more desirable foods (beer, leavened bread, wine) was known and exploited.

The elemental chemistry, amino acid composition, primary (amino acids sequence), secondary, tertiary, and quaternary

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_1,

structure of proteins required the strenuous effort of a host of brilliant biochemists, chemists, and physicists. In the past century, many of them have been awarded by Nobel prize. The first big achievement was made by Max Perutz who described the structure of hemoglobin tetramer [4], which gained him and his fellow John Kendrew (for the structure of the cognate protein myoglobin) the 1962 Nobel Prize in Chemistry.

In the same years, a second star molecule, DeoxyriboNucleic Acid (DNA), was gaining moment. First described by Miescher [5] in 1869 and analyzed by Levene [6] in 1919, DNA took the centre stage in 1953, thanks to a most famous paper by Watson and Crick [7], perhaps one of the clearest and more provoking scientific articles that everyone studying biology must read.

In this paper, James and Francis put down the basis of genetic code which since then has been the favorite issue of thousands of scientists all over the world, of many pharmaceutical and diagnostic firms, and even of lay people whose a frequent locution became: "It is in my (his, their, its) DNA".

This "fashion" reached its climax around the turning of millennium when the whole human genome was decrypted [8] and the individual DNA sequence made available by skilled entrepreneurs for a handful of dollars. It came out that only a tiny fraction ($\leq$2%) of the DNA filament was actually coding for proteins, inducing several scientists to preposterously call the remaining part "junk DNA" [9]. However, after a few years, the finding of new functions for the noncoding regions of the DNA molecule induced the scientific community to reconsider the matter [10].

In the meantime, it started a more accurate search for the role of RiboNucleic Acid (RNA), which till then considered no more than an efficient and compliant servant with three different attitudes (messenger, mRNA; transfer, tRNA; and ribosomal, rRNA). This simplistic view was soon abandoned by the discovery of several new types and functions of RNA. Just to mention a few, we today acknowledge the presence and function of heterogeneous nuclear (hnRNA), small interfering (siRNA), short hairpin (shRNA), piwi-interacting (piRNA), micro (miRNA), and small nucleolar (snoRNA) ribonucleic acids, not to mention double strand RNA (dsRNA) found in some viruses.

The most recent finding about RNA is its ability to serve as a catalyst in several reactions involving DNA, proteins, and RNA itself [11, 12]. This particular ability gave room to hypothesize a "RNA world" when ribonucleic acid might have been the unopposed king of all the living matter, before the appearance of proteins on earth and then the takeover by DNA as repository of genetic information [13, 14].

Nonetheless, it was evident that proteins, with their much greater possibility of variation, these being written with 20 different

letters, in comparison to the poorer four letters code of nucleic acids, could play much more characters in the life programme.

Notwithstanding, proteins appeared to be downgraded in a second row by the explosion of nucleic acids research and importance, as proven by the relative number of paper published, and by the fall of impact factors attributed to protein structure-function-devoted journals.

In these days, however, most likely to what happens with fashion, proteins are regaining a central role, thanks to rediscovery of an old concept, namely allostery. Allostery is again a Greek-derived word: $\alpha\lambda\lambda o\sigma$ (allos) = different and $\sigma\tau\varepsilon\rho\varepsilon o\sigma$ (stereos) = solid (object). Allostery is a term created by the team of Jacques Monod at the Institute Pasteur just more than 50 years ago.

As Jean-Pierre Changeux put it: the history and significance of the word "allosteric" is directly associated with the 1961 26th Cold Spring Harbor Symposium on Quantitative Biology entitled "Cellular regulatory mechanisms." The word "allosteric" was not orally pronounced during the meeting. It appeared for the first time in the printed version of the Proceedings: in the General Discussion written by Monod and Jacob as a conclusion of the meeting [15].

Nowadays, the term "allosteric" appears in more than 24,000 articles since the famous "number one" [16].

Among this ocean of scientific literature, there are several recent reviews [17–22] that render almost worthless any effort of writing a new one. But at least two articles deserve special mention for their significance in the development of the field, both connected with the evergreen model allosteric molecule, i.e., hemoglobin. The first, seminal for the quantitative treatment of allostery [23], was written by the two French researchers of Pasteur Institute, Jacques Monod and Jean Paul Changeux, and by an American scientist, Jeffries Wyman, who after wandering in USA, Europe, and Middle East found a convenient and comfortable niche in the Rome University at the Institute of Biological Chemistry. This apparently odd choice of him, who had a chair at Harvard, was due, besides the fascination of the Eternal City, to the presence in the Biochemistry Institute of the University (at those times the only one present in Rome) of two outstanding personalities: Alessandro Rossi Fanelli and Eraldo Antonini, both of them, together with their coworkers, already renowned for their studies on hemoglobin and the cognate protein myoglobin. So, it was not by chance or by preposterous self-esteem that hemoglobin gained the nickname "Roman molecule."

Anyhow, hemoglobin is by far the most studied allosteric protein and is a reference for every polymeric protein showing this kind of behavior, i.e., to change its activity upon binding at some distant site an effector molecule. Soon after the publication of the so-called MWC paper [23], Perutz gave structural ground to the theory [24]. Nonetheless, this theory was debated even among the same

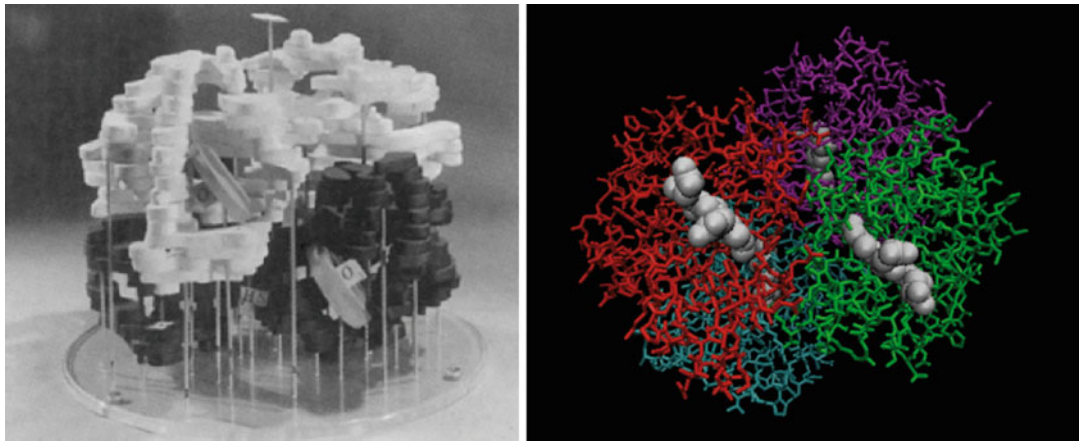

Fig. 1 The advent of computer science and the model of hemoglobin, the most representative oligomeric protein for allostery. Left panel (**a**): the original structure (at 5.5 Å resolution) solved by Max Perutz and John Kendrew, thanks to the introduction of computational biology (i.e., computer routines, since 1951) and the isomorphous replacement methodology (introduced by Max Perutz in the 1930s). Right panel (**b**): a re-elaboration of the human deoxyhemoglobin structure at 1.74 Å resolution (pdb file 2hhb), deposited by Max Perutz and collaborators in the protein data bank, in 1984

proponents [25, 26] and subsequently an alternative hypothesis was formulated, i.e., the so-called "sequential model" put forward by Daniel Koshland, George Némethy, and David Filmer [27]. Nowadays, the allosteric behavior of hemoglobin is still matter of research and comments [28].

It is important to recall that deciphering the structure of hemoglobin required a titanic computational effort to reach a still unsatisfactory resolution of 5.5 Å (consider that at this resolution it is impossible to establish the exact position of the amino acids side chains, Fig. 1a). Only more than thirty years later, Perutz and coworkers were able to increase the resolution down to 1.75 Å, allowing a more reliable modeling of the protein and its allosteric properties [29]. This important achievement was essentially due to the development of computers. It might be recalled that the first data on hemoglobin (or myoglobin) crystallographic structure required weeks of almost hand calculations and modelling that nowadays could be performed in few hours [30].

Thanks to this new computational power, further models of allostery are arising, making anew proteins the main characters of biological research [31]. These models do not consider as a fundamental property for an allosteric protein to be oligomeric, or to show well-defined "active, R" and "inactive, T" states, as initially described for hemoglobin and congeners. Instead, a concept of "dynamic allostery" started to gain momentum [32] till a general view of "conformation selection and population shift" became recognized to possibly explain all the observed allosteric phenomena [33]. Here, not only the binding of a small molecule may affect

the overall shape and activity of a protein, but a close interaction between two, or more, different proteins may change completely their conformation and physiological role. In this way, it was possible to understand the multiple role of some proteins, for instance when they get in contact with other proteins inside a membrane, giving rise to a receptor. This completely new approach renders very actual the definition of allostery as "the second secret of life", prophetically disclosed by Jacques Monod [34] and subsequently confirmed and extended by many others [35]. Thus, the supremacy of proteins seems today to be definitely accepted.

References

1. Funke O (1851) Über das Milzvenenblut. Zeitschrift für Ration Medizin 1:172–218
2. Mulder G (1838) Sur la composition de quelques substances animals. Bull des Sci Phys Nat en Neerl 1:104
3. Kühne W (1877) Über das Verhalten verschiedener organisirter und sog. ungeformter Fermente. Verhandlungen des naturhistorisch-medicinischen Vereins zu Heidelberg. New Ser 1:190–193
4. Perutz MF, Rossmann MG, Cullis AF et al (1960) Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5-{Å} resolution, obtained by X-ray analysis. Nature 185:416–422
5. Dahm R (2005) Friedrich Miescher and the discovery of DNA. Dev Biol 278:274–288. https://doi.org/10.1016/j.ydbio.2004.11.028
6. Levene P (1919) The structure of yeast nucleic acid. J Biol Chem 40:415–424
7. Watson JD, Crick FH, Pelz B et al (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737–738. https://doi.org/10.1126/science.aaf5508
8. Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. https://doi.org/10.1038/35057062
9. Ohno S (1972) So much 'junk' DNA in our genome. Brookhaven Symp Biol 23:366–370
10. Pennisi E (2012) ENCODE project writes eulogy for junk DNA. Science 337:1159–1161. https://doi.org/10.1126/science.337.6099.1159
11. Fedor MJ, Williamson JR (2005) The catalytic diversity of RNAs. Nat Rev Mol Cell Biol 6:399–412. https://doi.org/10.1038/nrm1647
12. Lee K-Y, Lee B (2017) Structural and biochemical properties of novel self-cleaving ribozymes. Molecules 22:678. https://doi.org/10.3390/molecules22040678
13. Robertson MP, Joyce GF (2012) The origins of the RNA World. Cold Spring Harb Perspect Biol 4:1. https://doi.org/10.1101/cshperspect.a003608
14. Higgs PG, Lehman N (2014) The RNA World: molecular cooperation at the origins of life. Nat Rev Genet 16:7–17. https://doi.org/10.1038/nrg3841
15. Changeux JP (2011) 50th anniversary of the word 'allosteric'. Protein Sci 20:1119–1124. https://doi.org/10.1002/pro.658
16. Dokholyan NV (2016) Controlling allosteric networks in proteins. Chem Rev 116 (11):6463–6487. https://doi.org/10.1021/acs.chemrev.5b00544
17. Cárdenas ML (2013) Michaelis and Menten and the long road to the discovery of cooperativity. FEBS Lett 587:2767–2771. https://doi.org/10.1016/j.febslet.2013.07.014
18. Cornish-Bowden A (2014) Understanding allosteric and cooperative interactions in enzymes. FEBS J 281:621–632. https://doi.org/10.1111/febs.12469
19. Swain JF, Gierasch LM (2006) The changing landscape of protein allostery. Curr Opin Struct Biol 16:102–108. https://doi.org/10.1016/j.sbi.2006.01.003
20. Motlagh HN, Wrabl JO, Li J, Hilser VJ (2014) The ensemble nature of allostery. Nature 508:331–339. https://doi.org/10.1038/nature13001
21. Gunasekaran K, Ma B, Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? Proteins 57:433–443
22. Nussinov R (2016) Introduction to protein ensembles and allostery. Chem Rev

116:6263–6266. https://doi.org/10.1021/acs.chemrev.6b00283
23. Monod J, Wyman J, Changeux JP (1965) On the nature of allosteric transitions: a plausible model. J Mol Biol 12:88–118. https://doi.org/10.1016/S0022-2836(65)80285-6
24. Perutz MF (1970) Stereochemistry of cooperative effects in haemoglobin. Nature 228:726–734. https://doi.org/10.1038/228726a0
25. Buc H (2013) The design of an enzyme: a chronology on the controversy. J Mol Biol 425:1407–1409. https://doi.org/10.1016/j.jmb.2013.03.015
26. Crick FHC, Wyman J (2013) A footnote on allostery. J Mol Biol 425:1500–1508. https://doi.org/10.1016/j.jmb.2013.03.012
27. Koshland DE, Némethy G, Filmer D (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits*. Biochemistry 5:365–385. https://doi.org/10.1021/bi00865a047
28. Brunori M (2014) Variations on the theme: allosteric control in hemoglobin. FEBS J 281:633–643. https://doi.org/10.1111/febs.12586
29. Fermi G, Perutz MF, Shaanan B, Fourme R (1984) The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. J Mol Biol 175:159–174. https://doi.org/10.1016/0022-2836(84)90472-8
30. Edwards D, Hubbard R (2006) Computer and protein crystallography. In: Ekins S (ed) Computer applications in pharmaceutical research and development. Wiley, New York, pp 277–300
31. Brunori M (2011) Allostery turns 50: is the vintage yet attractive? Protein Sci 20:1097–1099. https://doi.org/10.1002/pro.660
32. Cooper A, Dryden DTF (1984) Allostery without conformational change—a plausible model. Eur Biophys J 11:103–109. https://doi.org/10.1007/BF00276625
33. Liu J, Nussinov R (2016) Allostery: an overview of its history, concepts, methods, and applications. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1004966
34. Monod J (1977) Chance and necessity: essay on the natural philosophy of modern biology. Penguin Group, London
35. Fenton AW (2008) Allostery: an illustrated definition for the 'second secret of life'. Trends Biochem Sci 33:420–425. https://doi.org/10.1016/j.tibs.2008.05.009

Chapter 2

Disclosing Allostery Through Protein Contact Networks

Luisa Di Paola, Giampiero Mei, Almerinda Di Venere, and Alessandro Giuliani

Abstract

Proteins are located in the twilight zone between chemistry and biology, where a peculiar kind of complexity starts. Proteins are the smallest 'devices' showing a sensible adaptation to their environment by the production of appropriate behavior when facing a specific stimulus. This fact qualifies (from the 'effector' side) proteins as nanomachines working as catalysts, motors, or switches. However (from the sensor side), the need to single out the 'specific stimulus' out of thermal noise qualifies proteins as information processing devices. Allostery corresponds to the modification of the configuration (in a broad sense) of the protein molecule in response to a specific stimulus in a non-strictly local way, thereby connecting the sensor and effector sides of the nanomachine. This is why the 'disclosing' of allostery phenomenon is at the very heart of protein function; in this chapter, we will demonstrate how a network-based representation of protein structure in terms of nodes (aminoacid residues) and edges (effective contacts between residues) is the natural language for getting rid of allosteric phenomena and, more in general, of protein structure/function relationships.

Key words Protein contact networks, Network descriptors, Spectral clustering

1 Introduction

Allostery is a neologism modeled upon Greek language, which has to do with the ability of proteins to transmit a signal from one site to another in response to environmental stimuli. This ability is related to the transmission of information across the protein molecule from a sensor (allosteric) site to the effector (binding) site [1]. The molecule, hence, perceives ligand binding at a distance from the active site, or any other microenvironmental perturbation, like pH changes. The information transfer across protein molecules can be approached by many different methods going from experimental (change in affinity of the enzymatic systems upon allosteric stimulus exposition) to structural (comparison of X-ray or NMR structures correspondent to different activation states) and theoretical (molecular dynamics simulation of allosteric agent binding)

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_2,

perspectives. Here we propose an approach that already gave very interesting results in allostery elucidation [1–3], building upon the consideration of protein structures as network system having aminoacid residues as nodes and the presence of an effective non-covalent (and thus folding-induced) contact between residues as edges.

This approach is conveniently located at half-way between theoretically intensive and structural/experimental ones: while relying upon a fully quantitative formalization (network topological descriptors), the adopted mathematics is intuitive (simple count statistics on nodes and edges) and free from any physical constraint or hypothesis. On the other hand, the relation between network and 3D structure is univocal and clear and the statements coming from network analysis readily testable by experimental means.

Still more important is the fact that network formalism allows for a 'naturally multiscale' approach to allostery. Network graph-theoretical approaches are located half-way between bottom-up and top-down approaches focusing on the relation between the elements of the studied phenomenon. We can roughly describe the network approach as the answer to the question "What can we derive from the sole knowledge of the wiring diagram of a system?" [4]. A graph G is a mathematical object made of a finite set of vertices (or nodes) V and a collection of edges E connecting two vertices; in our case we deal with the simplest form of graphs: nondirected graphs whose edges can be traversed in both directions and have the same strength. This is formally equivalent to a binary two entries matrix having as rows (columns) the nodes and a 1/0 value at i, j cross (being 1 marking the presence and 0 the absence of an edge, Fig. 1).

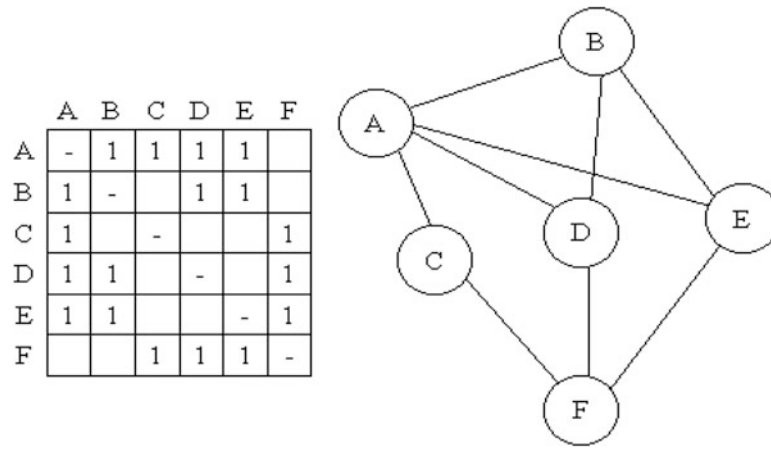

Fig. 1 The incidence (adjacency) matrix (left) is isomorphic to the network (right) formalization

It is immediately possible to compute topological descriptors at all the levels of definition from single residue (node) to the entire protein (whole network) passing by cluster of nodes (structural domains).

Thus, we can compute the degree of each node (how many edges correspond to a residue) that is a local, microscopic feature of the system or the "average shortest path" corresponding to the average length of minimal paths connecting all the node pairs (this is a wiring architecture global descriptor) or to group the residues into maximally connected clusters (a mesoscale feature) [5, 6]. The unique feature of graph formalization is that these different levels views are strictly intermingled and cannot by no way be considered as independent and/or simply corresponding to gross averages: a residue endowed with a low average shortest path with other residues (single node property) obtains its value thanks to the entire network wiring architecture (top-down causation), while in the same time influences the flux of information (energy) across the entire molecule (bottom-up causation). This natural interaction of scales allows to disentangle the different aspects of allostery from the recognition of the aminoacid residues more involved in signal transmission (microscopic layer) to the 'paths' linking sensor and effector sites (mesoscopic layer) and the modifications of the entire graph wiring architecture (macroscopic layer).

In the following, after a general introduction to complex network analysis and a thorough definition of the main topological descriptors, we will present some practical examples of network approaches to allostery.

2 Materials

2.1 Structural Data

The computational approach of protein contact networks relies on the availability of structural information on proteins.

The reference database for protein structures is the Protein Data Bank [7] (PDB, http://www.rcsb.org/), which also defines the PDB format, a standard for recording atom files. All files recorded on the Protein Data Bank repository follow the PDB format. Information in PDB files is organized in lines, named records. The PDB files include many types of records, recognizable by the line header and arranged in a given fashion, to convey all structural data through a standard format.

Atom coordinates are reported in the ATOM record, organized as reported in Table 1.

The information of interest for the protein contact networks constructions are the coordinates recorded in columns 31–54. Notice that alpha carbons are distinguished by other carbons in the residues and denoted by the character string "CA".

Table 1
Description of the coordinates section of PDB files

Columns	Data	Justification	Data type
1–4	"ATOM"		Character
7–11	Atom serial number	Right	Integer
13–16	Atom name	Left	Character
17	Alternate location indicator		Character
18–20	Residue name	Right	Character
22	Chain identifier		Character
23–26	Residue sequence number	Right	Integer
27	Code insertions of residues		Character
31–38	X orthogonal coordinate (Å)	Right	Real
39–46	Y orthogonal coordinate (Å)	Right	Real
47–54	Z orthogonal coordinate (Å)	Right	Real
55–60	Occupancy	Right	Real
61–66	Temperature factor	Right	Real
73–76	Segment identifier	Left	Character
77–78	Element symbol		Character

2.2 Protein-Ligand Binding, Allosteric Proteins, and Protein-Protein Interactions Databases

Protein-ligand databases provide information on protein-ligand binding thermodynamics and the structural data (PDB reference). Correlation between binding affinity data and structural (network topology) descriptors allows for a deep insight in the binding mechanism.

BindingDB [8] is a public, web-accessible database (http://www.bindingdb.org/) of experimental binding affinities, with a special regard to potential protein-drug targets. At the moment, the database contains 2291 protein-ligand crystal structures with BindingBD affinity measurements for proteins with 100% sequence identity and 5816 crystal structures allowing proteins to 85% sequence identity.

Analogously, PDBbind database [9] catalogued 5671 protein-ligand complexes out of 19,261 experimental structures (in 2003 at the moment of the database publication), whose 1359 matched with binding affinity data.

More recently, BioLIP [10] (http://zhanglab.ccmb.med.umich.edu/BioLiP/) is a database of biologically relevant protein-ligand complexes; the biological relevance of recorded complexes requires a careful manual verification. The current version (updated on May 26th, 2017) contains 81,811 protein structures from PDB, of those 23,492 with binding affinity data.

3 Methods

3.1 Protein Contact Networks Definition

Protein Contact Networks (PCNs) are graphs whose nodes (or vertices) are the protein residues and links (or edges) between the i-th and the j-th nodes (residues) occur if the distance between the two residues d_{ij} is higher than 4 and lower than 8 Å. The lower end excludes all covalent bonds (disulfide and peptidic bonds), which are not sensible to environment change (so to protein functionality), while the upper end gets rid of weaker non-covalent bonds (so not significant for protein functionality).

So, the first step is to extract from the PDB file the coordinate for all alpha-carbon atoms: with reference to Table 1, in lines with "ATOM" as header take only coordinates (columns 31–54) for alpha carbons (if "Atom Name"—column 13–16—is "CA"), keeping record as well of the "Residue name" (column 18–20).

As a result, a matrix $N \times 3$ is reporting X, Y, and Z coordinates of alpha carbons for all N residues, ordered according to the primary sequence. Another vector will keep trace of residue names in the sequence.

Starting from the coordinates matrix, it is possible to build up a distance matrix, whose generic element d_{ij} reports the Euclidean distance between residues i and j:

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z\right)^2} \tag{1}$$

(x_i, y_i, z_i) and (x_j, y_j, z_j), respectively, being the cartesian coordinates of residue i and j.

At this point, it is possible to build up adjacency matrix **A**, whose generic element is defined as:

$$A_{ij} = \begin{cases} 1 & if\ 4^{\circ}\mathrm{A} < d_{ij} < 8^{\circ}\mathrm{A} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The adjacency matrix **A** is the mathematical descriptor of unweighted, undirected graphs, from which all main topological descriptors can be derived (*see* **Note 1**). The adjacency matrix **A** can be visualized as matrix plot (Fig. 2). It is evident that the nature of **A** is sparse matrix.

3.2 Computation of Descriptors

Network topology translates into mathematical descriptors, derived from the adjacency matrix **A**. As follows, the definition of main descriptors and their relevance in protein structure and functionality description.

1. *Node degree*: the degree of the i-th node (residue) k_i computes the number of links the node participates in. It can be easily computed as the sum of elements of the i-th row or column:

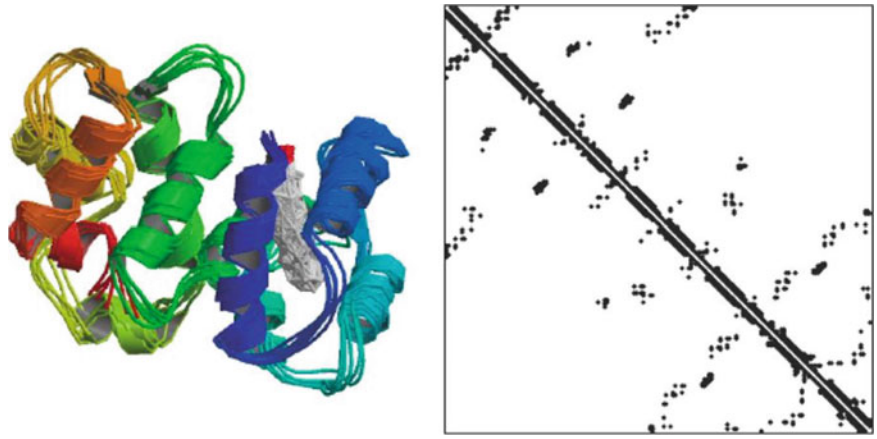

Fig. 2 The recoverin's three-dimensional structure (left) is translated into a network, graphically represented by means of the matrix plot of the adjacency matrix (right). Reprinted with permission from [11]

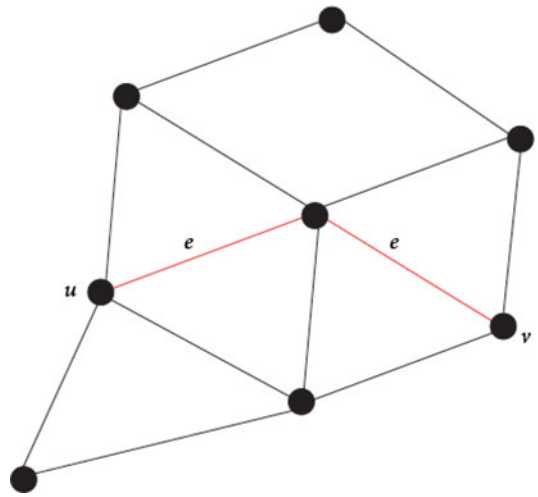

Fig. 3 Graphical representation of a graph with eight nodes and 13 total links: nodes u and v are connected by two links (shortest path). Reprinted with permission from [14]

$$k_i = \sum_j A_{ij} \tag{3}$$

The degree distribution helps classifying networks according to models (random, regular, small-world). Notice that protein contact networks escape classification, roughly described as small-world networks, albeit presenting key features of random networks (such as, Gaussian degree distribution).

2. *Shortest path*: the shortest path sp_{ij} between the i-th and j-th node describes the lowest number of links connecting the two residues (*see* Fig. 3). Algorithms to solve the shortest path

problems root in the information theory [12, 13]. On the web, libraries in Java and Python are available to compute shortest paths, mainly based on Dijkstra's algorithms [13].

The average shortest path, also known as network characteristic length, plays a key role in signal transmission throughout protein contact networks and is of a major relevance in protein functionality [15]. Notice that the shortest path is averaged over all possible node pairs $\frac{N\bar{n}(N-1)}{2}$

3. *Centrality metrics*: how central is a role is the first address to topological role of nodes in the network. This concept flanks the continuous quest for functional role of residues in proteins, which is also a topic in allostery identification. The simplest centrality metrics is the above-mentioned node degree, which describes the local structure of networks (and of proteins). However, allostery and, more in general, cooperativity have to do with global signatures of protein structures, which are represented by corresponding global network properties. In this perspective, the functional relevance of centrality must pass through a global analysis of protein structures and corresponding contact networks. Global centrality metrics pass by shortest paths computation [16]: the most relevant in the protein contact networks analysis are:
   (a) closeness centrality [17]: it is defined as the inverse of fairness, in turn defined for a node, as the sum of the shortest paths with all other nodes. Given a set of vertices V, the closeness centrality of a node i is defined as:

$$close(i) = \frac{1}{\sum_{u \in V, u \neq i} sp_{ui}} \tag{4}$$

Amitai and coworkers [18] demonstrated that residues in active sites show high closeness, but no clue comes as for residues acting in cooperative processes (allostery).

   (b) betweenness centrality (*see* **Note 2**): it describes for a node the number of shortest paths coming through it. Given a set of vertices V, the betweenness centrality of a node i is defined as:

$$betw(i) = \sum_{v \in V, v \neq i} \sum_{u \in V, u \neq i} \frac{\sigma_{uv}(i)}{\sigma_{uv}} \tag{5}$$

σ_{uv} being the total number of shortest paths connecting nodes u and v and $\sigma_{uv}(i)$ the number of shortest paths connecting the two nodes and also coming through the node i.

The betweenness centrality is among the most relevant descriptor to identify the role of residues (nodes) in the signal transmission network of protein structures. In terms of network robustness, the focused removal of high betweenness nodes is detrimental in the whole network connectivity.

4. *Graph energy*: it is defined as the sum of the absolute values of the adjacency matrix eigenvalues. This descriptor is strongly grounded into Chemical Graph Theory, aimed at describing chemico-physical properties of organic molecules from connectivity indices of chemical structural graphs [19]. Graph energy for protein contact networks highlights special features of enzymatic allostery [20] and of protein-protein interactions [21].

3.3 Clustering

Cooperativity in proteins is a direct effect of their modularity [22]. Thus, a key step in protein allostery analysis is to identify modules in protein structures, which somehow correspond to recognized functional domains.

Protein contact networks formalism strongly helps in this direction by means of network clustering: clusters in protein contact networks well match with protein domains [20].

Two methods have been devised to partition PCNs into clusters [23]: a geometrical method, based on the k-means algorithm and spectral clustering, which demonstrated to be very effective in identifying functional regions in proteins (*see* the azurin case discussed in ref. 23).

Spectral clustering is based on the spectral decomposition of the laplacian matrix **L**, defined as follows:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} \tag{6}$$

A being the adjacency matrix and **D** the degree matrix, i.e., a diagonal matrix whose diagonal is the degree vector. The eigenvalue decomposition is applied to the laplacian **L**: the eigenvector corresponding to the second minor of eigenvalue $\boldsymbol{v_2}$ is of interest for the clustering partition. Considering the partition in two clusters, for instance, nodes are divided into the two clusters according to the sign of the corresponding components of the vector $\boldsymbol{v_2}$.

Spectral clustering is a clustering based on binary partition, so the number of clusters that can be obtained is only a power of two (2,4,8, ...).

Spectral clustering allows to identify cluster of nodes maximally interacting with each other, rather than nodes close in the space (such as in geometrical clustering). In this respect, spectral clustering is specifically apt to identify functional regions [23].

Once clusters have been specified (*see* **Note 3**), two descriptors assign the topological role of nodes with regard to their communication attitude:

(a) the participation coefficient is defined for the node i in the cluster s as:

$$P_i = 1 - \left(\frac{k_{si}}{k_i}\right)^2 \tag{7}$$

k_{si} being the degree of node i computed with respect only of nodes pertaining to the same cluster s and k_i the node i degree. High P nodes, thus, have strong connections with nodes pertaining to other clusters, so they are likely to play a key role in signal transmission between the two clusters (domains in PCNs).

(b) the intramodule connectivity z-score z defined for the node i in the cluster s as:

$$z = \frac{k_i - \overline{k_s}}{\sigma_s} \tag{8}$$

$\overline{k_s}$ and σ_s being the average and the standard deviation of intramodule degree in cluster s. High z nodes are responsible for the cluster stability, but are not likely to participate in communication between clusters.

The clustering profile of networks is aptly represented by means of P-z maps: in general, once the protein contact network is built, it is possible to use web resources to perform clustering and represent P-z maps [24]. Incidentally, P-z maps for protein contact networks have a very specific shape ("dentist's chair") strongly conserved for a large number of proteins. Figure 4 reports a typical P-z map for protein contact networks.

Eventually, it is useful to map the variation of the participation coefficient $\Delta P = P_{\text{bound}} - P_{\text{unbound}}$ upon binding: this vector is simply the difference between the P vectors for the unbound and

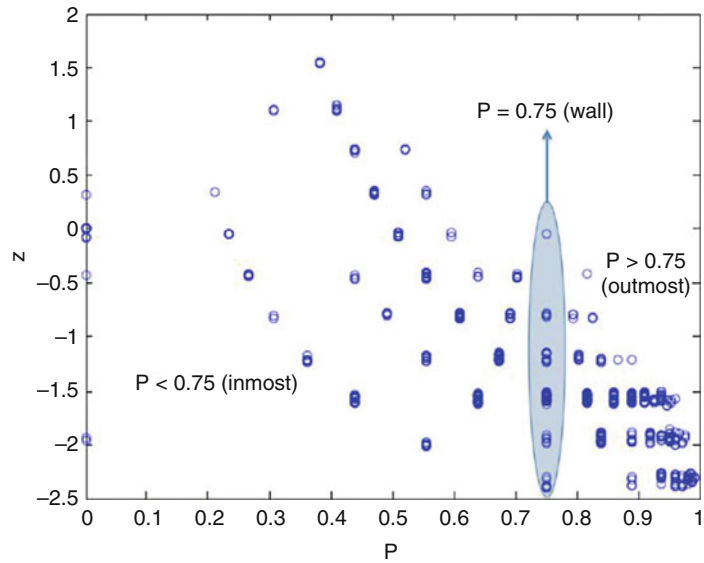

Fig. 4 A typical P-z map: P values of 0.75 represent a distinct break in the topological role of nodes: nodes with P higher than 0.75 share more links with nodes pertaining to the other clusters than with nodes on their same cluster. Reprinted with permission from [25]

bound (complex) form (if both are available). We previously demonstrated that these maps identify allosteric response of the protein structural networks [11, 20].

4 Applications

Herein, we present a short survey of possible applications of the method, with reference to some of our previous works and relevant literature in the field.

4.1 Allostery in Protein-Ligand Binding

The capital application of the method in allostery identification deals with the identification of allosteric activation in protein-ligand binding.

De Ruvo and coworkers highlighted a specific response in allosteric proteins, not present in non-allosteric proteins undergoing binding, such as albumin [11]. The paper focused on calcium-binding proteins, traditionally parted into sensors (allosteric) and buffers (non-allosteric). Starting on this well-defined classification, we reported projections of P onto protein sequence (Fig. 5).

In this work, we highlighted a sharp change of P values in regions close to the active site upon binding, while fainter changes are appreciable in non-allosteric forms (*see* Fig. 5).

In more recent works, we reported maps of ΔP (*see* **Note 4**) for different forms of a class of plant enzymes, which are in charge of many different reactions in common plant organisms. We compared also ΔP maps (*see* **Note 5**) with corresponding displacement, finding out an intriguing result: Fig. 6 reports results for cell invertase from *Arabidopsis Thaliana*. First of all, clustering well identifies functional domains (Fig. 6a). Then, upon binding, P increases in the region between the two domains (Fig. 6b). One may think it is only a result of a rigid displacement of the two domains, but the displacement map (Fig. 6c) tells a different story: only outer, more flexible regions move upon binding, while the interdomain region remains practically motionless.

This case is particularly intriguing since this allosteric activation does not result into a conformational change, falling in the category of entropically driven allosteric transitions—of great interest since only recently identified and classified [26].

4.2 Allostery in Protein-Protein Interactions

A very hot topic in the field of drug discovery is the identification of allosteric sites to modulate protein-protein interactions [27]. In this perspective, the allosteric drugs analysis strongly relies on the network paradigm [28].

In 2015, we presented a work focused on the analysis of the anthrax complexes [29]: it is a trimeric complex made up of a protective agent (PA), an edema factor (EF), and a lethal factor

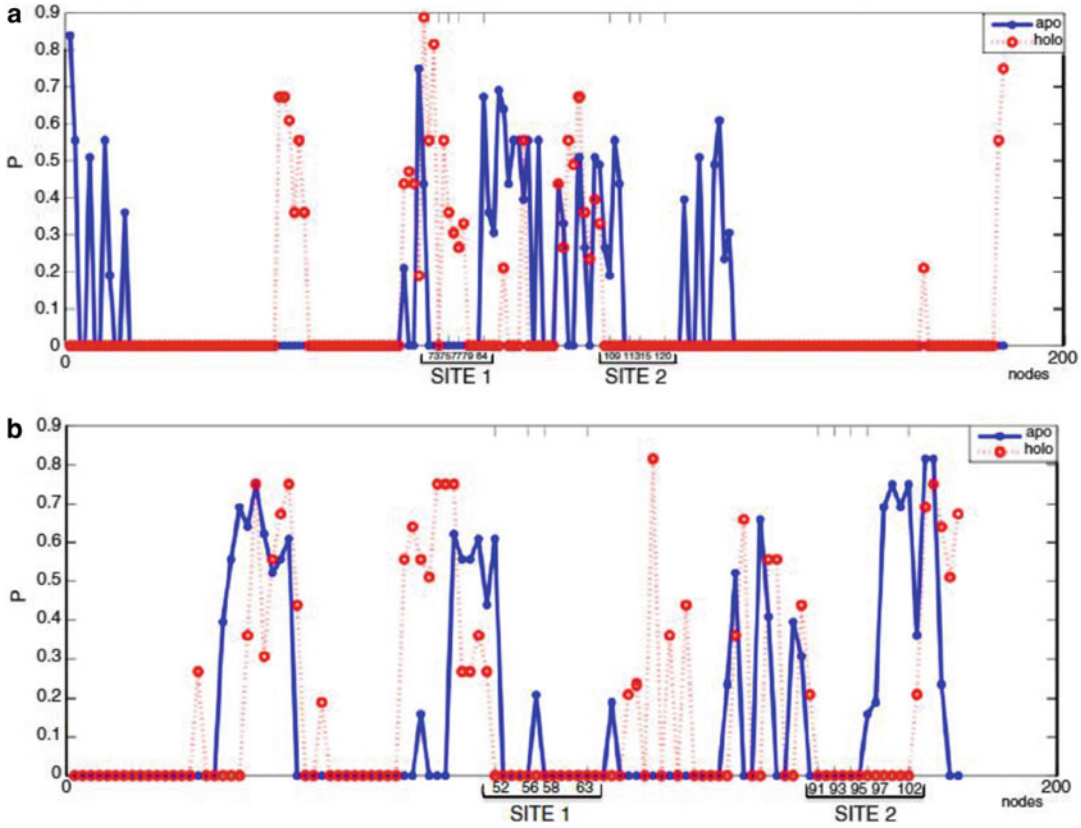

Fig. 5 *P* values distribution for apo (blue) and holo form (red) of recoverin (RC) (**a**) and of parvalbumin (PV) (**b**). Residues involved in the active sites are emphasized. Overall, *P* values abruptly change upon binding in RC, whereas, in the case of PV, *P* does not vary significantly. Reprinted with permission from [14]

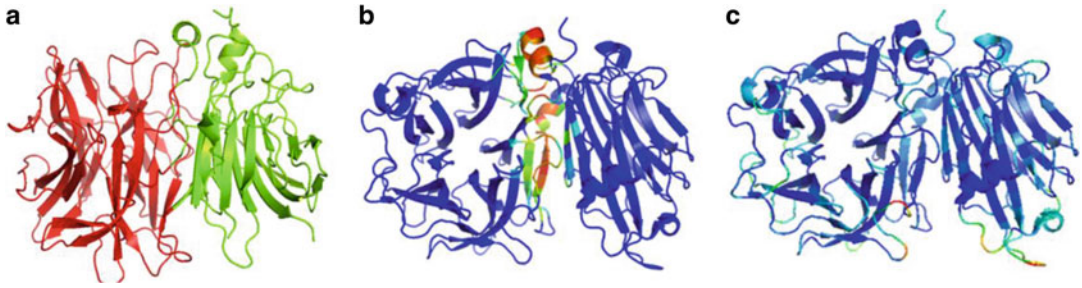

Fig. 6 Clustering results projected onto ribbon structures: application to cell invertase form *Arabidopsis thaliana.* (**a**) Clustering identifies domains in the apo form (PDB code 2 AC1); (**b**) ΔP projection upon binding with sucrose (PDB code 2 AC1); (**c**) displacement upon binding with sucrose. The region between the two domains activates upon binding, without displacement. This is an eminent case of allostery in play without relevant conformational transitions (only small displacement in outer, flexible regions). Reprinted with permission from [20]

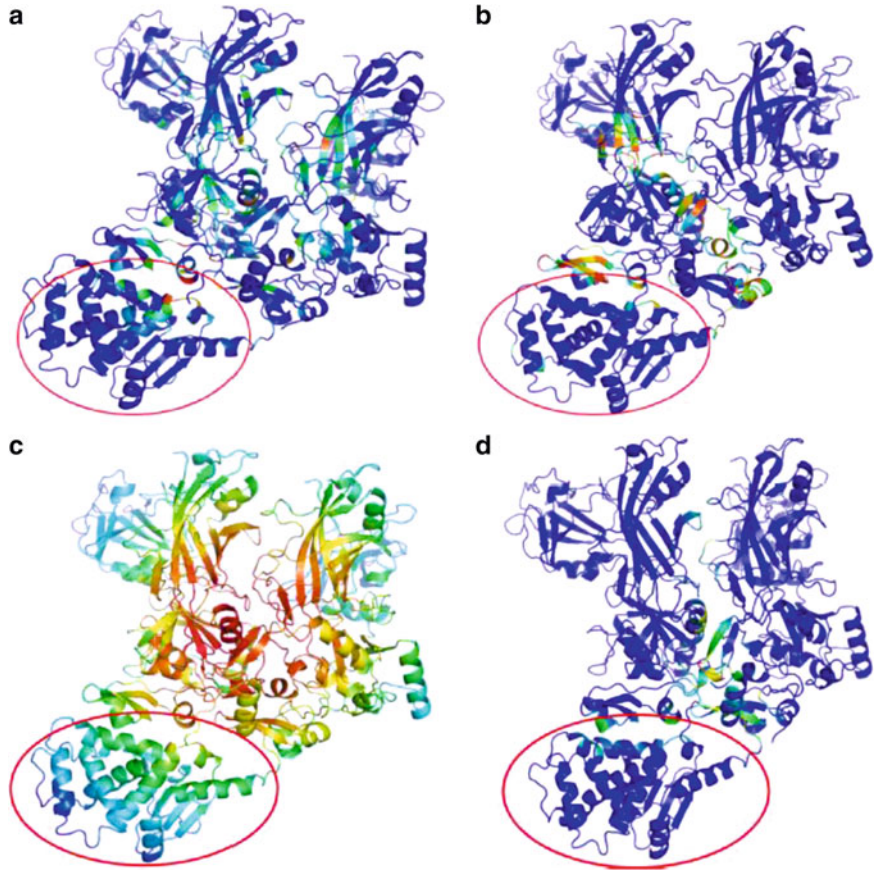

Fig. 7 Ribbon maps of network descriptors for the anthrax trimeric complex (PA-EF-LF). (**a**) betweenness centrality; (**b**) participation coefficient *P*; (**c**) closeness centrality; (**d**) interchain degree. The lethal factor LF is marked by the red circle. Reprinted with permission from [29]

(LF). The interface of protein-protein complexes was described by means of the interchain degree, i.e., the number of links nodes establish with nodes pertaining to a chain different from that they belong to (Fig. 7d). In this way, putative interface residues are highlighted. Further, *P* and betweenness centrality profiles identify probable allosteric sites (Fig. 7a, b). The closeness centrality reports only a kind of "rigidity" map, useless to identify functional nodes (Fig. 7c).

5 Notes

1. The distance cutoff to define links derives from chemico-physical considerations and statistical significance analysis [30]; however, a recent study reports a simple cutoff of 5 Å between residues center of mass as optimal for PCNs description [31]. This could be a useful option to define protein contact networks.

2. The betweenness centrality computation requires purposed algorithms; Brandes' algorithm is the most applied [32] and its version for Matlab, Java, and Python environments is available for free on web;
3. The clustering procedure is by far the most burdening part of the whole algorithm and relies on the solution for the eigenvalues problem; it is also a crucial step to define proper data structure to keep all useful information about clusters nodes. Writing an efficient algorithm for this part will strongly affect the whole program speed (shifting from many minutes to seconds);
4. When computing ΔP, it is necessary to verify that the apo (not bound) and holo (bound) forms are resolved with the same number of residues; in any case, it is advisable to verify their alignment through purposed software (*see* for instance, SuperPose http://wishart.biology.ualberta.ca/SuperPose/);
5. To create the ribbon maps, it is necessary to modify the PDB files, by replacing the B-factors in the ATOM section with the value of interest (i.e., node degree). To simplify the procedure, it is advisable to implement the whole algorithm in the same language of the molecular visualization system (Python for PyMol and Java for Jmol, for instance).

References

1. Di Paola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48. https://doi.org/10.1016/j.sbi.2015.03.001
2. Tsai CJ, del Sol A, Nussinov R (2009) Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. Mol BioSyst 5:207–216
3. De Ruvo M, Giuliani A, Paci P et al (2012) Shedding light on protein-ligand binding by graph theory: the topological nature of allostery. Biophys Chem 165–166:21–29. https://doi.org/10.1016/j.bpc.2012.03.001
4. Giuliani A, Filippi S, Bertolaso M (2014) Why network approach can promote a new way of thinking in biology. Front Genet. https://doi.org/10.3389/fgene.2014.00083
5. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393:440–442
6. Di Paola L, De Ruvo M, Paci P et al (2013) Protein contact networks: an emerging paradigm in chemistry. Chem Rev 113:1598–1613. https://doi.org/10.1021/cr3002356
7. Berman H, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res 28:235–242
8. Liu T, Lin Y, Wen X et al (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35:198–201
9. 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
10. Yang J, Roy A, Zhang Y (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res 41:1096–1103
11. De Ruvo M, Di Paola L, Giuliani A et al (2012) Shedding light on protein–ligand binding by graph theory: the topological nature of allostery. Biophys Chem 165–166:21–29. https://doi.org/10.1016/j.bpc.2012.03.001
12. Deo N, Pang C (1984) Shortest-path algorithms: taxonomy and annotation. Networks 14:275–323. https://doi.org/10.1002/net.3230140208

13. Johnson DB (1977) Efficient algorithms for shortest paths in sparse networks. J ACM 24:1–13. https://doi.org/10.1145/321992.321993
14. Hu G, Di Paola L, Liang Z, Giuliani A (2017) Comparative study of elastic network model and protein contact network for protein complexes: the hemoglobin case. Biomed Res Int 2017:2483264
15. Santoni D, Paci P, Paola LD, Giuliani A (2016) Are proteins just coiled cords? Local and global analysis of contact maps reveals the backbone-dependent nature of proteins. Curr Protein Pept Sci 17:26–29
16. Borgatti SP, Everett MG (2006) A graph-theoretic perspective on centrality. Soc Networks 28:466–484. https://doi.org/10.1016/j.socnet.2005.11.005
17. Sabidussi G (1966) The centrality index of a graph. Psychometrika 31:581–603. https://doi.org/10.1007/BF02289527
18. Amitai G, Shemesh A, Sitbon E et al (2004) Network analysis of protein structures identifies functional residues. J Mol Biol 344:1135–1146. https://doi.org/10.1016/j.jmb.2004.10.055
19. Bonchev DD, Rouvray DH (1990) Chemical graph theory: introduction and fundamentals. Gordon & Breach Science Publishers, London
20. Cimini S, Di Paola L, Giuliani A et al (2016) GH32 family activity: a topological approach through protein contact networks. Plant Mol Biol:1–10
21. Di Paola L, Mei G, Di Venere A, Giuliani A (2016) Exploring the stability of dimers through protein structure topology. Curr Protein Pept Sci 17:30–36. https://doi.org/10.2174/1389203716666150923104054
22. del Sol A, Araúzo-Bravo MJ, Amoros D, Nussinov R (2007) Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. Genome Biol 8:R92
23. Tasdighian S, Di Paola L, De Ruvo M et al (2013) Modules identification in protein structures: the topological and geometrical solutions. J Chem Inf Model 54:159–168
24. Cumbo F, Paci P, Santoni D et al (2014) GIANT: a cytoscape plugin for modular networks. PLoS One 9:e105001. https://doi.org/10.1371/journal.pone.0105001
25. Tasdighian S, Di Paola L, De Ruvo M et al (2014) Modules identification in protein structures: the topological and geometrical solutions. J Chem Inf Model 54:159–168. https://doi.org/10.1021/ci400218v
26. Tsai CJ, del Sol A, Nussinov R (2008) Allostery: absence of a change in shape does not imply that allostery is not at play. J Mol Biol 378:1–11
27. Nussinov R, Tsai C-J (2012) The different ways through which specificity works in orthosteric and allosteric drugs. Curr Pharm Des 18:1311–1316. https://doi.org/10.2174/138920012799362855
28. Csermely P, Nussinov R, Szilágyi A (2013) From allosteric drugs to allo-network drugs: state of the art and trends of design, synthesis and computational methods. Curr Top Med Chem 13:2–4. https://doi.org/10.2174/1568026611313010002
29. Di Paola L, Platania CBM, Oliva G et al (2015) Characterization of protein–protein interfaces through a protein contact network approach. Front Bioeng Biotechnol 3:170. https://doi.org/10.3389/fbioe.2015.00170
30. Di Paola L, De Ruvo M, Paci P et al (2012) Protein contact networks: an emerging paradigm in chemistry. Chem Rev 113:1598–1613
31. Viloria JS, Allega MF, Lambrughi M, Papaleo E (2017) An optimal distance cutoff for contact-based Protein Structure Networks using side chain center of masses. Sci Rep 7:2838. https://doi.org/10.1038/s41598-017-01498-6
32. Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25:163–177. https://doi.org/10.1080/0022250X.2001.9990249

Chapter 3

Identification of Allosteric Effects in Proteins by Elastic Network Models

Guang Hu

Abstract

Allostery is a fundamental regulatory mechanism in the majority of biological processes of molecular machines. Allostery is well-known as a dynamic-driven process, and thus, the molecular mechanism of allosteric signal transmission needs to be established. Elastic network models (ENMs) provide efficient methods for investigating the intrinsic dynamics and allosteric communication pathways in proteins. In this chapter, two ENM methods including Gaussian network model (GNM) coupled with Markovian stochastic model, as well as the anisotropic network model (ANM), were introduced to identify allosteric effects in hemoglobins. Techniques on model parameters, scripting and calculation, analysis, and visualization are shown step by step.

Key words Normal mode analysis, Global motion, Commute time, Allosteric site, Communication pathway

1 Introduction

One of the central goals in current biology is to understand the allosteric effects in molecular machines [1, 2]. Allostery is a fundamental process that regulates the function of proteins via a local perturbation of one site, such as ligand binding, mutations, or covalent modifications, which can induce a communication across the structure to another spatially distant site [3]. Identification of allosteric effects in proteins may help in efficient drug discovery and protein design [4]. After the first incorporation of the concept of allostery in describing the cooperative transition of hemoglobin [5], different models have been proposed to understand the molecular mechanism of allosteric regulations [6]. However, two fundamental aspects in allosteric effects are still needed to be uncovered. On the first hand, the identification of potential allosteric sites and how to quantify their allosteric ability remains an enigma. On the other hand, the mechanism that underlies distal communications

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_3,

pathways between allosteric sites and other active sites including catalytic sites or ligand-binding sites also needs to be established.

Computational approaches including both sequence and structure-based methods have been widely used to investigate the molecular basis of allosteric regulation and communication [7–11]. Sequence-based models can describe allosteric sites in terms of conservation and coevolution properties and enumerate potential communication pathways by constructing evolutionary networks [12]. Network-based structural studies (also called PSN) have also demonstrated that the topology and connectivity of protein structures provide a robust framework for understanding allosteric effects in terms of local and global graph-based parameters [13]. Recent evidence supports the view that allosteric communication is facilitated by the intrinsic dynamics of the biomolecules [14]. These methods, combined with molecular dynamics (MD) simulations, have been recently applied to elucidate allosteric communication pathways of diverse protein systems [15–19]. In the most recent studies, Verkhivker et al. [20, 21] have combined evolutionary analysis, MD simulations of the molecular chaperone of Hsp90 with the network analysis, and perturbation response scanning (PRS) approach to probe key sites in allosteric mechanisms.

An alternative dynamical approach, Elastic network model (ENM) [22], was first introduced by Tirion to study the intrinsic dynamics of proteins by using normal mode analysis. In ENMs, a protein structure was considered as a network consisting of a set of residues interconnected by elastic springs. Currently, two main types of ENMs are widely used, which are the Gaussian network model (GNM) [23] and the anisotropic network model (ANM) [24]. Based on ENMs, several theoretical approaches have been made toward understanding the molecular basis of allosteric communication. ENM combined with dynamics perturbation analysis (DPA) [25] and PRS [26, 27] were developed to detect the key residues whose perturbation couple structural dynamic changes at distal distance. By using ENM to calculate the correlations between the fluctuations in spring length of pairwise residues, a new method was also proposed to identify allosteric residues [28]. In addition, a thermodynamic method based on ENM was proposed to predict the allosteric sites on the protein surface [29]. ENM and PSN are two coarse-grain methods, and the integration of them may help to understand the pathways of communication between spatially distant sites [30]. The first attempt to combine ENM and PCN has been reported to investigate allosteric communication pathways in the PDZ2 domain [31]. The successful applications of combining ENM and PSN in the prediction of structural dynamics and allosteric sites have been proved in several protein systems [32] and bacterial ribosome [33].

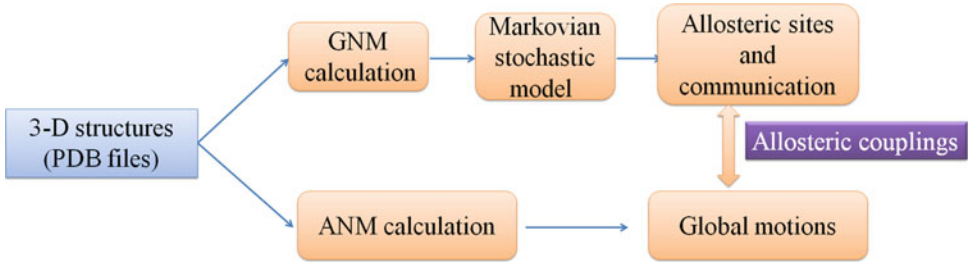

Fig. 1 A simplified representation showing the identification of allosteric effects in proteins by ENMs

It should be noted that introducing a Markovian process into coarse-grained models (such as GNM) has offered opportunities to assess signal propagation in proteins. In 2006 [34], Bahar's group proposed a novel GNM approach to elucidate allosteric communication pathways, by coupling Markov stochastic model based on information theory and spectral graph methods. In this framework, hitting and commute times were defined based on GNM fluctuations to measure the communication abilities of residues [35]. The relationship between allosteric communication pathways and the intrinsic structural dynamics of proteins was described by their coupling and provides a new avenue for further examination of protein allostery from local dynamical changes to global motions [36]. Applying this method has found that metal-binding sites are designed to achieve allosteric signaling properties [37]. In our recent work, we have shown the ability of using GNM works with Markovian model to study DNMT3A and highlighted the key role of dimer interface in allosteric communication [38].

In this chapter, two types of ENM methods were introduced to study the allosteric effects of a typical allosteric protein, hemoglobin (Hbs). Theoretical basics of GNM-Markov model and ANM calculation are outlined briefly. We showed techniques and computer scripting to elucidate the overall protein structure by using a combination of GNM and Markov stochastic modeling. Furthermore, the most cooperative modes of motions are predicted by ANM calculations. The functional coupling between global dynamics and signal transduction pathways could give more insight of mechanical mechanism of allostery. A simplified representation of the methods used in the chapter is shown in Fig. 1.

2 Materials

1. The structures of two Hbs [39] were downloaded from protein data back (https://www.rcsb.org/). Hbs with two quaternary conformations are used: T-Hb (PDB code: 2dn2) and the high-affinity state R-Hb (PDB code: 2dn1). Both structures are composed of four subunits: α_1 and α_2 subunits of

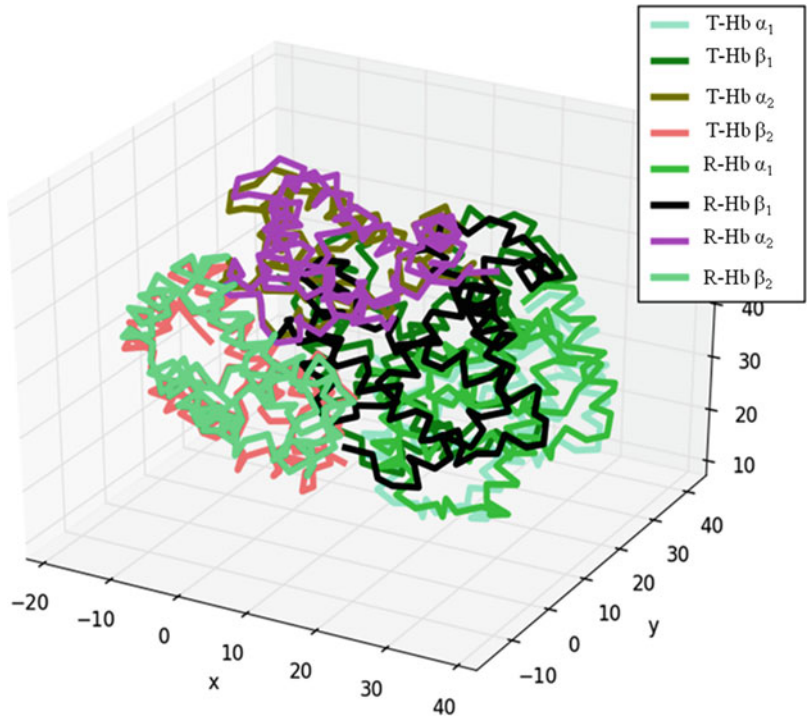

Fig. 2 The structure alignment of Hbs with T and R states

140 residues, β_1 and β_2 subunits of 145 residues. The structural alignment (Fig. 2) of two states shows the difference caused by the torsional rotation of $\alpha_1\beta_1$ dimer and $\alpha_2\beta_2$ dimer, with RMSD of 2.41 Å.

2. Both GNM and ANM calculations were performed using ProDy [40], which is a free and open-source Python package for protein structural dynamics analysis (http://prody.csb.pitt.edu/). The installation of ProDy may require latest versions of Biopython, NumPy, and Matplotlib. *See* **Note 1** for more tools.
3. VMD 1.91 or the newer releases (http://www.ks.uiuc.edu/Research/vmd/) are required for visualization and rendering. Normal Mode Wizard (NMWiz) is a VMD plugin, which is developed to visualize GNM modes and ANM motions.
4. Matlab (http://www.mathworks.com/products/matlab/) was used to calculate to identify shortest path between two sites in proteins. Two Matlab scripts used in this section can be downloaded at http://sysbio.suda.edu.cn/pdbgraph/.

3 Methods

Normal mode analysis has become a gold standard for studying protein dynamics and allosteric regulation of protein systems. Here, two ENMs are employed to perform normal mode analysis. *See* **Note 2** and **Note 3** for their main advantages and limitations.

3.1 Predicting Hinge Sites by GNM

3.1.1 Theory of GNM

GNM [41] describes a protein as a network of C_α connected by springs of uniform force constant γ if they are located within a cutoff distance r_c. In GNM, the interaction potential for a protein of N residues is

$$V_{\mathrm{GNM}} = -\frac{\gamma}{2}\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(R_{ij}-R_{ij}^{0}\right)\cdot\left(R_{ij}-R_{ij}^{0}\right)\Gamma_{ij}\right] \tag{1}$$

where R_{ij}^0 and R_{ij} are the equilibrium and instantaneous distance between residues i and j, and Γ is the $N \times N$ Kirchhoff matrix, which is written as:

$$\Gamma_{ij} = \begin{cases} -1 & i \neq j, R_{ij} \leq r_c \\ 0 & i \neq j, R_{ij} \geq r_c \\ -\sum\limits_{i,i\neq j}\Gamma_{ij} & i = j \end{cases}. \tag{2}$$

Then, square fluctuations are given by.

$$\left\langle (\Delta R_i)^2 \right\rangle = \left(3kT/\gamma\right)\tilde{n}\left[\Gamma^{-1}\right]_{ii} \text{ and } \left\langle \Delta R_i \tilde{n} \Delta R_j \right\rangle = \left(3kT/\gamma\right)\tilde{n}\left[\Gamma^{-1}\right]_{ij} \tag{3}$$

The normal modes are extracted by eigenvalue decomposition: $\Gamma = U\Lambda U^T$, U being the orthogonal matrix whose k^{th} column u_k is the k^{th} mode eigenvector. Λ is the diagonal matrix of eigenvalues, λ_k. Hinge sites for a protein are defined by GNM at the fluctuation minima of the lowest modes. Hinge sites not only correspond to key residues for maintaining collective behaviors of proteins, but also have been proved to be implicated in allosteric mechanisms [42].

3.1.2 GNM Calculation

GNM calculations are carried out with ProDy. Using T-Hb as an example, the calculation steps are listed as follows:

1. Import of all related content from ProDy:

```
$from prody import *
$from pylab import *
$from numpy import *
$ion ()
```

2. Defining the *T*-Hb structure by parsing the PDB file with only Ca atoms:

```
$THb = parsePDB ('2dn2.pdb', subset = 'calpha')
```

3. Defining the class of GNM analysis:

```
$gnm_T=GNM ('T-Hb')
```

4. Construction of Kirchhoff matrix of atomic coordinates:

```
$gnm_T.buildKirchhoff (T-Hb)
```

5. Calculation of GNM modes (20 modes by default) by diagonalization of Kirchhoff matrix:

```
$gnm_T.calcModes ()
```

6. Calculation of square fluctuations for the first and the second GNM modes to identify hinge sites.

```
$sq_1=calcSqFlucts (gnm_T [0])
$sq_2=calcSqFlucts (gnm_T [1])
```

7. Saving square fluctuations for the first and the second GNM modes.

```
$np.savetxt ('sq_1.txt', sq_1)
$np.savetxt ('sq_2.txt', sq_2)
```

The square fluctuations of the T-Hb based on the first and second GNM modes are shown in Fig. 3a. The hinge residues for mode 1 are distributed in the α_1-β_2 interface, including Thr38, Thr41, Leu91, Val93, Pro95, and Ser 138 in α_1 (the blue line), while for mode 2 are distributed in the α_1-β_1 interface, including Thr118, Ala120, Ala123 in α_1, and His112, Phe118 in β_1 (the red line). Based on the first and second GNM modes, similar protocol can be performed to obtain fluctuations of the R-Hb, which are shown in Fig. 3c. The hinge residues for mode 1 (the blue line) are not only distributed in the α_1-β_2 interface including Val93, Asp94, Pro95, Thr137, Ser138 in α_1, but also the α_2-β_1 interface including Asp94, His97in β_1, while for mode 2 (the red line) are also distributed in the α_1-β_1 interface, including Phe117, Val121 in α1, and Ala 115 in β1. The distributions of these residues predicted by the first GNM mode (green beads) and the second GNM mode (yellow beads) in the three-dimensional structures of T and R-Hbs are displayed in Fig. 3b, d, respectively.

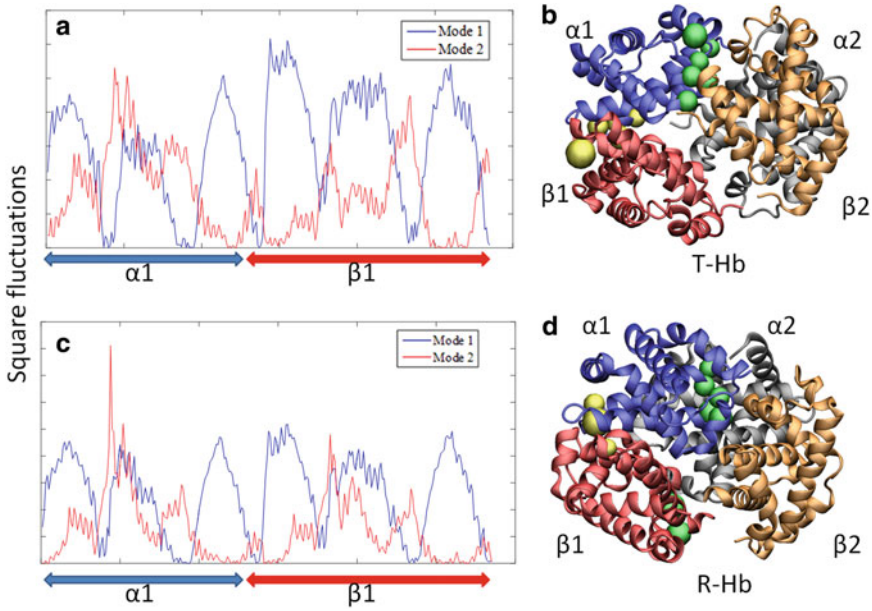

Fig. 3 Results for predicted hinge site in Hbs. (**a**) The minima of square fluctuations based on the two lowest GNM modes predict hinges that are located at different interfaces in T-Hb, and (**b**) green and yellow beads denote hinges predicted by the first and the second GNM modes, respectively. Similar results (**c**, **d**) were found for R-Hb

3.2 Identifying Key Sites for Communications

3.2.1 Markov Stochastic Model

The Markov stochastic model coupled with GNM was used for exploring the signal transductions of perturbations in proteins. The affinity matrix A describes the interactions between residue pairs connected in GNM; its generic element a_{ij} is defined as:

$$a_{ij} = \frac{N_{ij}}{\sqrt{N_i N_j}} \tag{4}$$

where N_{ij} is the number of atom-atom contacts between residues i and j based on a cutoff distance of 4 Å, N_i is the number of side chain atoms in residue i. The density of contacts at each node i is given by:

$$d_i = \sum_{j=1}^{N} a_{ij} \tag{5}$$

The Markov transition matrix M, whose element $m_{ij} = d_j^{-1} a_{ij}$ determines the conditional probability of transmitting a signal from residue j to residue i in one time step. Accordingly, the hitting time for the transfer of a signal from residue j to i is given by

$$H(i,j) = \sum_{k=1}^{N} \left\{ \left[\Gamma^{-1}\right]_{kj} - \left[\Gamma^{-1}\right]_{ij} - \left[\Gamma^{-1}\right]_{ki} - \left[\Gamma^{-1}\right]_{ii} \right\} \tilde{n} d_k \tag{6}$$

where Γ is Kirchoff matrix obtained by GNM. The average hit time for the i-th residue $<H(i)>$ is the average of $H(i, j)$ over all starting points i. The commute time is defined by the sum of the hitting times in both directions, that is:

$$C(i,j) = H(i,j) + H(j,i) \tag{7}$$

Accordingly, commute times provide a metric of the efficiency of allosteric communication.

3.2.2 Calculations of Hitting Time and Commute Time

The commute time enables us to identify residues that are more sensitive to allosteric communication across the protein. The application of Markov stochastic model to Hbs for computing commute time includes the following commands:

1. Import the program for computing the hitting times and commute times into ProDy:

```
$from IT_HitCommute import *
```

2. Write the Kirchhoff matrix from GNM.

```
$K_T=gnm_T.getKirchhoff ()
```

3. Pass the Kirchhoff matrix to hit/commute time.

```
$hc_T=IT_HitCommute (K_T)
```

4. Calculation and save of hit time matrix.

```
$H_T=hc_T.buildHitTimes (K_T)
$np.savetxt ('H_T.txt', H_T)
```

5. Calculation and save of commute time matrix.

```
$C_T=hc_T.buildCommuteTimes ()
$np.savetxt ('C_T.txt', C_T)
```

The commute time map of ?-Hb is displayed in Fig. 4a, where the blue and red regions correspond to short and long commute time. The average values of each row or column of the commute time map were also calculated to evaluate the communication abilities of each residue. In addition, the minima of the average commute time indicate the key residues allostery in Hbs. As shown in Fig. 4b, the profiles of average commute times for α1 chain in both T- and R-Hbs show that Val10, Leu29, Arg31, Thr39, Cys104, Val107, His122, and Leu125 in α1 chain are residues with highest communication abilities. Figure 4c predicts that

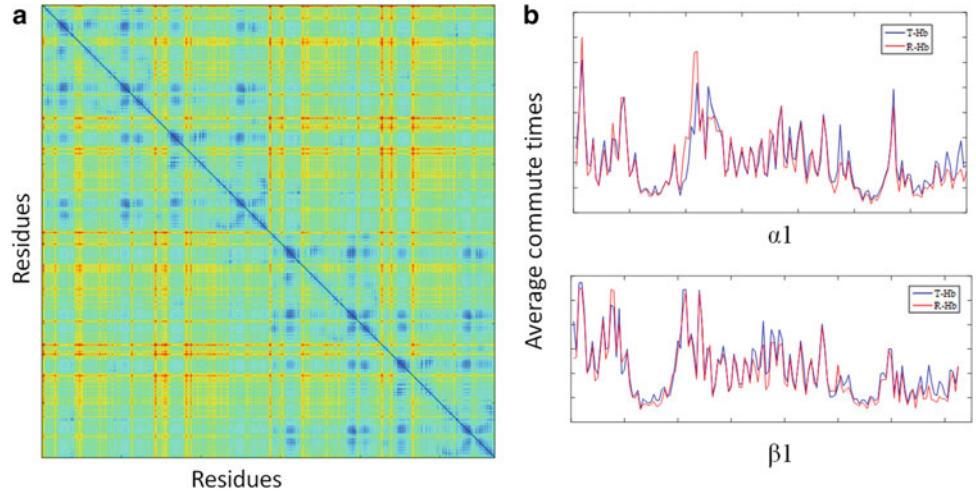

Fig. 4 Allosteric properties for Hbs. (**a**) The hitting time map for *T*-Hb. (**b**) The average commute time profiles for α_1 chain in *T*- (blue line) and *R*-Hbs (red line). (**c**) The average commute time profiles for β_1 chain in *T*- and *R*-Hbs

Ala27, Val109, Cys112, and Gln127 in **β**1 chain are potential allosteric sites in both states. It is worth noting that the two α chains and two β chains have the same profile shapes. The distributions of these residues showed that Arg31, Cys104, Val107, and His122 in α1 chain and Cys112 and Gln127 in β1 chain are located at α_1-β_1 interface. Likely, the same region was also found at α_2-β_2 interface.

3.3 The Identification of Communication Pathways

3.3.1 The Determination of Start and End Points

For the identification of communication pathways, the first step is the determination of start and end points of these pathways. T-Hb is the unliganded form, while R is O_2–bound form, with binding sites located at Val62, His58, and His87 in α chain and Val67, His63, and His92 in β chain. In order to investigate the allosteric communication pathways induced by O_2 binding in the R state, the signal transmission between ligand-binding sites and allosteric residues located at interfaces in R-Hb were considered. Thus, the binding sites in R-Hb were chosen as start points, and the allosteric sites with low average commute time at interface were chosen as end points. For the particular case studies, the Val62 in α_1 chain and His 92 in β_1 chain were set as start points, while Arg31 and Asp94 in α_1 chain, as well as Ala 115 and Val33 in β_1 chain located at $\alpha_1\beta_1$ interface, were set as end points.

3.3.2 Shortest Pathway Calculation

For the communication pathways calculation, a protein was transformed into a graph whose topology is decided by Kirchhoff matrix, with each edge weighted by the commute time $C(i, j)$. Then, Dijkstra's algorithm was used to find the shortest pathway between nodes in the graph. Shortest pathway calculation is

implemented in Matlab. To follow this tutorial, you will need the following execute files:

Cafrompdb.m: this is a function for reading a protein structure from PDB file.

Graph.m: this is a function for transforming a protein structure into a graph.

1. Start Matlab.
2. Copy cafrompdb.m and to Graph.m the work directory, and use Cafrompdb.m and Graph.m to construct graphs from PDB files, and type the flowing command lines in the Matlab workspace.

```
$[G, A, C]=Graph(pdb_file, C, cutoff, pdb_path)
$ G = sparse (G);
```

G is a connected graph generated from the adjacent matrix (G), in which each edge is weighted by commute time (C_{ij}).

3. Calculate shortest pathway using the Matlab built-in functions.

```
$[dist, path, pred] = graphshortestpath(G, S, T)
```

For example, the pathway between Val62 in α1 chain to Asp94 α1 chain located at $\alpha_1\beta_1$ interface can be calculated as:

```
Input: [G,A,C]=Graph ('2dn1.pdb', 'C_R.txt', 7, './')
G=sparse(G);
 [dist, path]=graphshortestpath (G, 62, 94)

Output: dist=5.6316e+03
 path=62 25 28,104,101 97 94
```

Table 1 displays four examples of such pathways in R-Hb, starting from Val62 in α_1 chain and His 92 in β_1 chain, to Arg31 and Asp94 in α_1 chain, and Ala115 and Val33 in $\boldsymbol{\beta}_1$ chain located at $\alpha_1\beta_1$ interface.

Table 1
Proposed paths of communication between ligand-binding and allosteric sites within the monomer in R-Hbs

α_1	Val62 → Gly25 → Ala28 → Arg31
α_1	Val62 → Gly25 → Ala28 → Cys104 → Leu101 → Asn97 → Asp94
β_1	His92 → Val98 → Pro100 → Arg104 → Asn108 → Cys112 → Ala115
β_1	His 92 → Val98 → Pro100 → Phe103 → Leu106 → Leu31 → Val 33

3.4 Global Motions

In the previous ENM study [43], the allosteric communication effects are often well-described by low-frequency modes that identify most cooperative motions. In this section, ANM method was used to investigate global motions of T- and R-Hbs, and their conformational change, toward gaining a mechanistic understanding of the allosteric couplings.

3.4.1 Theory of ANM

In ANM [44], the interaction potential for a protein of N residues is

$$V_{\mathrm{ANM}} = \frac{\gamma}{2}\sum_{i,j}^{N}\left(\left|R_{ij}\right| - \left|R_{ij}^{0}\right|\right)^{2} \tag{8}$$

The motion of the ANM mode of proteins is determined by the 3 $N \times 3$ N *Hessian matrix H*. The eneric element is given as:

$$H_{ij} = \begin{bmatrix} \frac{\partial^2 V}{\partial X_i \partial X_j} & \frac{\partial^2 V}{\partial X_i \partial \Upsilon_j} & \frac{\partial^2 V}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V}{\partial \Upsilon_i \partial X_j} & \frac{\partial^2 V}{\partial \Upsilon_i \partial \Upsilon_j} & \frac{\partial^2 V}{\partial \Upsilon_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Z_i \partial X_j} & \frac{\partial^2 V}{\partial Z_i \partial \Upsilon_j} & \frac{\partial^2 V}{\partial Z_i \partial Z_j} \end{bmatrix} \tag{9}$$

where X_i, Υ_i, and Z_i represent the Cartesian components of residues i, V is the potential energy of the system. r_c used here is 13 Å. Accordingly, ANMs not only provide the information about the amplitudes, but also about the direction of residue fluctuations.

The similarity between two ANM modes, u_k and v_l, evaluated for proteins with two different conformations can be quantified in terms of inner product of their eigenvectors, i.e.:

$$O(u_k, v_l) = u_k \tilde{n} v_l \tag{10}$$

The degree of overlap between k^{th} ANM modes u_k and the experimentally observed conformation change Δr of Hbs among different states is quantified by $\left(\Delta r \tilde{n} u_k / |\Delta r|\right)$. Therefore, the cumulative overlap *CO(m)* between Δr and the directions spanned by a subsets of m ANM modes is calculated as:

$$CO(m) = \sqrt{\sum_{k=1}^{m}\left(\Delta r \tilde{n}\frac{u_k}{|\Delta r|}\right)^{2}} \tag{11}$$

3.4.2 NM Calculation

The ANM calculation was also performed in ProDy, as follows:

1. Defining the class for ANM analysis for T- and R-Hbs.

```
$anm_T, T-Hb=calcANM (T-Hb)
$anm_R, R-Hb=calcANM (R-Hb)
```

2. Saving the ANM model for visualization in VMD and NMWiz:

```
$writeNMD ('ANM_T.nmd', anm_T, T-Hb)
$writeNMD ('ANM_R.nmd', anm_R, R-Hb)
```

3. Calculation of the subspace overlap between the first 5 ANM modes of T- and R-Hbs.

```
$calcOverlap (anm_T [:5], anm_R [:5]
$showOverlapTable (anm_T[:5], anm_R[:5])
```

The overlap map between the five ANM slowest modes (eigenvectors) was calculated to compare the global dynamics of *T*- and *R*-Hbs (Fig. 5a). The upper limit of 1 indicates perfect overlap (red), and 0 indicates no overlap (blue). It can be see that the reordering of the first two modes was found, which means that the motion of the first mode of *T*-Hb corresponds to the motion of the second mode of *R*-Hb (overlap of 0.92), while the first mode of *R*-Hb shifts to the second mode of *T*-Hb (overlap of 0.90).

4. Visualization of ANM results in NMD format by using VMD plugin Normal Mode Wizard.

```
$Start VMD
$Select Extensions→Analysis→Normal Mode Wizard
$Select 'Load NMD File'
```

Two ANM modes are:

ANM Mode 1.

The first ANM mode of the *T*-Hbs (also the second ANM mode of the *R*-Hb) shows the tong-like motion (Fig. 5b), which is

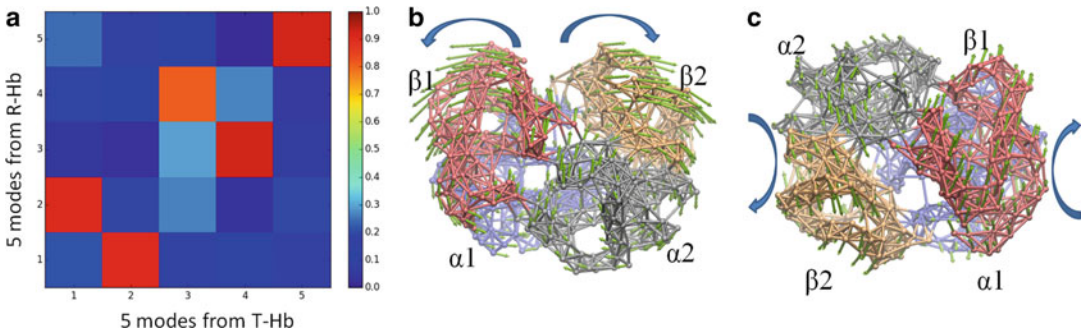

Fig. 5 ANM results for Hbs. (**a**) Overlaps between the five slowest ANM modes of T- and R-Hbs. (**b**) The tong-like motion, corresponding to the first mode of T-Hb or the second mode of R-Hb. (**c**) The hinge-binding rotation, corresponding to the second mode of T-Hb or the first mode of R-Hb. The figures are rendered by NMWiz in VMD, and arrows indicate the opening and the rotation directions

consistent with the opening motion of β_1 and β_2 in opposite directions, while two α chains are relatively stable, mainly in a rigid-body-type motion. This observation is consistent with the pervious molecular dynamics study, which has shown that β chains are more strongly linked to the quaternary transition than α chains [45].

ANM Mode 2.

The second ANM mode of the *T*-Hbs (also the first ANM mode of the *R*-Hb) shows another global motion involving quaternary changes of two dimers (Fig. 5c), namely, $\alpha_1\beta_1$ dimer, exhibiting a torsional rotation in an opposite direction with $\alpha_2\beta_2$ dimer, coordinated by the hinges at the α_1-β_2 and β_1-α_2 interfaces. This hinge-bending rotation defines the intrinsic dynamics of Hbs, which may facilitate their allosteric communication pathways via hinge regions.

4 Notes

1. Main advantages of ENMs lie at its simplicity, whose modes at low frequencies are enough to capture intrinsic dynamics and allosteric properties of biomolecular systems. Thus, ENM methods are applicable to large systems, as well as for high-throughput investigation of protein data.
2. There are some limitations in ENMs since ENMs only consider Cα atoms without considering specific interactions. It can be known that allostery should take into account the roles of flexible regions such as loops. ENMs may not be suitable for modeling the allosteric effects of these flexible regions.
3. Additionally, there are some other normal mode analysis-based web servers to predict allosteric sites and signal propagation pathways in proteins and their complexes, such as PARS [46], SPACER [47], AlloPred [48], and DynOmics [49] servers.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31872723) and a Project Funded by the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions.

The author thanks Prof. Ivet Bahar for giving the opportunity to study Elastic network models and ProDy in her lab. The author also thanks Drs. Hongchun Li and Chakra Chennubhotla for providing programming codes.

References

1. Lorimer GH, Horovitz A, McLeish T (2018) Allostery and molecular machines. Philos Trans R Soc Lond Ser B Biol Sci 373 (1749):20170173
2. Nussinov R (2016) Introduction to protein ensembles and allostery. Chem Rev 116 (11):6263–6266
3. Ribeiro AA, Ortiz V (2016) A chemical perspective on allostery. Chem Rev 116 (11):6488–6502
4. Wagner JR et al (2016) Emerging computational methods for the rational discovery of allosteric drugs. Chem Rev 116 (11):6370–6390
5. Pauling L (1935) The oxygen equilibrium of hemoglobin and its structural interpretation. Proc Natl Acad Sci U S A 21(4):186–191
6. Liu J, Nussinov R (2016) Allostery: an overview of its history, concepts, methods, and applications. PLoS Comput Biol 12(6): e1004966
7. Collier G, Ortiz V (2013) Emerging computational approaches for the study of protein allostery. Arch Biochem Biophys 538(1):6–15
8. Feher VA et al (2014) Computational approaches to mapping allosteric pathways. Curr Opin Struct Biol 25:98–103
9. Verkhivker GM (2018) Computational modeling of the Hsp90 interactions with cochaperones and small-molecule inhibitors. Methods Mol Biol 1709:253–273
10. Greener JG, Sternberg MJ (2017) Structure-based prediction of protein allostery. Curr Opin Struct Biol 50:1–8
11. Schueler-Furman O, Wodak SJ (2016) Computational approaches to investigating allostery. Curr Opin Struct Biol 41:159–171
12. Suel GM et al (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 10 (1):59–69
13. Di Paola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48
14. Gunasekaran K, Ma B, Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? Proteins 57(3):433–443
15. Blacklock K, Verkhivker GM (2014) Allosteric regulation of the Hsp90 dynamics and stability by client recruiter cochaperones: protein structure network modeling. PLoS One 9(1): e86547
16. Verkhivker GM (2017) Network-based modelling and percolation analysis of conformational dynamics and activation in the CDK2 and CDK4 proteins: dynamic and energetic polarization of the kinase lobes may determine divergence of the regulatory mechanisms. Mol BioSyst 13(11):2235–2253
17. Stetz G, Verkhivker GM (2017) Computational analysis of residue interaction networks and coevolutionary relationships in the Hsp70 chaperones: a community-hopping model of allosteric regulation and communication. PLoS Comput Biol 13(1):e1005299
18. Zhou R et al (2015) Molecular mechanism underlying PRMT1 dimerization for SAM binding and methylase activity. J Chem Inf Model 55(12):2623–2632
19. Stetz G, Tse A, Verkhivker GM (2017) Ensemble-based modeling and rigidity decomposition of allosteric interaction networks and communication pathways in cyclin-dependent kinases: differentiating kinase clients of the Hsp90-Cdc37 chaperone. PLoS One 12(11): e0186089
20. Stetz G, Tse A, Verkhivker GM (2018) Dissecting structure-encoded determinants of allosteric cross-talk between post-translational modification sites in the Hsp90 chaperones. Sci Rep 8(1):6899
21. Verkhivker GM (2018) Dynamics-based community analysis and perturbation response scanning of allosteric interaction networks in the TRAP1 chaperone structures dissect molecular linkage between conformational asymmetry and sequential ATP hydrolysis. Biochim Biophys Acta 1866(8):899–912
22. Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 77 (9):1905–1908
23. Li H et al (2016) iGNM 2.0: the Gaussian network model database for biomolecular structural dynamics. Nucleic Acids Res 44 (D1):D415–D422
24. Eyal E, Lum G, Bahar I (2015) The anisotropic network model web server at 2015 (ANM 2.0). Bioinformatics 31(9):1487–1489
25. Ming D, Wall ME (2005) Quantifying allosteric effects in proteins. Proteins 59(4):697–707
26. Atilgan C, Atilgan AR (2009) Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. PLoS Comput Biol 5(10):e1000544
27. Zheng W, Brooks BR, Thirumalai D (2009) Allosteric transitions in biological nanomachines are described by robust normal modes

of elastic networks. Curr Protein Pept Sci 10 (2):128–132

28. Erman B (2013) A fast approximate method of identifying paths of allosteric communication in proteins. Proteins 81(7):1097–1101
29. Su JG et al (2014) Prediction of allosteric sites on protein surfaces with an elastic-network-model-based thermodynamic method. Phys Rev E Stat Nonlinear Soft Matter Phys 90 (2):022719
30. Hu G et al (2017) Comparative study of elastic network model and protein contact network for protein complexes: the hemoglobin case. Biomed Res Int 2017:2483264
31. Raimondi F et al (2013) A mixed protein structure network and elastic network model approach to predict the structural communication in biomolecular systems: the PDZ2 domain from tyrosine phosphatase 1E as a case study. J Chem Theory Comput 9 (5):2504–2518
32. Yao XQ, Skjaerven L, Grant BJ (2016) Rapid characterization of allosteric networks with ensemble normal mode analysis. J Phys Chem B 120(33):8276–8288
33. Guzel P, Kurkcuoglu O (2017) Identification of potential allosteric communication pathways between functional sites of the bacterial ribosome by graph and elastic network models. Biochim Biophys Acta 1861(12):3131–3141
34. Chennubhotla C, Bahar I (2006) Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES. Mol Syst Biol 2:36
35. Chennubhotla C, Bahar I (2007) Signal propagation in proteins and relation to equilibrium fluctuations. PLoS Comput Biol 3 (9):1716–1726
36. Chennubhotla C, Yang Z, Bahar I (2008) Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. Mol BioSyst 4 (4):287–292
37. Dutta A, Bahar I (2010) Metal-binding sites are designed to achieve optimal mechanical and signaling properties. Structure 18 (9):1140–1148
38. Liang Z et al (2018) Deciphering the role of dimer interface in intrinsic dynamics and allosteric pathways underlying the functional transformation of DNMT3A. Biochim Biophys Acta 1862(7):1667–1679
39. Park SY et al (2006) 1.25 A resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. J Mol Biol 360(3):690–701
40. Bakan A, Meireles LM, Bahar I (2011) ProDy: protein dynamics inferred from theory and experiments. Bioinformatics 27 (11):1575–1577
41. Bahar I, Atilgan AR, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 2(3):173–181
42. Sumbul F, Acuner-Ozbabacan SE, Haliloglu T (2015) Allosteric dynamic control of binding. Biophys J 109(6):1190–1201
43. Rodgers TL et al (2013) Modulation of global low-frequency motions underlies allosteric regulation: demonstration in CRP/FNR family transcription factors. PLoS Biol 11(9): e1001651
44. Atilgan AR et al (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80(1):505–515
45. Hub JS, Kubitzki MB, de Groot BL (2010) Spontaneous quaternary and tertiary T-R transitions of human hemoglobin in molecular dynamics simulation. PLoS Comput Biol 6 (5):e1000774
46. Panjkovich A, Daura X (2012) Exploiting protein flexibility to predict the location of allosteric sites. BMC Bioinformatics 13:273
47. Goncearenco A et al (2013) SPACER: Server for predicting allosteric communication and effects of regulation. Nucleic Acids Res 41 (Web Server issue):W266–W272
48. Greener JG, Sternberg MJ (2015) AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. BMC Bioinformatics 16:335
49. Li H et al (2017) DynOmics: dynamics of structural proteome and beyond. Nucleic Acids Res 45(W1):W374–W380

Chapter 4

Locating and Navigating Energy Transport Networks in Proteins

Korey M. Reid and David M. Leitner

Abstract

We review computational methods to locate energy transport networks in proteins that are based on the calculation of local energy diffusion in nanoscale systems. As an illustrative example, we discuss energy transport networks computed for the homodimeric hemoglobin from *Scapharca inaequivalvis*, where channels for facile energy transport, which include the cluster of water molecules at the interface of the globules, have been found to lie along pathways that experiments reveal are important in allosteric processes. We also review recent work on master equation simulations to model energy transport dynamics, including efforts to relate rate constants in the master equation to protein structural dynamics. Results for apomyoglobin involving relations between fluctuations in the length of hydrogen bonds and the energy flux between them are presented.

Key words Energy transport networks, Allostery, Water cluster, Nonbonded networks

1 Introduction

Progress in locating energy transport pathways in proteins both computationally and experimentally has been proceeding at a rapid pace [1]. Time-resolved IR and Raman techniques, e.g., have provided detailed pictures of the nature and rate of energy transport in peptides and proteins [2–7], including recent advances in identifying transport through individual amino acids of several heme proteins [8–10]. Energy transport pathways have since some time been identified by molecular simulations [11, 12], with recent focus on the development of coarse-graining approaches [13–25], some of which have exploited analogies to thermal transport in other molecular materials [26, 27]. Network analysis has been applied to facilitate identification of pathways and residues that control protein dynamics [28–51], where a variety of definitions of a network have been adopted, including those that incorporate distance, conformational fluctuations, and energy criteria. The energy transport channels of a protein form a network, and analysis

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_4,

of the energy transport network reveals how a protein responds to local structural changes, possibly pointing to pathways along which allosteric transitions occur. In this chapter we review an approach to calculate energy transport networks in proteins that is based on calculation of local thermal transport in nanoscale materials. We illustrate the method with the example of a homodimeric hemoglobin, where prominent energy transport channels were found to lie along pathways important in allostery. In addition to locating energy transport networks, the calculated local energy diffusion coefficients can be used to model energy transport by master equation simulations, as we review here. We also examine the possibility of relating the rate constants to dynamic fluctuations of hydrogen bonds, where we present calculations exploring such a connection in apomyoglobin.

That energy transport pathways exist, i.e., energy does not simply flow isotropically through a globular protein, is an inherent property of the geometry of a folded protein [52–60], which resembles that of a percolation cluster at threshold. Some channels are relatively long range, which may contribute to function such as allostery [41, 51, 61–65], by which proteins regulate reactions that occur in remote regions of the molecule [66–69]. In efforts to elucidate protein dynamics, strategies have been adopted to identify pathways or ensembles of pathways [64, 65, 70–75] along which transitions between different states of the protein occur. Whether or not vibrational energy transport channels point to pathways involved in allosteric transitions, energy relaxation pathways regulate chemical reaction dynamics. Optical studies of energy relaxation in myoglobin have since some time produced a detailed picture of events that follow excitation of the heme and ligand photolysis, elucidating chemical dynamics of that protein [76, 77]. However, the extent to which the relaxation pathways identified in myoglobin play a role in allostery in hemoglobins remains unclear, due to the diversity of orientations of the monomeric units of different hemoglobins [78]. It would thus be desirable to identify energy transport in individual hemoglobins to examine the extent to which they overlap pathways or ensembles of pathways along which allosteric transitions take place.

Towards this goal, and as an illustrative example of the energy transport networks that can be computed for a protein, we summarize recent computational work identifying networks of energy transport channels in the allosteric homodimeric hemoglobin from *Scapharca inaequivalvis*, HbI [79]. When HbI is in the unliganded state the crystallographic structure reveals a cluster of 17 water molecules at the interface between the two globules, whereas 11 are found in the liganded state. The free energy of ligand binding in HbI and the origin of cooperativity is mainly entropic [80, 81]. Ligand-linked tertiary structural changes occur upon ligand binding, including rotation of Phe97 into the interface

between the globules, which is otherwise tightly packed against the proximal histidine, His101, in the unliganded structure. Cooperativity depends on a number of residues at the interface in contact with the tightly bound [82] cluster of water molecules [83–85], and the Lys30-Asp89 salt bridge [86], which is farther from the water cluster but crucial to the stability of the homodimer. Crystal structures reveal differences between the hydrogen bonding arrangement of the waters and side chains at the interface of the unliganded and liganded states [87]. Modification of this arrangement by point mutation apparently influences cooperativity [83, 85]. Overall the ligand-linked changes are mainly tertiary in HbI. Quaternary changes that take place, among the last steps [88], are much smaller than those in tetrameric human hemoglobin.

Having identified networks along which energy transport occurs we model the flow of energy by master equation simulations. Rate constants in the master equation can be obtained from local energy diffusion coefficients or more directly by fitting results of all-atom simulations of energy flow to a master equation. We discuss here a recent comparison between results of a master equation simulation and results of all-atom nonequilibrium simulations of energy flow in the villin headpiece subdomain HP36, where the rate constants used in the master equation simulations were related to the energy diffusion coefficients obtained by coarse-graining thermal transport in the protein [17]. A more recent study on HP36 by Stock and coworkers suggested that the rate constants in the master equation simulations are related to dynamic fluctuations of the protein [23]. We consider that possibility here for the hydrogen bonds of apomyoglobin. The hydrogen bond dynamics is analyzed from results of molecular dynamics (MD) simulations, and the rate constants for energy transfer along the bond are obtained using the same trajectory in a calculation of the local energy current, using the methodology developed by Yamato and coworkers [13, 14].

In the following section we summarize a coarse-graining approach to locate energy transport channels in proteins and discuss one application to the identification of energy transport networks in the homodimeric hemoglobin, HbI. We then review results comparing the dynamics along protein energy transport networks obtained by master equation simulations using rate constants obtained from a communication map with results of all-atom simulations of the villin headpiece subdomain, HP36. Possible scaling relations between rate constants in a master equation for hydrogen bonds in a protein and the dynamics of that hydrogen bond are then discussed. Concluding remarks are given in Subheading 3.

2 Computational Methods

2.1 Communication Maps: Locating Energy Transport Networks

To identify pathways along which energy transport is facile one can carry out nonequilibrium simulations, which have been done extensively by all-atom MD simulations [1, 12, 89]. Simulations in harmonic approximation via wave packet propagation [57, 58] have also been run to examine harmonic and anharmonic contributions to energy transport. The latter approach was used, e.g., to calculate vibrational energy transport in HbI. Those results are plotted in Fig. 1 for the case where energy is first introduced to one of the hemes and the flow of energy calculated over the next few picoseconds. Anisotropic transport is observed; the cluster of water molecules at the interface apparently serves as an energy transport channel from one globule to the other.

More coarse-grained approaches have also been introduced, including the local conductivity analysis of Yamato and coworkers [13, 14], which we shall adopt below to examine a relation between hydrogen bond dynamics and energy flow between two hydrogen bonded residues. An alternative coarse-graining method yields a network weighted by local energy diffusion coefficients calculated in terms of normal modes [19]. The weights for the network are expressed in terms of the matrix elements of the energy current operator, **S**, which in harmonic approximation can be written in terms of the Hessian matrix, **H**, and eigenmodes, **e**, of the object [90]. The mode diffusivity, in turn, can be expressed in terms of the matrix elements of **S** [90]. We break each matrix element up into contributions from individual residues. The contribution to the energy flux between residues A and A' to matrix element $S_{\alpha\beta}$ is [19]

$$S_{\alpha\beta}^{\{AA\prime\}} = \frac{i\hbar\left(\omega_\alpha + \omega_\beta\right)}{4V\sqrt{\omega_\alpha\omega_\beta}} \sum_{r,r'\in(x,y,z)} \sum_{l,l'\in AA'} e_l^\alpha H_{rr'}^{ll'}(\mathbf{R}_l - \mathbf{R}_{l'}) e_{l'}^\beta, \quad (1)$$

where $\mathbf{R}_l$ is the position of atom l and r is a coordinate (x, y, or z). We sum the atoms l together in a given region, A, and sum atoms l' together in region A'. V is the volume of the space spanned by the two regions. While such a volume remains somewhat ambiguous, it cancels out in the definition of the local energy diffusivity, Eq. 2.

For mode α the energy diffusivity is a sum over the squares of matrix elements of the heat current operator, i.e., $D_\alpha \propto \sum_{\beta\neq\alpha} \left|S_{\alpha\beta}\right|^2 \delta\left(\omega_\alpha - \omega_\beta\right)$. Considering only energy flow between residues A and A', we approximate the local energy diffusivity in mode α using the harmonic model as

$$D_\alpha^{\{AA\prime\}} = \frac{\pi V^2}{3\hbar^2\omega_\alpha^2} \sum_{\beta\neq\alpha} \left|S_{\alpha\beta}^{\{AA\prime\}}\right|^2 \delta\left(\omega_\alpha - \omega_\beta\right). \quad (2)$$

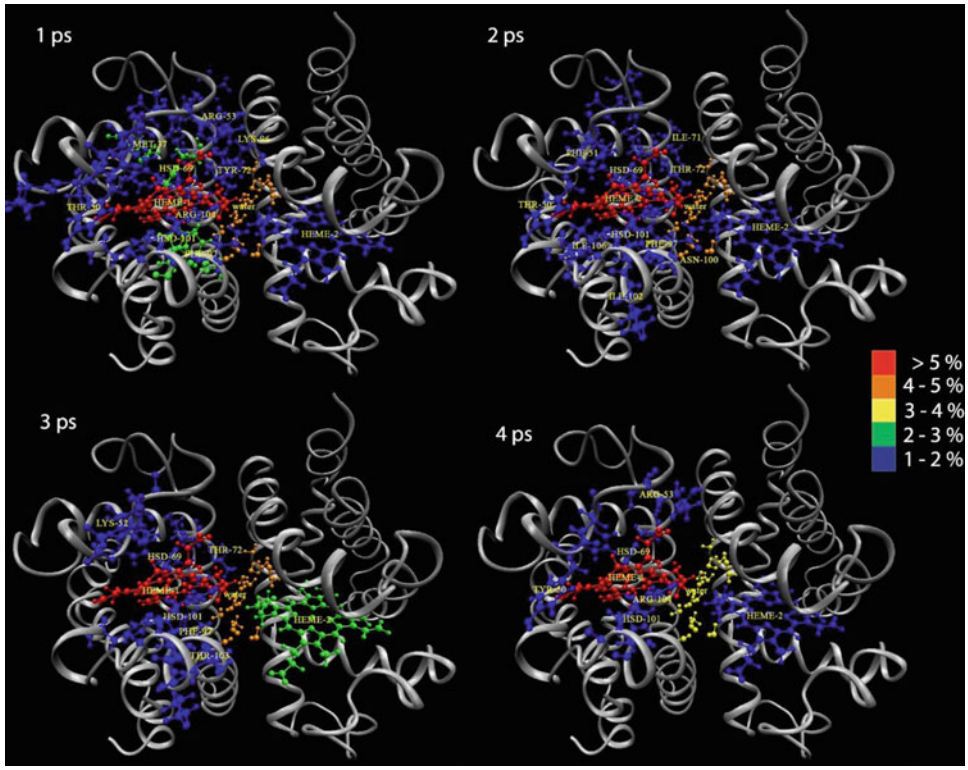

Fig. 1 Simulations of vibrational energy flow in Hbl, starting with all the energy in one of the hemes, shown as the red one at 1 ps. The percentages indicated correspond to percent kinetic energy of the whole system contained in a residue or the interfacial waters. Any part of the protein not highlighted by a color is relatively cold. Reprinted with permission from R. Gnanasekaran, J. K. Agbo and D. M. Leitner, "Communication maps computed for homodimeric hemoglobin: Computational study of water-mediated energy transport in proteins," *J. Chem. Phys.* 135, 065103, Copyright (2011), American Institute of Physics

$D_\alpha^{\{AA\prime\}}$ is the mode-dependent energy diffusivity between regions A and A'. For a local thermal diffusion coefficient to be well defined we need to assume that thermalization occurs within each residue. Thermalization in molecules has been the focus of considerable attention [91–119], in part because it mediates chemical reaction kinetics [120–131], and it appears to largely hold on the scale of peptides [132–138]. In practice a region, A, is a residue or a cofactor such as a heme, or perhaps a cluster of water molecules in the protein. We note that when A and A' span the molecule, Eq. 2 gives the mode diffusivity [90], from which the coefficient of thermal conductivity, κ, can be expressed for the whole system, $\kappa = \sum_\alpha C_\alpha D_\alpha$, where C_α is the heat capacity per unit volume of the molecule ($V = 1$, as it is not relevant below) for mode α, given by

$$C_\alpha = k_B(\beta\hbar\omega_\alpha)^2 \frac{e^{\beta\hbar\omega_\alpha}}{(e^{\beta\hbar\omega_\alpha} - 1)^2}. \tag{3}$$

We calculate a thermal average as we would to obtain the thermal diffusivity for the protein, i.e.,

$$D_{AA'} = \frac{\sum_{\alpha} C_{\alpha}(T) D_{\alpha}^{\{AA\prime\}}}{\sum_{\alpha} C_{\alpha}(T)}, \tag{4}$$

where C_{α} is calculated with Eq. 3, which incorporates the thermal population of the modes and is the only quantum effect that is accounted for in the energy transport. Assuming energy diffusion between pairs of residues, the time constant between A and A' per degree of freedom, $\tau_{AA'}$, is calculated as

$$\tau_{AA'} = d_{AA'}^2 / 2 D_{AA'} \tag{5}$$

where $d_{AA'}$ is the distance between A and A', which in practice we take to be the distance between the center of mass of the two residues. Local energy diffusion occurs along a path between these two centers of mass of regions A and A'. The energy diffusion thus occurs essentially along a one-dimensional path, so we include the factor of 2 as appropriate for diffusion in one dimension.

2.2 Communication Maps: Illustrative Example

We recently constructed an energy transport network for the homodimeric hemoglobin from *Scapharca inaequivalvis*, HbI, where we obtained the transition times between residues with Eq. 5 [22]. In addition to a network where all edges were weighted by $\tau_{AA'}$ we also identified networks of nonbonded residues and the water cluster subject to cutoff times for $\tau_{AA'}$, specifically 2 and 3 ps. Any nonbonded residue pair, or a residue and the water cluster, lies within a nonbonded network (NBN) if they are linked by an edge with a value of $\tau_{AA'}$ that is below the cutoff. While there are many such nonbonded pairs, a criterion whereby at least five nodes must be so connected was used to form an NBN, which indicates pathways along which rapid response to local strain occurs in the protein via nonbonded interactions.

In Fig. 2 we illustrate the energy transport NBNs for the deoxy (top 2 images) and oxy (bottom 2 images) states. The threshold values for τ are 2 ps (two images shown on left) and 3 ps (two images shown on right). Consider first deoxy HbI, plotted as the two images on the top. For the short time cutoff (left) we observe two regions, one (red) that includes the heme, the water cluster, and several residues in the middle of the E helix and the upper portion of the F helix, where more information about the specific residues is detailed in Ref. [22]. Both the proximal and distal histidines belong to the same NBN as the heme and water cluster, a network that spans both globules. The other NBN (purple) includes the salt bridge formed by Lys30 and Asp89, as well as other residues of the upper portion of the B helix, the lower portion of the E helix and a few residues of the F helix. This NBN also spans

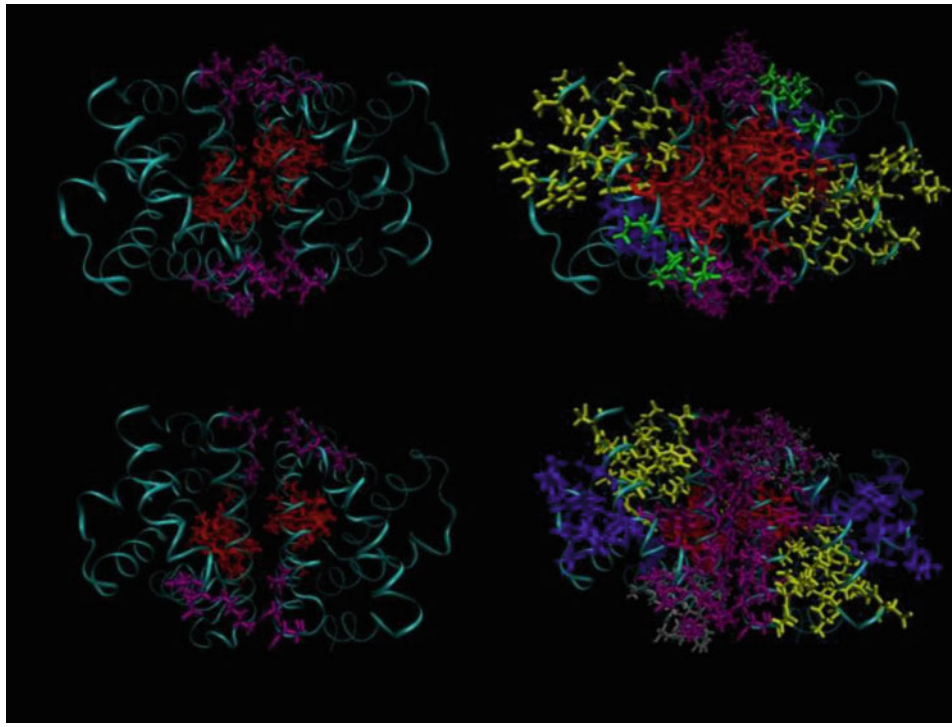

Fig. 2 Nonbonded networks (NBNs) for unliganded (top) and liganded (bottom) HbI. An NBN is defined for at least five connected nonbonded residues where τ is less than 2 ps (left) or 3 ps (right). The most robust NBNs, found using the smaller τ, include the one spanning both globules and including the Lys30-Asp89 salt bridge (purple), and another (red) that includes the hemes, distal and proximal histidines, and other nearby residues. For the unliganded structure it also includes the cluster of water molecules at the interface. Reprinted with permission from D. M. Leitner, "Water-mediated energy dynamics in a homodimeric hemoglobin," *J. Phys. Chem. B* **120**, 4019–4027 (2016). Copyright (2016) American Chemical Society

both globules. When we extend the cutoff to longer times (right) both of these NBNs grow and new ones appear. In addition to the much-expanded red network, which includes the hemes and water cluster, the upper parts of the E, F, and H helices, and the moderately expanded purple network, which includes the salt bridges, three other NBNs localized on each globule form. One of these NBNs (yellow) includes residues from the lower portion of the B helix, residues of the lower portion of the H helix, and a few residues of the E helix. Another (blue) includes a few residues in the upper part of the B helix, the C helix, and the G helix. A third new NBN (green) includes the middle of the B helix.

The NBNs for unliganded HBI are distinct from those of liganded HbI, shown as the bottom two images in Fig. 2. At the shorter cutoff (lower left) we again find only two NBNs, but only one that spans both globules, the purple network that includes the Lys30-Asp89 salt bridge, as well as Asp28, Asn32, Asn86, and Val93. The NBN that includes the heme (red) does not include

the cluster of water molecules at the interface, which is smaller (11 molecules) than in the unliganded protein (17 molecules). The red NBN consists of the heme, His69, Leu73, Leu77, Ala98, His101, and Arg104, as well as a few residues from the E and F helix. At the longer time cutoff (lower right) there is again only one NBN spanning both globules (purple), which includes the Lys30-Asp89 salt bridge and water cluster, as well as the lower portion of the B helix, the upper part of the D helix, and parts of the E and F helices, including Phe97. The red network, which includes the heme, grows only slightly beyond the NBN obtained with the shorter cutoff. Three other NBNs appear each confined to one globule. The yellow and blue NBNs partially overlap the NBNs of the unliganded protein of the same color. Another network (silver) does not overlap NBNs of deoxy HbI. The yellow NBN includes the lower part of the B helix, the upper part of the E helix, and the upper part of the H helix. The blue NBN includes the upper portion of the B helix, the G helix, and the lower part of the H helix. The silver NBN includes parts of the B, C, and E helices.

These NBNs constitute groups of residues that respond to local strain via nonbonded interactions. Both unliganded and liganded states contain an interglobule network with the Lys30-Asp89 salt bridge at its core, while the unliganded protein also contains an interglobule network that includes the hemes and nearby residues bridged by the cluster of water molecules at the interface. For the unliganded protein the more immediate response of the water cluster to local strain at each heme is consistent with expulsion of water molecules that accompanies the allosteric transition to the liganded state, as discussed further below.

Of course, the more complete network also includes the main chain, along which energy transport occurs readily. A more general analysis must therefore include all interactions. One means to quantify information flow along the entire network is calculation of the betweenness centrality, C_B. If we take a residue, heme, and water cluster to be a node, ν, of a network, then C_B is defined by [34, 139].

$$C_B(\nu) = \frac{2}{(N-1)(N-2)} \sum_{s=1}^{N-1} \sum_{t=s+1}^{N} \frac{\sigma_{st}(\nu)}{\sigma_{st}}, \tag{6}$$

where N is the number of nodes in the network, σ_{st} is the number of shortest paths linking nodes s and t, and $\sigma_{st}(\nu)$ is the number of shortest paths between s and t that also include node ν. The betweenness centrality has been used to locate hubs in information flow related to protein dynamics [34], though other centrality measures may also be usefully adopted [51]. For a weighted network, where values for the edges are the time constants given by Eq. 5, we can locate the shortest path between nodes s and t using the Dijkstra algorithm [140].

We have computed values for the betweenness centrality, C_B, for all the nodes of unliganded and liganded structures HbI and plot the largest values for each structure in Fig. 3. The largest value for the unliganded structure, 0.15, lies on the cluster of water molecules at the interface, highlighting again the centrality of this feature to the global network of energy transport in unliganded HbI. The next largest values, 0.12, lie on the Lys30-Asp89 salt bridge, with somewhat smaller but comparable values on nearby residues, including Arg67, Leu84, Asp85, Asn86, Pro87, and Asp88. The hubs of information flow on the energy transport network as quantified by C_B lie in regions critical to the stability of the protein, and to the allosteric regulation of HbI.

Mutation studies that influence interactions between the water cluster and the protein reveal significant effects on cooperativity [83–85]. Mutation of Lys30 to Asp30 destabilizes the protein altering the mechanism of cooperativity, which then involves dissociation of the two globules upon oxygen binding, and reformation of the dimer upon dissociation [86]. The largest values of the betweenness centrality calculated for the energy transport network apparently identify two regions that control allostery in this protein. Both of these regions controlling allosteric regulation identified here were also identified as NBNs.

The largest values of C_B calculated for the liganded protein, also plotted in Fig. 3, are 0.16 and 0.13 for, respectively, Lys30 and Asp89, with somewhat smaller values obtained for several residues of the E and F helices at the interface between the proteins, including residues Tyr75, Leu77, Gln78, Asn79, Gln83, Leu84, Asp85, Val91, Cys92, and Val93. The Lys30-Asp89 salt bridge forms the basis of the only interglobule network for the liganded protein, and its pivotal role in cooperativity was noted above. The residues around Cys92 on the F helix form a hinge with residues around Arg67 on the E helix that has been observed in time-resolved crystallography experiments to serve as a pivot point for structural change following ligand photolysis [141]. We find sizable values of the betweenness centrality, 0.11 and 0.06, respectively, for Cys92 and Arg67. More details concerning network analysis of this allosteric protein can be found in Ref. [22].

2.3 Master Equation Simulations of Energy Dynamics on a Network

We turn now to dynamics along a network, where we simulate energy dynamics with a master equation, using as rate constants the time constants obtained with the local energy diffusion coefficients, Eq. 5. In a recent study of the 36-amino acid fragment from the villin headpiece subdomain, HP36, the results of a master equation simulation using the rate constants obtained from communication maps were compared with results of all-atom nonequilibrium simulations. Here we summarize work carried out in that study [17]. The master equation is,

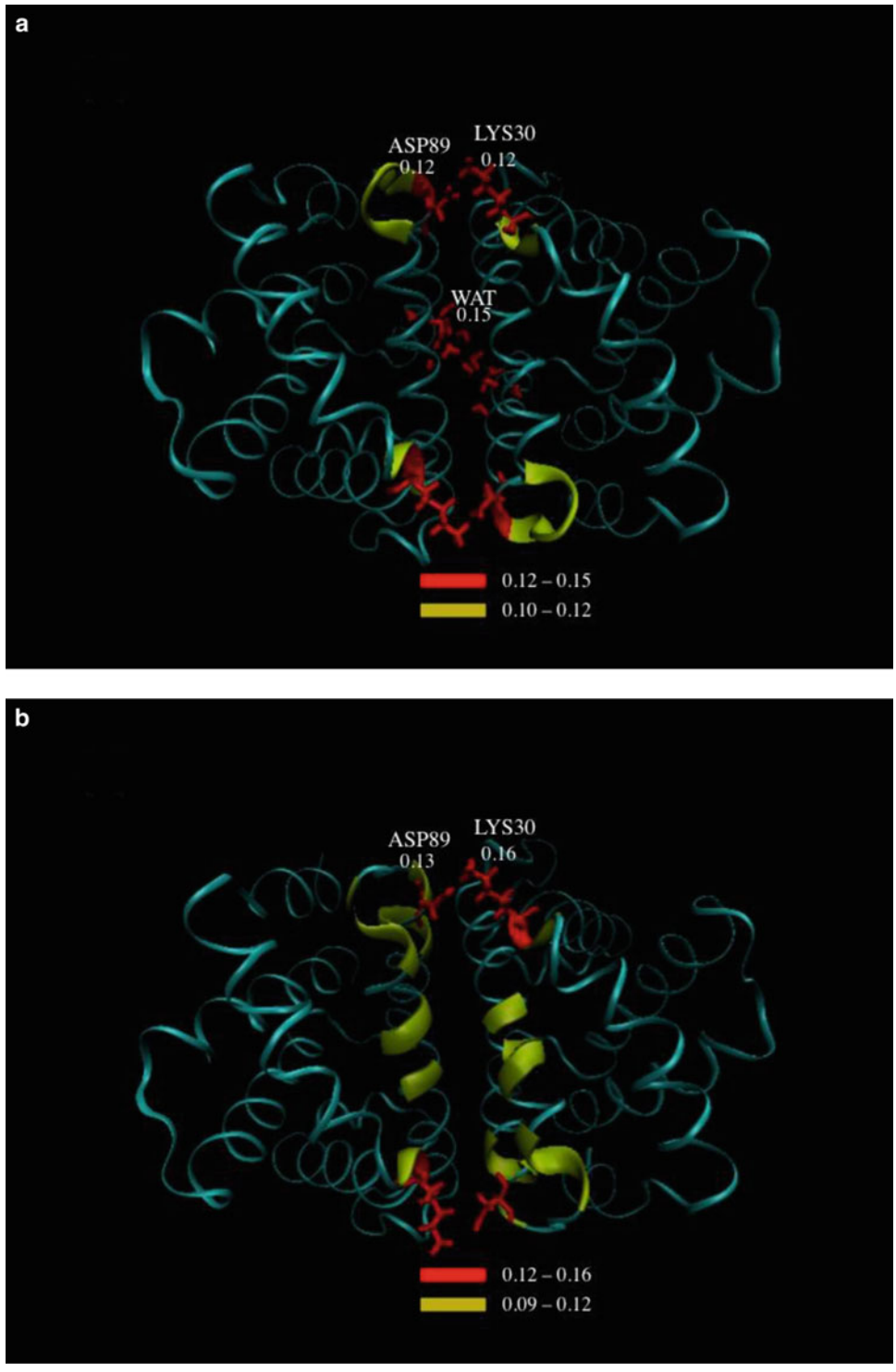

Fig. 3 Sizable values of the betweenness centrality, C_B, for nodes of the energy transport network for the (**a**) unliganded and (**b**) liganded states. Nodes with the largest values (red) are labeled and include side chains. Other nodes with sizable values are indicated in yellow. Reprinted with permission from D. M. Leitner, "Water-mediated energy dynamics in a homodimeric hemoglobin," *J. Phys. Chem. B* **120**, 4019–4027 (2016). Copyright (2016) American Chemical Society

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{k}\mathbf{P}(t), \tag{7}$$

where $\mathbf{P}$ is a vector with elements corresponding to the population of each residue and $\mathbf{k}$ is the matrix of transition probabilities between residues. The elements of the matrix, $\{k_{ij}\}$, are the rate constants for energy transfer between a pair of residues, i and j. The solutions of the master equation describing the time evolution of the population of the residues are given by

$$\mathbf{P}(t) = \exp(\mathbf{k}t)\mathbf{P}(0), \tag{8}$$

The elements of the rate matrix were obtained using Eq. 5. Damping due to coupling to the solvent was also included in some of the simulations reported in Ref. [17], which matched closely results of all-atom nonequilibrium simulations of hydrated villin, but we summarize here only the results without damping.

The detailed analysis of energy flow in HP36 revealed some shortcuts in sequence space. Initial excitation of the protein was taken to be near the middle of the sequence, at residue 16. Because of the hydrogen bond between residues 15 and 4, shown in Fig. 4, the authors examined the population of residues near 4. In Fig. 5a, $P(t)$ is plotted [17] for residues 3–7 obtained from the master equation simulation, where the hydrogen bond between residues 4 and 15 gives rise to rapid energy transport to residue 4. Energy is also seen to reach residues 3 and 7 relatively quickly, followed by residues 5 and 6, which, like the others, are seen to reach their equilibrium populations of ≈0.028 somewhat after 20 ps. Since the system studied here is closed, the population of each residue

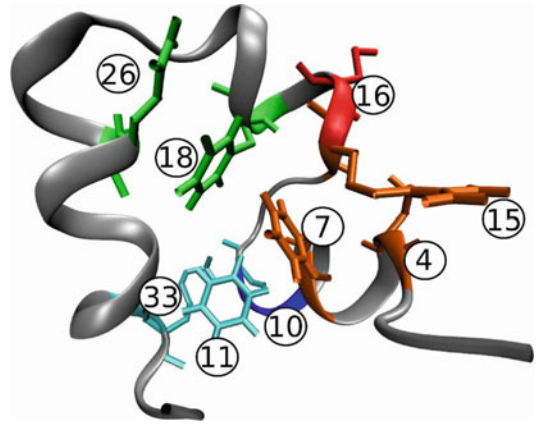

Fig. 4 Villin headpiece subdomain (HP36) with some of the residues discussed in text highlighted. Reprinted with permission from D. M. Leitner, S. Buchenberg, P. Brettel, G. Stock, "Vibrational energy flow in the villin headpiece subdomain: Master equation simulations," *J. Chem. Phys.* 142, 075101, Copyright (2015), American Institute of Physics

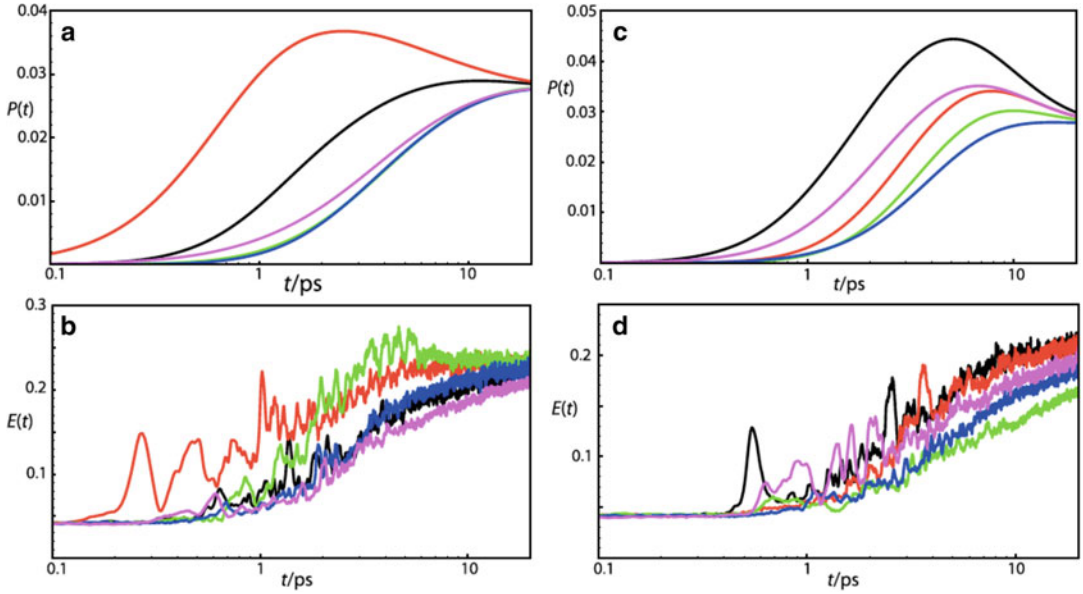

Fig. 5 (**a**) Master equation simulation of $P(t)$ and (**b**) all-atom nonequilibrium MD simulation of kinetic energy per degree of freedom, $E(t)$, for residues 3 (black), 4 (red), 5 (green), 6 (blue), and 7 (magenta) of HP36 when residue 16 is heated initially. Rapid heating of residue 4 arises from shortcut due to hydrogen bond between residues 4 and 15. (**c**) Master equation simulation of $P(t)$ and (**d**) all-atom simulation of kinetic energy per degree of freedom, $E(t)$, for residues 22 (black), 23 (red), 24 (green), 25 (blue), and 26 (magenta) of HP36 when residue 16 is heated initially. Rapid heating of residue 26 arises from shortcut due to hydrogen bond between residues 18 and 26. Reprinted with permission from D. M. Leitner, S. Buchenberg, P. Brettel, G. Stock, "Vibrational energy flow in the villin headpiece subdomain: Master equation simulations," *J. Chem. Phys.* 142, 075101, Copyright (2015), American Institute of Physics

converges to the inverse of the number of residues in the protein, which for the 36-residue villin headpiece subdomain is about 0.028.

The results of the master equation simulations were compared with the population of residues 3–7 obtained by all-atom nonequilibrium simulations, plotted in Fig. 5b. Overall energy flow into and out of the residues in this part of the protein occurs at times similar to the times found in the master equation simulation, where some modest differences were attributed to the time needed to heat residue 16 from the attached azobenzene in the all-atom simulations, not accounted for in the master equation simulations, among other factors. The two simulations were found to provide a consistent picture for all residues at early times, i.e., below 1 ps, with some differences seen in the heating and cooling of some of the individual residues beyond 1 ps. The results of the two simulations converged again at longer times, beyond about 10 ps, as equilibrium is approached.

Another shortcut in sequence space due to a hydrogen bond appears in a second region of the protein and was also examined. *P*

(t) for residues 22–26 obtained by master equation simulations is shown in Fig. 5c, which can be compared with the time-dependent energy obtained by the all-atom nonequilibrium simulations, plotted in Fig. 5d. Figure 5c shows that energy transport to residue 22 is fastest in this region, followed by residue 26, then followed by residues 23, 24, and 25, the latter two appearing around the same time. The sequence could be explained by the local energy diffusion coefficients calculated between residue 16 and the residues of this part of the sequence, as well as values of the other local energy diffusion coefficients corresponding to this part of the protein. At early times, similar trends are seen in the all-atom simulations, plotted in Fig. 5d, and again some differences are seen between 1 ps and the equilibration times beyond 10 ps, and are further discussed in Ref. [17].

2.4 Rate Constants: Scaling Energy Transport Through Hydrogen Bonds with Hydrogen Bond Fluctuations

While the local thermal diffusion coefficients appear to provide a means to estimate the rate constants in a master equation, an alternative based on a scaling relation between fluctuations of nonbonded residues and rate constants was found for pairs of hydrogen bonded residues of villin [23]. The results of the all-atom nonequilibrium simulations, which were carried out at low temperature (less than 100 K), were fit to a master equation, and the resulting rate constants, when introduced to the master equation, reproduced the results of the all-atom simulations very closely [23]. Interestingly, Stock and coworkers found that those rate constants scale with $1/\left\langle \delta r_{ij}^2 \right\rangle$, where $\left\langle \delta r_{ij}^2 \right\rangle$ is the variance in the distance between the two atoms, i and j, forming the hydrogen bond. Some theoretical justification for the scaling relation is discussed in Ref. [23]. Briefly [23, 59, 142, 143], the master equation given by Eq. 7 for the diffusion of energy among residues has the same form as the equation of motion for lattice vibrations with nearest-neighbor interactions, $m_\alpha \frac{d^2 u_\alpha}{dt^2} = \sum_{\beta \neq \alpha} k_{\beta\alpha} u_\beta$, where m_α and u_α are the mass and displacement, respectively, at site α, and here $k_{\alpha\beta}$ is the force constant connecting the masses. The only difference between the equations is the presence of a first- and second-order time derivative, respectively. A solution to the diffusion equation is obtained from the vibrational problem by substituting t for ω^{-2} [142], which is proportional to $\left\langle \delta r_{ij}^2 \right\rangle$, so that the transition rate between i and j, which is proportional to the energy flux between these residues, is proportional to $1/\left\langle \delta r_{ij}^2 \right\rangle$.

Here we examine scaling relations for hydrogen bonds in apomyoglobin by comparing the energy flux and hydrogen bond fluctuations obtained by classical MD simulations at 300 K. More details can be found in Ref. [143]. The simulations were carried out as follows: The solvated protein system investigated, apomyoglobin, was created from the starting structure of bilverdin apomyoglobin (PDB 1BVD). We carried out the MD simulations using the AMBER16 MD package and the amber ff14SB force

field [144]. The protein structure was first minimized with steepest descent followed by conjugate gradient for 500 steps total. The system was solvated with 6661 solvent molecules with a NaCl concentration of 0.2 M in water using the TIP3P water model. The final system was minimized once more for 2000 steps to remove bad contacts. Consecutive position-restrained 1 ns canonical ensemble and 1 ns isothermal–isobaric ensemble simulations were performed with a heavy atom restraint force constant of 500 kcal/(mol Å^2) with the Berendsen thermostat and barostat [145] at 300 K and 1 bar, respectively. The system was then simulated under isothermal–isobaric ensemble for 700 ps converging the system density, followed by a 5 ns production simulation. System snapshots were taken at 100 ps intervals along the 5 ns trajectory and were further simulated using the microcanonical ensemble (NVE) for 150 ps using 0.5 fs time steps, outputting the velocity and position trajectory files every 10 fs.

The trajectories of the MD simulations were then analyzed. To compute the inter-residue flux for each NVE simulation we used the CURP (CURrent calculations for Proteins) program developed by Yamato and coworkers [13]. The flux between residues is multiplied by k_BT, where T was 300 K in the MD simulations, and the results are reported in units of (kJ/mol)^2 ps^{-1}. The polar contact dynamics was calculated over the same trajectory. A polar contact search was carried out on each simulation looking for all donor nitrogen or oxygen, donor proton and oxygen acceptor triplets and for all triplets where the donor proton and oxygen acceptor distances are $\leq$0.28 nm for at least 99% of the NVE trajectory. For the pairs that met the distance criteria we calculated $1/\left\langle \delta r_{ij}^2 \right\rangle$. For each trajectory the polar contact dynamics calculated was paired with the corresponding energy flux. We then defined hydrogen bonds as having a donor-proton-acceptor angle greater than or equal to 150° over the duration of the simulation.

For each pair found in any one of the simulations that satisfied the hydrogen bond criterion, we plot the value of the flux against $1/\left\langle \delta r_{ij}^2 \right\rangle$ in Fig. 6. As in the HP36 work [23], amino acid pairs within 4 in sequence space are not included, since those pairs are within helices along which energy transport is anyway relatively fast, the rate constants for which follow a different relation appropriate for energy transfer along the backbone [23]. The data are separated into two groups, one for hydrogen bond pairs that lie 5–9 amino acids away and another group where they are further away in sequence space. The first group consists only of pairs that are 5 and 6 away in sequence space, and 3 of the four pairs identified involve backbone–backbone hydrogen bonds. They are shown in Fig. 7, and are seen to be at turns just beyond one of the helices. The points corresponding to those hydrogen bonds appear to be well fit by a line in Fig. 6 given by flux $= 76.4 + 0.0273/\left\langle \delta r_{ij}^2 \right\rangle$,

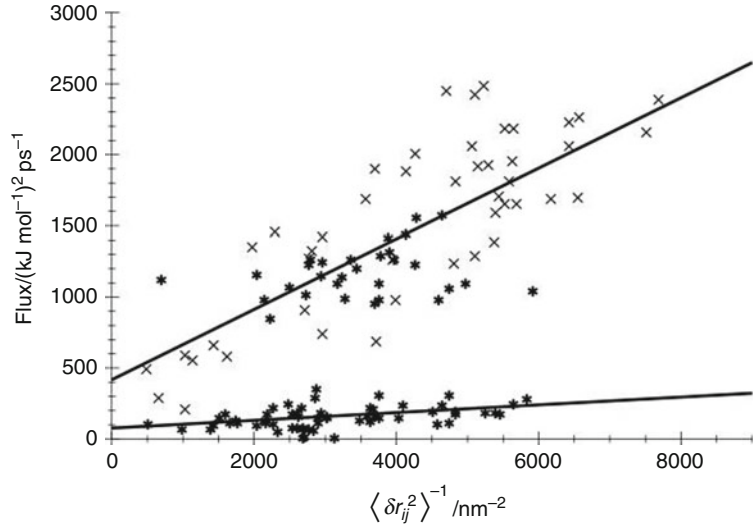

Fig. 6 Flux vs. $1/\left\langle \delta r_{ij}^2 \right\rangle$, where $\left\langle \delta r_{ij}^2 \right\rangle$ is the variance in the distance between the two atoms, *i* and *j*, forming the hydrogen bond. Pairs within 5 and 9 residues in sequence space are indicated by * and those 10 or more residues away by X. Two linear regions are seen, one apparently corresponding to backbone–-backbone hydrogen bonds, which have a smaller slope, and the rest corresponding to side chain–backbone hydrogen bonds. Linear fits to each set are shown and discussed in the text

which is shown in the figure. A second set of points also appears to fall fairly close to a line. They include one of the hydrogen bonds close in sequence space and the rest, which are further away. All of these hydrogen bonds are formed between a side chain and backbone, and a fit gives flux $= 417.1 + 0.248/\left\langle \delta r_{ij}^2 \right\rangle$, which is also shown in the figure. A linear relation between flux and $1/\left\langle \delta r_{ij}^2 \right\rangle$ was justified by the relation between diffusion along an elastic network of residues. We find here a different slope for hydrogen bonds formed, respectively, between side chains and those formed between backbone and side chain. Further work has been reported in Ref. [143].

3 Concluding Remarks

Energy transport networks can be located computationally by the thermal transport approach reviewed here. For the dimeric hemoglobin discussed above the energy transport channels that have been identified and characterized overlap regions involved in allosteric transitions. Simulations of energy transport dynamics can be carried out with master equation approaches using rate constants obtained from the local energy diffusion coefficients computed for

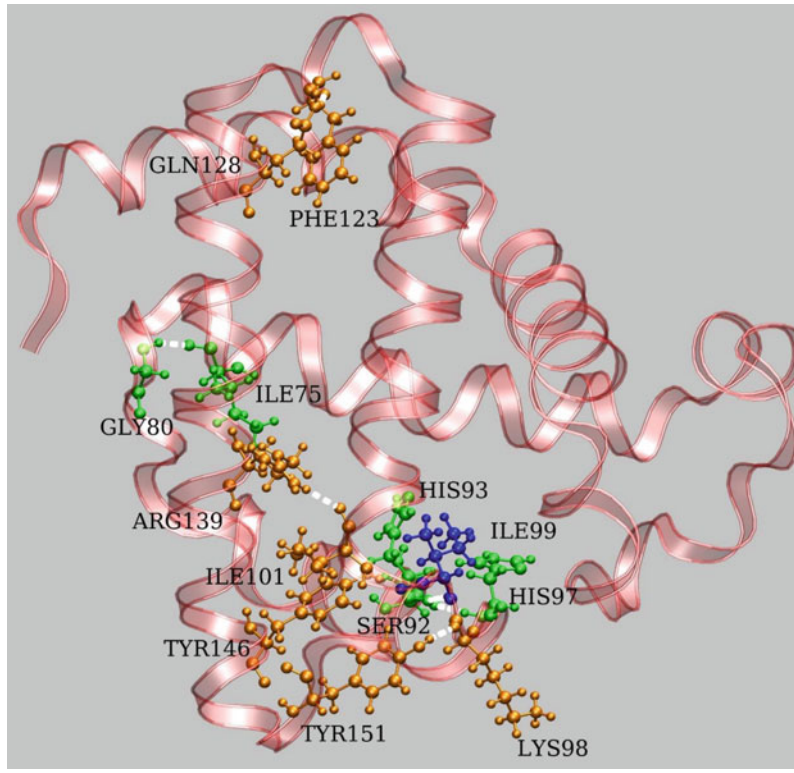

Fig. 7 Apomyoglobin, highlighting hydrogen bonds contributing to the results in Fig. 6. Only those hydrogen bonds where the distance between amino acids in sequence is greater than 4 were considered. Two scaling relations were found, one for backbone–backbone hydrogen bonds (participating residues are green) and another for side chain–backbone hydrogen bonds (orange). ILE99 participates in both groups (blue)

the protein. There are also indications that at least some of the rate constants may be related to protein structural fluctuations, and can thus be obtained from relatively short MD simulations. This was originally pointed out in a computational study of energy flow in the villin headpiece subdomain [23], HP36, and we have here presented newer results exploring scaling relations between hydrogen bond dynamics and energy transfer along hydrogen bonds in apomyoglobin.

Future work will explore connections between dynamics of other kinds of nonbonded contacts and energy flow between them. For instance, work on apomyoglobin indicates that ionic contacts, which are less localized, exhibit much greater variability than hydrogen bonds [143]. Moreover, less localized interactions will couple dynamically to the hydration layer surrounding a protein. Protein and water dynamics are coupled, as revealed, e.g., by THz measurements and molecular simulations [146–171]. Fluctuations of contacts closer to the surface will undoubtedly be influenced by the water dynamics as well. Despite these complicating

features, future computational studies of energy transport in larger proteins and protein complexes will be facilitated by any connections between protein dynamics and energy transfer that can be identified. Work in these directions is in progress [172–175].

Acknowledgements

The authors are grateful to Prof. Takahisa Yamato for making available his program CURP and for a number of helpful discussions. Some of the work reviewed here is the result of a collaboration DML has enjoyed with Gerhard Stock and Sebastian Buchenberg on modeling energy dynamics in proteins. Support from NSF grants CHE-1361776 and CHE-1854271 is gratefully acknowledged.

References

1. Leitner DM, Straub JE (2009) Proteins: energy, heat and signal flow. CRC Press, Boca Raton, FL
2. Nguyen PH, Hamm P, Stock G (2009) Non-equilibrium molecular dynamics simulation of photoinduced energy flow in peptides: theory meets experiment. In: Leitner DM, Straub JE (eds) Proteins: energy, heat and signal flow. CRC Press, Boca Raton, FL, pp 149–168
3. Hassan S, Schade M, Shaw CP, Levy P, Hamm P (2014) Response of villin headpiece-capped gold nanoparticles to ultrafast laser heating. J Phys Chem B 118:7954–7962
4. Botan V, Backus EHG, Pfister R, Moretto A, Crisma M, Toniolo C, Nguyen PH, Stock G, Hamm P (2007) Energy transport in peptide helices. Proc Natl Acad Sci U S A 104:12749–12754
5. Backus EHG, Nguyen PH, Botan V, Pfister R, Moretto A, Crisma M, Toniolo C, Stock G, Hamm P (2008) Energy transport in peptide helices: a comparison between high- and low-energy excitations. J Phys Chem B 112:9091–9099
6. Backus EHG, Nguyen PH, Botan V, Moretto A, Crisma M, Toniolo C, Zerbe O, Stock G, Hamm P (2008) Structural flexibility of a helical peptide regulates vibrational energy transport properties. J Phys Chem B 112:15487–15492
7. Backus EH, Bloem R, Pfister R, Moretto A, Crisma M, Toniolo C, Hamm P (2009) Dynamical transition in a small helical peptide and its implication for vibrational energy transport. J Phys Chem B 113:13405–13409
8. Kondoh M, Mizuno M, Mizutani Y (2016) Importance of atomic contacts in vibrational energy flow in proteins. J Phys Chem Lett 7:1950–1954
9. Fujii N, Mizuno M, Mizutani Y (2011) Direct observation of vibrational energy flow in cytochrome c. J Phys Chem B 115:13057–13064
10. Fujii N, Mizuno M, Ishikawa H, Mizutani Y (2014) Observing vibrational energy flow in a protein with the spatial resolution of a single amino acid residue. J Phys Chem Lett 5:3269–3273. https://doi.org/10.1021/jz501882h
11. Sagnella DE, Straub JE, Thirumalai D (2000) Timescales and pathways for kinetic energy relaxation in solvated proteins: application to carbonmonoxy myoglobin. J Chem Phys 113:7702–7711
12. Bu L, Straub JE (2003) Simulating vibrational energy flow in proteins: relaxation rate and mechanism for heme cooling in cytochrome c. J Phys Chem B 107:12339–12345
13. Ishikura T, Iwata Y, Hatano T, Yamato T (2015) Energy exchange network of interresidue interactions within a thermally fluctuating protein: a computational study. J Comput Chem 36:1709–1718. https://doi.org/10.1002/jcc.23989
14. Ishikura T, Yamato T (2006) Energy transfer pathways relevant for long-range intramolecular signaling of photosensory protein revealed by microscopic energy conductivity analysis. Chem Phys Lett 432:533–537
15. Xu Y, Leitner DM (2014) Vibrational energy flow through the green fluorescent protein-

water interface: communication maps and thermal boundary conductance. J Phys Chem B 118:7818–7826

16. Xu Y, Leitner DM (2014) Communication maps of vibrational energy transport in photoactive yellow protein. J Phys Chem A 118:7280–7287
17. Leitner DM, Buchenberg S, Brettel P, Stock G (2015) Vibrational energy flow in the villin headpiece subdomain: master equation simulations. J Chem Phys 142:075101
18. Agbo JK, Gnanasekaran R, Leitner DM (2014) Communication maps: exploring energy transport through proteins and water. Isr J Chem 54:1065–1073
19. Leitner DM (2009) Frequency resolved communication maps for proteins and other nanoscale materials. J Chem Phys 130:195101
20. Gnanasekaran R, Agbo JK, Leitner DM (2011) Communication maps computed for homodimeric hemoglobin: computational study of water-mediated energy transport in proteins. J Chem Phys 135:065103
21. Agbo JK, Xu Y, Zhang P, Straub JE, Leitner DM (2014) Vibrational energy flow across heme-cytochrome c and cytochrome c-water interfaces. Theor Chem Accounts 133:1504
22. Leitner DM (2016) Water-mediated energy dynamics in a homodimeric hemoglobin. J Phys Chem B 120:4019–4027
23. Buchenberg S, Leitner DM, Stock G (2016) Scaling rules for vibrational energy transport in proteins. J Phys Chem Lett 7:25–30
24. Martínez L, Figueira ACM, Webb P, Polikarpov I, Skaf MS (2011) Mapping the intramolecular vibrational energy flow in proteins reveals functionally important residues. J Phys Chem Lett 2:2073–2078
25. Leitner DM, Yamato T (2018) Mapping energy transport networks in proteins. In: Parrill AL, Lipkowitz KB (eds) Rev. Comp. Chem, vol 31. Wiley, New York, pp 63–114
26. Leitner DM (2013) Thermal boundary conductance and rectification in molecules. J Phys Chem B 117:12820–12828
27. Rubtsova NI, Rubtsov IV (2015) Vibrational energy transport in molecules studied by relaxation-assisted two-dimensional infrared spectroscopy. Ann Rev Phys Chem 66:717–738
28. Sethi A, Eargle J, Black AA, Luthey-Schulten Z (2009) Dynamical networks in tRNA: protein complexes. Proc Natl Acad Sci U S A 106:6620–6625
29. Ribeiro AAST, Ortiz V (2015) Energy propagation and network energetic coupling in proteins. J Phys Chem B 119:1835–1846
30. Ribeiro AAST, Ortiz V (2014) Determination of signaling pathways in proteins through network theory: importance of the topology. J Chem Theor Comput 10:1762–1769
31. DiPaola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48
32. Feher VA, Durrant JD, Wart ATV, Amaro RE (2014) Computational approaches to mapping allosteric pathways. Curr Opin Struct Biol 25:98–103
33. Gursoy A, Keskin O, Nussinov R (2008) Topological properties of protein interaction networks from a structural perspective. Biochem Soc Trans 36:1398–1403
34. Lee Y, Choi S, Hyeon C (2014) Mapping the intramolecular signal transduction of G-protein coupled receptors. Proteins 82:727–743
35. Miao Y, Nichols SE, Gasper PM, Metzger VT, McCammon JA (2013) Activation and dynamic network of the M2 muscarinic receptor. Proc Natl Acad Sci U S A 110:10982–10987
36. Del-Sol A, Fujihashi H, Amoros D, Nussinov R (2006) Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol Sys Biol 2:2006.0019
37. Atilgan AR, Turgut D, Atilgan C (2007) Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication. Biophys J 92:3052–3062
38. Woods KN (2014) Using THz time-scale infrared spectroscopy to examine the role of collective, thermal fluctuations in the formation of myoglobin allosteric communication pathways and ligand specificity. Soft Matter 10:4387–4402
39. Woods KN, Pfeffer J (2015) Using THz spectroscopy, evolutionary network analysis methods, and MD simulation to map the evolution of allosteric communication pathways in c-type lysozymes. Mol Biol Evol 2:271–274
40. Achoch M, Dorantes-Gilardi R, Wymant C, Feverati G, Salamatian K, Vuillon L, Lesieur C (2016) Protein structural robustness to mutations: an in silico investigation. Phys Chem Phys 16:13770–13780
41. Ribeiro AAST, Ortiz V (2016) A chemical perspective on allostery. Chem Rev 116:6488–6502
42. Vuillon L, Lesieur C (2015) From local to global changes in proteins: a network view. Curr Opin Struct Biol 31:1–8

43. DiPaola L, DeRuvo M, Paci P, Santoni D, Giuliani A (2013) Protein contact networks: an emerging paradigm in chemistry. Chem Rev 113:1598–1613
44. Dokholyan NV (2016) Controlling allosteric networks in proteins. Chem Rev 116:6463–6487
45. Livi L, Maiorino E, Pinna A, Sadeghian A, Rizzi A, Giuliani A (2016) Analysis of heat kernel highlights the strongly modular and heat-preserving structure of proteins. Physica A 441:199–214
46. Livi L, Maiorino E, Giuliani A, Rizzi A, Sadeghian A (2016) A generative model for protein contact networks. J Biomol Struct Dynamics 34:1441–1454
47. Khor S (2017) Comparing local search paths with global search paths on protein residue networks: allosteric communication. J Complex Networks 5:409–432
48. Khor S (2016) Protein residue networks from a local search perspective. J Complex Networks 4:245–278
49. Avd V, Lorkowski A, Ma N, Gray GM (2017) Computer simulations of the retinoid X receptor: conformational dynamics and allosteric networks. Curr Top Med Chem 17:731–741
50. Amor BRC, Schaub MT, Yaliriki SN, Barahona M (2016) Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. Nat Commun 7:12477. https://doi.org/10.1038/ncomms12477
51. Censoni L, dosSantosMuniz H, Martínez L (2017) A network model predicts the intensity of residue-protein thermal coupling. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx124
52. Banerji A, Ghosh I (2011) Fractal symmetry of protein interior: what have we learned? Cell Mol Life Sci 68:2711–2737
53. Reuveni S, Klafter J, Granek R (2012) Dynamic structure factor of vibrating fractals. Phys Rev Lett 108:068101
54. Reuveni S, Granek R, Klafter J (2010) Anomalies in the vibrational dynamics of proteins are a consequence of fractal-like structure. Proc Natl Acad Sci U S A 107:13696–13670
55. Granek R (2011) Proteins as fractals: role of the hydrodynamic interaction. Phys Rev E 83:020902
56. Enright MB, Yu X, Leitner DM (2006) Hydration dependence of the mass fractal dimension and anomalous diffusion of vibrational energy in proteins. Phys Rev E 73:051905
57. Enright MB, Leitner DM (2005) Mass fractal dimension and the compactness of proteins. Phys Rev E 71:011912
58. Yu X, Leitner DM (2003) Anomalous diffusion of vibrational energy in proteins. J Chem Phys 119:12673–12679
59. Leitner DM (2008) Energy flow in proteins. Ann Rev Phys Chem 59:233–259
60. Chowdary P, Gruebele M (2009) Molecules: what kind of bag of atoms? J Phys Chem A 113:13139–13143
61. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 10:59–69
62. Ota N, Agard DA (2005) Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. J Mol Biol 351:345–354
63. Sharp K, Skinner JJ (2006) Pump-probe molecular dynamics as a tool for studying protein motion and long range coupling. Proteins 65:347–361
64. Lu C, Knecht V, Stock G (2016) Long-range conformational response of a PDZ domain to ligand binding and release: a molecular dynamics study. J Chem Theor Comput 12:870–878
65. Buchenberg S, Knecht V, Walser R, Hamm P, Stock G (2014) Long-range conformational transition of a photoswitchable allosteric protein: molecular dynamics simulation study. J Phys Chem B 118:13468–13476. https://doi.org/10.1021/jp506873y
66. Cui Q, Karplus M (2008) Allostery and cooperativity revisited. Protein Sci 17:1295–1307. https://doi.org/10.1110/ps.03259908
67. Tsai C-J, delSol A, Nussinov R (2009) Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. Mol BioSystems 5:207–216
68. Changeux J-P (2012) Allostery and the Monod-Wyman-Changeux model after 50 years. Annu Rev Biophys 41:103–133
69. Diaz-Franulic I, Poblete H, Miño-Galaz G, González C, Latorre R (2016) Allosterism and structure in thermally activated transient receptor potential channels. Annu Rev Biophys 45:371–398
70. Smock RG, Gierasch LM (2009) Sending signals dynamically. Science 324:198–203
71. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286:295–299

72. LeVine MV, Weinstein H (2014) NbIT—a new information theory-based analysis of allosteric mechanisms reveals residues that underlie function in the leucine transporter LeuT. PLoS Comput Biol 10:e1003603. https://doi.org/10.1371/journal.pcbi.1003603
73. Motlagh HN, Wrabl JO, Li J, Hilser VJ (2014) The ensemble nature of allostery. Nature 508:331–339
74. Liu T, Whitten ST, Hilser VJ (2006) Ensemble-based signatures of energy propagation in proteins: a new view of an old phenomenon. Proteins 62:728–738
75. Zhuralev PI, Papoian GA (2010) Protein functional landscapes, dynamics, allostery: a tortuous path towards a universal theoretical framework. Q Rev Biophys 43:295–332
76. Nagy AM, Raicu V, Miller RJD (2005) Nonlinear optical studies of heme protein dynamics: implications for proteins as hybrid states of matter. Biochim Biophys Acta 1749:148–172
77. Miller RJD (1991) Vibrational energy relaxation and structural dynamics of heme proteins. Ann Rev Phys Chem 42:581–614
78. Royer WE, Knapp JE, Strand K, Heaslet HA (2001) Cooperative hemoglobins: conserved fold, diverse quaternary assemblies and allosteric mechanisms. Trends Biochem Sci 26:297–304
79. Royer WE, Zhu H, Gorr TA, Flores JF, Knapp JE (2005) Allosteric hemoglobin assembly: diversity and similarity. J Biol Chem 280:27477–27480
80. Laine JM, Amat M, Morgan BR, Royer WE, Massi F (2014) Insight into the allosteric mechanism of Scapharca dimeric hemoglobin. Biochemist 53:7199–7210
81. Ikeda-Saito M, Yonetani T, Chiancone E, Ascoli F, Verzili D, Antonini E (1983) Thermodynamic properties of oxygen equilibria of dimeric and tetrameric hemoglobins from Scapharca inaequivalvis. J Mol Biol 170:1009–1018
82. Gnanasekaran R, Xu Y, Leitner DM (2010) Dynamics of water clusters confined in proteins: a molecular dynamics simulation study of interfacial waters in a dimeric hemoglobin. J Phys Chem B 114:16989–16996
83. Pardanani A, Gambacurta A, Ascoil F, Royer WE (1998) Mutational destabilization of the critical interface water cluster in Scapharca dimeric hemoglobin: structural basis for altered allosteric activity. J Mol Biol 284:729–739
84. Pardanani A, Gibson QH, Colotti G, Royer WE (1997) Mutation of residue Phe 97 to Leu disrupts the central allosteric pathway in Scapharca dimeric hemoglobin. J Biol Chem 272:13171–13179
85. Royer WE, Pardanani A, Gibson QH, Peterson ES, Friedman JM (1996) Ordered water molecules as key allosteric mediators in a cooperative dimeric hemoglobin. Proc Natl Acad Sci 93:14526–14531
86. Ceci P, Giangiacomo L, Boffi A, Chiancone E (2002) The mutation K30D disrupts the only salt bridge at the subunit interface of the homodimeric hemoglobin from Scapharca inaequivalvis and changes the mechanism of cooperativity. J Biol Chem 277:6929–2933
87. Royer WE, Hendrickson WA, Chiancone E (1990) Structural transitions upon ligand binding in a cooperative dimeric hemoglobin. Science 249:518–521
88. Elber R (2007) A milestoning study of the kinetics of an allosteric transition: atomically detailed simulations of deoxy Scapharca hemoglobin. Biophys J 92:L85–L87
89. Sagnella DE, Straub JE (2001) Directed energy "funneling" mechanism for heme cooling following ligand photolysis or direct excitation in solvated carbonmonoxy myoglobin. J Phys Chem B 105:7057–7063
90. Allen PB, Feldman JL (1993) Thermal conductivity of disordered harmonic solids. Phys Rev B 48:12581–12588
91. Pandey HD, Leitner DM (2016) Thermalization and thermal transport in molecules. J Phys Chem Lett 7:5062–5067
92. Pandey HD, Leitner DM (2017) Vibrational energy transport in molecules and the statistical properties of vibrational modes. Chem Phys 482:81–85
93. Leitner DM (2005) Heat transport in molecules and reaction kinetics: the role of quantum energy flow and localization. Adv Chem Phys 130B:205–256
94. Leitner DM (2015) Quantum ergodicity and energy flow in molecules. Adv Phys 64:445–517
95. Leitner DM, Wolynes PG (1996) Vibrational relaxation and energy localization in polyatomics: effects of high-order resonances on flow rates and the quantum ergodicity transition. J Chem Phys 105(24):11226–11236
96. Leitner DM, Wolynes PG (1996) Statistical properties of localized vibrational eigenstates. Chem Phys Lett 258:18–24
97. Yu X, Leitner DM (2003) Vibrational energy transfer and heat conduction in a protein. J Phys Chem B 107:1698–1707
98. Leitner DM, Gruebele M (2008) A quantum model of restricted vibrational energy flow on

the way to the transition state in unimolecular reactions. Mol Phys 106:433–442

99. Bigwood R, Gruebele M, Leitner DM, Wolynes PG (1998) The vibrational energy flow transition in organic molecules: theory meets experiment. Proc Natl Acad Sci U S A 95:5960–5964
100. Gruebele M (2000) Molecular vibrational energy flow: a state space approach. Adv Chem Phys 114:193–261
101. Keshavamurthy S (2013) Scaling perspective on intramolecular vibrational energy flow: analogies, insights and challenges. Adv Chem Phys 153:43–110
102. Manikandan P, Keshavamurthy S (2014) Dynamical traps lead to the slowing down of intramolecular vibrational energy flow. Proc Natl Acad Sci U S A 111:14354–14359
103. Leitner DM (1993) Real-symmetric random matrix ensembles of Hamiltonians with partial symmetry-breaking. Phys Rev E 48:2536–2546
104. Leitner DM, Cederbaum LS (1994) Some properties of invariant random-matrix ensembles and their connection to ergodic and nonergodic Hamiltonian systems. Phys Rev E 49:114–121
105. Leitner DM, Wolynes PG (1997) Quantization of the stochastic pump model of Arnold diffusion. Phys Rev Lett 79:55–58
106. Leitner DM, Pandey HD (2015) Quantum bottlenecks and unidirectional energy flow in molecules. Ann der Phys 527:601–609
107. Leitner DM, Pandey HD (2015) Asymmetric energy flow in liquid alkylbenzenes: a computational study. J Chem Phys 143:144301
108. Fujisaki H, Stock G (2008) Dynamic treatment of vibrational energy relaxation in a heterogeneous and fluctuating environment. J Chem Phys 129(13):134110
109. Leitner DM (2001) Vibrational energy transfer in helices. Phys Rev Lett 87:188102
110. Leitner DM (2001) Vibrational energy transfer and heat conduction in a one-dimensional glass. Phys Rev B 64:094201
111. Alicki R, Leitner DM (2015) Size-dependent accuracy of nanoscale thermometers. J Phys Chem B 119:9000–9005. https://doi.org/10.1021/jp508047q
112. Leitner DM (2012) Mode damping rates in a protein chromophore. Chem Phys Lett 530:102–106
113. Leitner DM, Köppel H, Cederbaum LS (1994) Effects of symmetry breaking on spectra of chaotic Hamiltonian systems. Phys Rev Lett 73:2970–2973
114. Leitner DM, Köppel H, Cederbaum LS (1996) Statistical properties of molecular spectra and molecular dynamics: analysis of their correspondence in NO2 and C2H4+. J Chem Phys 104:434–443
115. Maisuradze GG, Yu X, Leitner DM (2007) Normal mode analysis and calculation of the cooling rates of the chromophore vibrations during isomerization of photoactive yellow protein. J Biol Phys Chem 7:25–29
116. Xu Y, Gnanasekaran R, Leitner DM (2013) The dielectric response to photoexcitation of GFP: a molecular dynamics study. Chem Phys Lett 564:78–82
117. Pandey HD, Leitner DM (2018) Small saccharides as a blanket around proteins: a computational study. J Phys Chem B 122:7277–7285. https://doi.org/10.1021/acs.jpcb.8b04632
118. Pandey HD, Leitner DM (2017) Influence of thermalization on thermal conduction through molecular junctions: computational study of PEG oligomers. J Chem Phys 147:084701
119. Buldum A, Leitner DM, Ciraci S (1999) Thermal conduction through a molecule. Europhys Lett 47:208–212
120. Leitner DM, Matsunaga Y, Li C-B, Komatsuzaki T, Shojiguchi A, Toda M (2011) Non-brownian phase space dynamics of molecules, the nature of their vibrational states, and non-RRKM kinetics. Adv Chem Phys 145:83–122
121. Komatsuzaki T, Berry RS, Leitner DM (2011) Advancing theory for kinetics and dynamics of complex, many-dimensional systems: clusters and proteins, Adv. chem. phys, vol 145. Wiley, Hoboken
122. Leitner DM, Wolynes PG (2006) Quantum theory of enhanced unimolecular reaction rates below the ergodicity threshold. Chem Phys 329:163–167
123. Leitner DM, Wolynes PG (1997) Quantum energy flow during molecular isomerization. Chem Phys Lett 280:411–418
124. Leitner DM, Levine B, Quenneville J, Martínez TJ, Wolynes PG (2003) Quantum energy flow and trans-stilbene photoisomerization: an example of a non-RRKM reaction. J Phys Chem A 107:10706–10716
125. Leitner DM (1999) Influence of quantum energy flow and localization on molecular isomerization in gas and condensed phases. Int J Quantum Chem 75:523–531
126. Nordholm S (1989) Photoisomerization of stilbene—a theoretical study of deuteration

shifts and limited internal vibrational redistribution. Chem Phys 137(1–3):109–120

127. Agbo JK, Leitner DM, Myshakin EM, Jordan KD (2007) Quantum energy flow and the kinetics of water shuttling between hydrogen bonding sites on trans-formanilide (TFA). J Chem Phys 127:064310–064311
128. Agbo JK, Leitner DM, Evans DA, Wales DJ (2005) Influence of vibrational energy flow on isomerization of flexible molecules: incorporating non-RRKM kinetics in the simulation of dipeptide isomerization. J Chem Phys 123:124304
129. Agbo JK, Jain A, Leitner DM (2010) Quantum localization, dephasing and vibrational energy flow in a trans-formanilide (TFA)-H2O complex. Chem Phys 374:111–117
130. Patra S, Keshavamurthy S (2015) Classical-quantum correspondence in a model for conformational dynamics: connecting phase space reactive islands with rare events sampling. Chem Phys Lett 634:1–10
131. Toda M (2005) Global aspects of chemical reactions in multidimensional phase space. Adv Chem Phys 130A:337–399
132. Hamm P, Lim M, Hochstrasser RM (1998) Structure of the amide I band of peptides measured by fs nonlinear-infrared spectroscopy. J Phys Chem B 102:6123–6138
133. Zhang Y, Fujisaki H, Straub JE (2009) Direct evidence for mode-specific vibrational energy relaxation from quantum time-dependent perturbation theory. 1. Five-coordinate ferrous iron porphydin model. J Chem Phys 130:025102
134. Zhang Y, Fujisaki H, Straub JE (2009) Mode specific vibrational energy relaxation of amide I and II modes in N-methylacetamide/water clusters: the intra- and inter-molecular energy transfer mechanisms. J Phys Chem A 113:3051–3060
135. Austin RH, Xie A, Lvd M, Redlich B, Lingård P-A, Frauenfelder H, Fu D (2005) Picosecond thermometer in the amide I band of myoglobin. Phys Rev Lett 94:128101
136. Peterson KA, Rella CW, Engholm JR, Schwettman HA (1999) Ultrafast vibrational dynamics of the myoglobin amide I band. J Phys Chem B 103:557–561
137. Moritsugu K, Miyashita O, Kidera A (2003) Temperature dependence of vibrational energy transfer in a protein molecule. J Phys Chem B 107:3309–3317
138. Moritsugu K, Miyashita O, Kidera A (2000) Vibrational energy transfer in a protein molecule. Phys Rev Lett 85:3970–3973
139. Newman M (2005) A measure of betweenness centrality based on random walks. Soc Networks 27:39–54
140. Dijkstra EW (1959) A note on two problems in connection with graphs. Numerische Math 1:269–271
141. Ren Z, Srajer V, Knapp JE, Royer WE (2012) Cooperative macromolecular device revealed by meta-analysis of static and time-resolved structures. Proc Natl Acad Sci U S A 109:107–112
142. Nakayama T, Kousuke Y, Orbach RL (1994) Dynamical properties of fractal networks: scaling, numerical simulations, and physical realizations. Rev Mod Phys 66:381–443
143. Reid KM, Yamato T, Leitner DM (2018) Scaling of rates of vibrational energy transfer in proteins with equilibrium dynamics and entropy. J Phys Chem B 122:9331–9339. https://doi.org/10.1021/acs.jpcb.8b07552
144. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2016) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theor Compute 11:3696–3713
145. Berendsen H, Postma J, vanGunsteren W, DiNola A, Haak J (1984) Molecular-dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
146. Meister K, Ebbinghaus S, Xu Y, Duman JG, DeVries A, Gruebele M, Leitner DM, Havenith M (2013) Long-range protein–water dynamics in hyperactive insect antifreeze proteins. Proc Natl Acad Sci U S A 110:1617–1622
147. Leitner DM, Havenith M, Gruebele M (2006) Biomolecule large amplitude motion and solvation dynamics: modeling and probes from THz to X-rays. Int Rev Phys Chem 25:553–582
148. Ebbinghaus S, Kim S-J, Heyden M, Yu X, Heugen U, Gruebele M, Leitner DM, Havenith M (2007) An extended dynamical solvation shell around proteins. Proc Natl Acad Sci U S A 104:20749–20752
149. Ebbinghaus S, Kim SJ, Heyden M, Yu X, Gruebele M, Leitner DM, Havenith M (2008) Protein sequence- and pH-dependent hydration probed by terahertz spectroscopy. J Am Chem Soc 130:2374–2375
150. Heugen U, Schwaab G, Bründermann E, Heyden M, Yu X, Leitner DM, Havenith M (2006) Solute induced retardation of water dynamics: hydration water probed directly

by THz spectroscopy. Proc Natl Acad Sci U S A 103:12301–12306

151. Heyden M, Bründermann E, Heugen U, Niehues G, Leitner DM, Havenith M (2008) The long range influence of carbohydrates on the solvation dynamics of water—answers from THz spectroscopic measurements and molecular modelling simulations. J Am Chem Soc 130:5773–5779
152. Heyden M, Havenith M (2010) Combining THz spectroscopy and MD simulations to study protein-hydration coupling. Methods 52:74–83
153. Leitner DM, Gruebele M, Havenith M (2008) Solvation dynamics of biomolecules: modeling and terahertz experiments. HFSP J 32:314–323
154. Luong TQ, Xu Y, Bründermann E, Leitner DM, Havenith M (2016) Hydrophobic collapse induces changes in the collective protein and hydration low frequency modes. Chem Phys Lett 651:1–7
155. Xu Y, Gnanasekaran R, Leitner DM (2012) Analysis of water and hydrogen bond dynamics at the surface of an antifreeze protein. J At Mol Opt Phys 2012:125071–125076
156. Meister K, Duman JG, Xu Y, DeVries AL, Leitner DM, Havenith M (2014) The role of sulfates on antifreeze protein activity. J Phys Chem B 118:7920–7924
157. Schmidt DA, Birer Ö, Funkner S, Born B, Gnanasekaran R, Schwaab G, Leitner DM, Havenith M (2009) Rattling in the cage: ions as probes of sub-picosecond water network dynamics. J Am Chem Soc 131:18512–18517
158. Xu Y, Bäumer A, Meister K, Bischak C, DeVries AL, Leitner DM, Havenith M (2016) Protein-water dynamics in antifreeze protein III activity. Chem Phys Lett 647:1–6
159. Acbas G, Niessen KA, Snell EH, Markelz AG (2014) Optical measurements of long-range protein vibrations. Nat Commun 5:3076
160. Knab JR, Chen J-Y, Markelz AG (2006) Hydration dependence of conformational dielectric relaxation of lysozyme. Biophys J 90:2576–2581
161. Yu X, Park J, Leitner DM (2003) Thermodynamics of protein hydration computed by molecular dynamics and normal modes. J Phys Chem B 107:12820–12829
162. Pandey HD, Leitner DM (2017) Thermodynamics of hydration water around an antifreeze protein: a molecular simulation study. J Phys Chem B 121:9498–9507
163. LeBard DN, Matyushov DV (2010) Ferroelectric hydration shells around proteins: electrostatics of the protein-water interface. J Phys Chem B 114:9246–9258
164. Martin DR, Matyushov DV (2012) Non-Gaussian statistics and nanosecond dynamics of electrostatic fluctuations affecting optical transitions in proteins. J Phys Chem B 116:10294–10300
165. Martin DR, Matyushov DV (2015) Dipolar nanodomains in protein hydration shells. J Phys Chem Lett 6:407–412
166. Matyushov DV (2010) Terahertz response of dipolar impurities in polar liquids: on anomalous dielectric absorption of protein solutions. Phys Rev E 81:021914
167. Heyden M, Tobias DJ (2013) Spatial dependence of protein-water collective hydrogen bond dynamics. Phys Rev Lett 111:218101
168. Laage D, Elsaesser T, Hynes JT (2017) Water dynamics in the hydration shells of biomolecules. Chem Rev 117:10694–10725
169. Xu Y, Havenith M (2015) Perspective: watching low-frequency vibrations of water in biomolecules by THz spectroscopy. J Chem Phys 143:170901
170. Wirtz H, Schäfer S, Hoberg C, Reid KM, Leitner DM, Havenith M (2018) Hydrophobic collapse of ubiquitin generates rapid protein-water motions. Biochemistry 57:3650–3657
171. Wellig S, Hamm P (2018) Solvation layer of antifreeze proteins analyzed with a Markov state model. J Phys Chem B 122:11014–11022. https://doi.org/10.1021/acs.jpcb.8b04491
172. Leitner DM, Pandey HD, Reid KM (2019) Energy transport across interfaces in biomolecular systems. J Phys Chem B 123:9507–9524
173. Reid KM, Yamato T, Leitner DM (2020) Variation of energy transfer rates across protein-water contacts with equilibrium structural fluctuations of a homodimeric hemoglobin. J Phys Chem B 124:1148–1159
174. Leitner DM, Yamato T (2020) Recent developments in the computational study of protein structural and vibrational energy dynamics. Biophysical Reviews 12:317–322
175. Leitner DM, Hyeon C, Reid KM (2020) Water-mediated biomolecular dynamics and allostery. J Chem Phys 152:240901

Chapter 5

Probing Allosteric Mechanism with Long-Range Rigidity Transmission Across Protein Networks

Adnan Sljoka

Abstract

Allosteric transmission refers to regulation of protein function at a distance. "Allostery" involves regulation and/or signal transduction induced by a perturbation event. Allostery, which has been coined the "second secret of life," is a fundamental property of most dynamics proteins. Most of critical questions surrounding allostery are largely unresolved. One of the key puzzles is to describe the physical mechanism of distant coupled conformational change. Another hot research area surrounding allostery is detection of allosteric pathways or regions (residues) in the protein that are the most critical for transmission of allosteric information. Using techniques inspired by mathematical rigidity theory and mechanical linkages, we have previously proposed a mechanistic model and description of allosteric transmission and an accompanying computational method, the Rigidity Transmission Allostery (RTA) algorithm. The RTA algorithm and method are designed to predict if mechanical perturbation of rigidity, for example, due to ligand binding, at one site of the protein can transmit and propagate across a protein structure and in turn cause a change in available conformational degrees of freedom and a change in conformation at a second distant site, equivalently resulting in allosteric transmission. The RTA algorithm is computationally very fast and can rapidly scan many unknown sites for allosteric transmission, identifying potential novel allosteric sites and quantify their allosteric effect. In this chapter we will discuss the rigidity-based mechanistic model of allosteric communication. As a case illustrative study, we will demonstrate RTA analysis on a G protein coupled receptor (GPCR) human adenosine A_{2A} receptor. Our method gives important implications and a novel prospective for general mechanistic description of allosteric communication.

Key words Allostery, Protein flexibility, Degrees of freedom, Rigidity theory, Big data, Pebble game algorithm, Molecular theorem, FIRST, Rigidity-transmission allostery, RTA algorithm

Abbreviations

DOF	Degrees of freedom
FIRST	Floppy inclusions and rigid substructure topography
GPCR	G protein coupled receptor
PDB	Protein Data Bank
RTA algorithm	Rigidity transmission allostery algorithm
TM	Transmembrane

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_5,

1 Introduction

Allosteric control of protein function refers to regulation at a distance. Allostery is one of the most powerful and predominant means of regulating protein activity and has been referred to as "the second secret of life." [1]. Allosteric regulation is a universal phenomenon that is initiated by a perturbation through binding of an effector molecule at an allosteric site that is topographically distinct and remote from orthosteric/active binding site. Binding of an effector molecule at the allosteric site(s) triggers a local conformational change that can propagate a substantial distance to cause a rearrangement and a change in conformation and dynamics at a distant functional active site, subsequently resulting in modification of protein function. Allostery is a common event in a cell and it occurs in all dynamics proteins, in RNA and DNA polymers [2–4]. Initial perturbation can arise due to covalent (i.e., phosphorylation, enzyme-substrate reaction, point mutation) and noncovalent (binding of drugs, proteins, ions, etc.) modifications at the allosteric site [2]. It is integral to the control of metabolic and signaling pathways, and it provides organisms the ability to adapt to constant changes in cellular and environmental conditions [1–4]. As allosteric interactions at a remote site lead to conformational and ultimately a change in functional site and deregulation of a protein function, allostery has direct relevance to cellular function and disease [2]. Remarkably, even after 50 years since the concept of allostery was first introduced [5], most of the critical questions surrounding allostery remain unresolved. One of the key puzzles is to provide a mechanistic description of allosteric transmission between remote sites in a protein, specifically how a structural change in conformation at an allosteric site (in some cases a subtle change) can induce a change in conformation at a distant active site. Moreover, what region in the protein is important for allosteric transmission (i.e., what are the allosteric pathways?). Since allostery is a crucial biological phenomenon for understanding biological systems, disease, and design of novel allosteric drugs, decoding the mechanism of allosteric transmission remains one of the key long-standing unsolved problems in biological sciences.

In this chapter we describe and summarize the mechanistic description and physical model for allosteric transmission called Rigidity Transmission Allostery (RTA) analysis. RTA is based on concepts in mathematical rigidity theory [6, 7] building on our initial work on mathematical models and algorithms for studying allostery [8] with further theoretical considerations in [9]. We will demonstrate the RTA analysis on a crystal structure of a GPCR receptor. RTA algorithm was recently used to predict and quantify allosteric interactions between remote sites in protein structures

and it can be extended to identify the allosteric pathways and detect potential novel allosteric sites [10–12].

In the next two subsections we give a brief review of how to predict flexibility of proteins with rigidity theory methods, focusing on method FIRST [13]. In Subheading 2 we discuss rigidity-based prediction of allosteric communication using the RTA analysis.

1.1 Analyzing Molecular/Protein Rigidity with Mathematical Rigidity Theory

To understand how proteins function including allosteric transmission requires deep knowledge of protein flexibility and its dynamics. Protein motions take place on a wide range of time scales from rapid bond-vibrations on the femtosecond range to large-amplitude collective motions occurring on milliseconds-seconds range [14, 15]. A typical protein can contain thousands of conformational degrees of freedom, whose conformational fluctuations among the structural ensemble members are rapid, transient, and result in structures that are mostly spectroscopically undistinguishable compared to the ground state [14–17]. The ultimate desire is to observe proteins move in real time at atomistic level as they accomplish their function, but despite many advances in experimental measures of dynamics and biophysical and computational methods including molecular dynamic simulations we are still far from actualizing this goal [14, 17]. The computational time needed to investigate large-scale functionally relevant motions including those of allosteric transmissions with MD simulations, even with special-purpose commodity computer clusters such as Anton, is beyond practical wide-range applications [18]. To tackle this challenge, there is a clear need to come up with alternate and fast computational methods that simplify the force fields which can still provide accurate and efficient protein flexibility predictions that are in agreement with experimental measures. Numerous advances in the mathematical rigidity theory [6–8, 19, 20] over the last 35 years have facilitated developments of several emerging technologies [8, 10–12, 16, 21–24] for fast computational predictions of both protein flexibility and their dynamics.

Rigidity theory examines the rigidity/flexibility of frameworks which are specified by geometric constraints (distances, directions, etc.) on a collection of points and rigid bodies [6, 7] which has many applications to both natural structures (molecules, crystals, etc.) and engineered structures (bridges, robots, etc.) [13, 16, 22, 23, 25]. Proteins are modeled as constrained geometric molecular frameworks (a mechanical linkage in kinematics and robotics vocabulary) consisting of atoms and an assortment of linking intermolecular forces (constraints) [7]. In a molecular framework model of a protein (in rigidity theory referred to as a *body-bar framework* [7, 26] (*see* Fig. 1a) we assume the angles between the bonds of an atom (body) are fixed allowing dihedral angles to freely rotate,

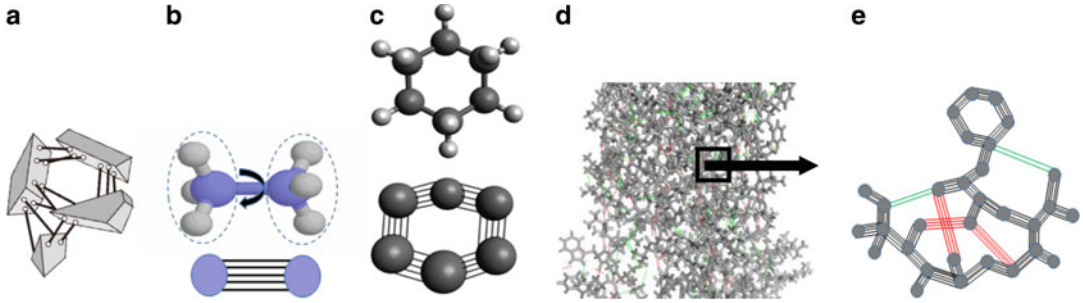

Fig. 1 (**a**) A general 3D body-bar framework composed of rigid bodies whose motioms are restricted by connecting bar constraints, where each (independent) bar removes a single DOF. (**b**) Molecular framework of ethane has a single internal DOF and can be modelled as a body-bar framework (multigraph). Each carbon atom together with its locked bonds is modeled as a fully rigid body with 6 trivial DOF and is represented as a vertex (node) and the rotatable bond between two carbon atoms as a set of five bars (edges) leaving 1 internal DOF between the two rigid bodies (i.e. $6 + 6 - 5 = 7$ DOF; 6 trivial rigid body DOF and 1 internal DOF). (**c**) A cyclohexane and its body-bar multigraph representation. (**d**) Protein structure in stick representation (body-hinge) with gray, red and green lines corresponding to covalent bonds (hinges), hydrogen bonds and hydrophobic contacts, respectively. (**e**) Body-bar multigraph representation of a protein framework

and the locked dihedral angles associated with double and peptide bonds together with non-covalent interactions impose additional constraints (bars). Each atom is treated as a fully rigid body with six trivial degrees of freedom (DOF) and rotatable bonds (hinges) as a set of five bars (edges), where each bar removes a single DOF, leaving one bond rotational DOF (*see* Fig. 1b, c). Double or peptide bonds are modeled as a set of 6 bars between the two atoms locking the rotational DOF [7]. Non-covalent interactions (hydrogen bonds, hydrophobic contacts, etc.) are modeled as a set of 1–5 bars that further restrict the protein's internal conformational DOF [16, 22]. Depending on the energy strength and persistence of a hydrogen bond, and if an ensemble of structures is available, the number of bars can be appropriately adjusted [16]. Hydrophobic contacts are modeled between any contacting pairs of carbon-carbon, carbon-sulfur, or sulfur-sulfur atoms [13]. A molecular body-bar framework is said to be rigid if every motion results in framework that is isometric to the original one (i.e., the framework only has rigid-body motions); otherwise, the framework is flexible [7]. A molecular theorem [7, 19] states that (generic) rigidity of a molecular framework is only a property of the underlying topology (i.e., graph, network), which also prescribes a necessary and sufficient mathematical counting certificate for rigidity. In other words, we only need to count the number of atoms (vertices) and bars (edges) in the body-bar graph and its distribution throughout the subgraphs to determine the rigidity of a corresponding molecular model of a protein.

1.2 Protein Rigidity/Flexibility Analysis with Method FIRST

Given a PDB structure or an ensemble of structures, the program FIRST [13] (and its various spinoff methods—Kinari, DCA, CNA and others [20]) converts the structure to a body-bar multigraph (network) model of a protein, consisting of vertices (atoms) and edges (covalent bonds, hydrogen bonds, hydrophobic contacts, and electrostatic interactions) (Fig. 1d, e). The strength of each hydrogen bond is calculated using an energy potential [13]. A user selects a hydrogen bond energy cutoff value such that all bonds weaker than this cutoff are ignored and the final constraint body-bar multigraph is obtained. FIRST then applies the pebble game algorithm [8, 27] on the multigraph which checks the combinatorial characterization of rigidity prescribed in the molecular theorem [7, 19]. The pebble game determines if each constraint (bar/edge) is "independent" (i.e., removes a DOF from the network) or is otherwise "redundant." Pebbles are synonymous with conformational degrees of freedom and a removal of a pebble indicates the inserted constraint (edge) is independent. The pebble game finally decomposes the protein into rigid clusters and flexible regions. A rigid cluster moves as a single rigid body with its trivial 6 DOF (a combination of 3 rotations and 3 translations). A typical protein normally consists of several rigid regions connected by flexible linkers (Fig. 2). Given such a decomposition of a protein into rigid and flexible connections, fast Monte-Carlo methods such as FRODA [23] (which give 100,000 speedups compared to MD simulation) go a step further and can simulate the actual protein motions and explore their dynamics. Sampling of conformational space and dynamics can be done on very large systems, such as ribosome and even viral capsids [28], and we have recently been extending these techniques and applied it on intrinsically flexible proteins which have a substantial amount of disorder and an extremely high number of internal DOF [24].

In Fig. 2a–d we have shown the output of FIRST on a human adenosine A_{2A} GPCR at several hydrogen bond energy cutoffs. Hydrogen bonds can be removed one by one (i.e., by lowering of hydrogen bond energy cutoff) in the order of increasing strength, while maintaining all other covalent and hydrophobic interactions intact, and then repeating the analysis as hydrogen bonds are removed while recalculating rigid and flexible regions. Change in rigidity can be visualized in the hydrogen bond "dilution plot" (Fig. 2e). FIRST can predict the rigid clusters and flexible connections (known as the *rigid cluster decomposition*) in less than a second on a standard PC/laptop. Many studies have demonstrated that FIRST gives accurate predictions of flexibility and rigidity in proteins that are in agreement with experiments [10, 11, 16, 22].

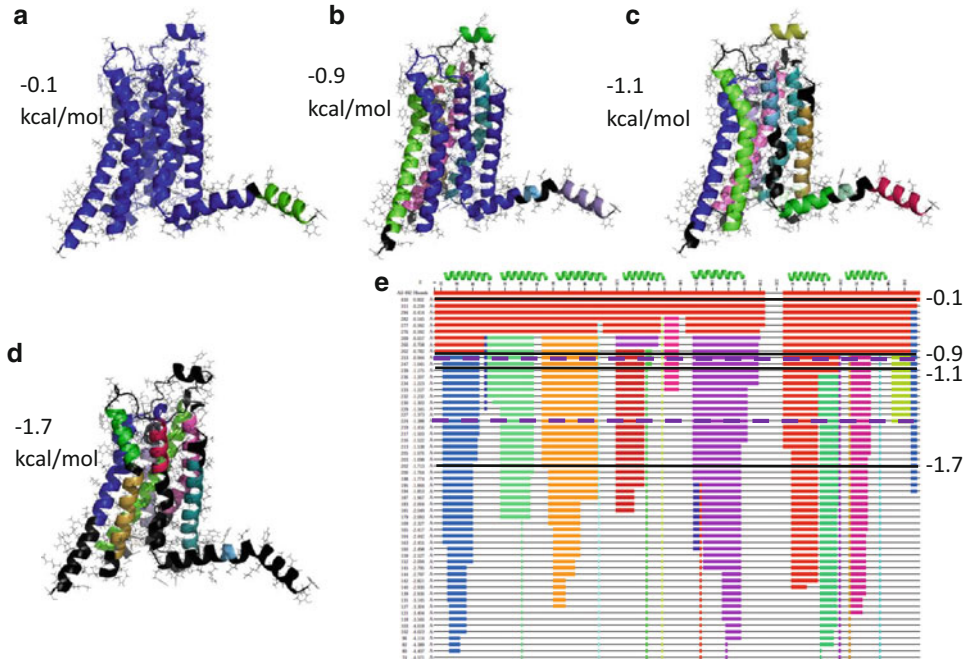

Fig. 2 Rigid cluster decomposition with program FIRST on human A_{2A} adenosine receptor (pdb 2ydo) at various hydrogen bond energy cutoffs. Blue is the largest rigid cluster, black regions are flexible parts of the protein. (**a**) At −0.1 kcal/mol hydrogen bond cutoff, the protein is mainly composed of a single large rigid cluster. At this cutoff, most hydrogen bonds are modeled in the network including very weak/transient hydrogen bonds. (**b**) As hydrogen bonds are diluted, the receptor breaks into several rigid clusters. (**c**) Individual helices are separated into rigid components. (**d**) As stronger hydrogen bonds are removed, helices become flexible. (**e**) Hydrogen bond dilution plot. Columns on the left are updated and display the hydrogen bond energy levels and total number of remaining hydrogen bonds. Corresponding energy cutoff lines are highlighted in black at −0.1, −0.9, −1.1 and −1.7 kcal/mol. Flexible regions are shown as thin black lines, with coloured blocks indicating distinct rigid clusters. Initially with inclusion of all potential hydrogen bonds, the protein is quite rigid (red block) and as hydrogen bonds are gradually broken with increasing energy, the protein decomposes into several rigid clusters, many which correspond to TM helices. Purple dashed lines indicate the start (−0.944 kcal/mol) and end of allosteric transmission (−1.387 kcal/mol) (*see* Fig. 5c)

2 Methods

2.1 Rigidity Transmission Model as a Mechanistic Description of Allostery

In rigidity transmission model of allostery, a change in rigidity induced by a binding event(s) which results in a change and transmission in DOF (i.e., a conformational change) across protein network to a remote distant site(s) gives a mechanical description for allosteric coupling between distant sites in a protein. The mathematical and mechanical model of allosteric communication is founded on our initial foundation work in mathematical allostery in rigidity theory introduced in [8]. Further mathematical properties were further considered by Whiteley et al. [9] in a special class of geometric frameworks. Recently, we have applied this

mechanistic view of allostery for deciphering allosteric coupling in enzymes, receptors, antibodies, and other protein structures [10–12, 21].

Our model of allostery is designed to predict if a mechanical perturbation of rigidity and a change in conformational DOF (mimicking a binding event) at a given site A on a protein can percolate and transmit across a protein structure and result in a change and transmission in rigidity and available conformational DOF at a second remote site B. A key step is to introduce the local perturbation of rigidity at site A by adding extra constraints (edges) to site A up to its rigidification. (Note that in the description of RTA algorithm below, no actual edges need to be added, but the same effect can be obtained, which is mathematically verified.) Upon the initial perturbation (rigidification) of site A, if this results in a reduction of conformational DOF at site B, then A "transmits degrees of freedom' (DOF) to B and the two sites are in 'rigidity-transmission communication." The maximum possible reduction in DOF in site B quantifies the strength of the allosteric transmission signal, where larger the reduction, the stronger the allosteric transmission signal. As a mechanistic description, the presence of rigidity-transmission allostery (transmission of DOF) between sites A and B can be mathematically verified to be equivalent to a statement that a change in shape (conformation) in site A (i.e., mechanically change the shape as binding might) will lead to rearrangement and change of shape and conformation of the second site B [8]. Thus, rigidity-based allostery captures the essence of coupled conformational change between distant sites inherent in allosteric communication.

In Fig. 3b, c we have illustrated the concept of DOF transmission and a change in shape propagation between two remote sites in a 2-dimensional bar and joint framework toy model, which is built with bars (rods) which fix the distance between the connecting flexible joints. This framework has a single internal DOF and responds specifically to a stimulus at a second distant site. When we introduce a subtle change in the distance between the two joints in site A (analogous to simulating ligand binding), this initial shape change propagates across the framework and results in a change in shape and conformation at the distant site B. Equivalently, fixing the distance between the end joints in A (i.e., insert a bar connecting u and v) and rigidifying site A will in turn rigidify site B, stopping the motion in B. Hence, there is a transmission of one DOF between A and B. As an analogy to allostery in a hypothetical protein, a small ligand that fits in site A can pull on the two vertices, which in turn leads to a change in conformation and a closing motion at site B, allowing site B to more likely dock its binding ligand partner (i.e., a hypothetical analogue to positive allosteric modulation).

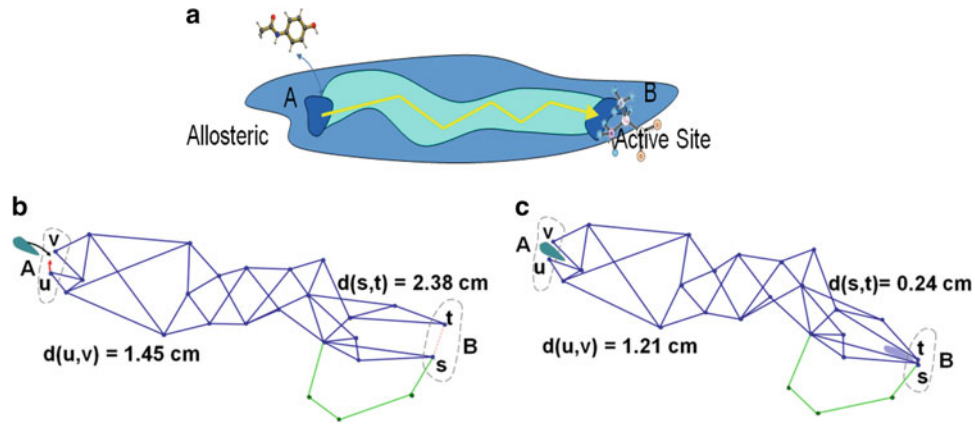

Fig. 3 (**a**) The goal of RTA analysis is to check if upon perturbation (rigidification) of site A (mimicking ligand binding): (1) are the two remote sites A and B in rigidity-transmission communication (i.e. is there a transmission of DOF from A to B—a coupled conformational change), (2) to identify the strengh of the allosteric transmission signal between A and B and (3) to find the pathway that is critical for this allosteric transmission. In (**b**) and (**c**) we illustrate allostery in a 2-dimensional bar and joint framework model and a hypothetical example of positive allosteric transmission. This framework has one non-trivial DOF (excluding green edges) and this single DOF can transmit between A and B. In other words, a change in shape in site A (i.e. moving u and v closer together using the single DOF) simulating ligand binding will propage across the framework and cause a change in shape in site B (increasing likelyhood for binding). The green edges are not relevant for this communication and would not form part of the allosteric pathway, their removal/insertion has no effect on A and B cross-talk

In rigidity-transmission allostery model, once we have added constraints in site A (up to its rigidification), the effect on B is one of the following: (1) no-change in DOF is observed at B (i.e., A and B are not in communication); (2) there is a reduction in DOF at B but some internal DOF still remain in B, or (3) all internal DOF in B are removed (i.e., B becomes rigid). The effect of a change in DOF propagating across the network to reach site B, upon initial perturbation at site A, forces the second site B to move in a different conformational space than prior to perturbation. In terms of the conformational ensemble view of protein structure, conformational selection, and energy landscapes, presence of rigidity-transmission allostery will lead to a change in shape, stabilizing certain conformations and biasing (restricting) of distribution of states that can be sampled within the conformational ensemble [29].

We now describe the RTA procedure for allostery computation in proteins.

2.2 Rigidity-Transmission Allostery (RTA) Analysis

For simplicity, we assume we are given two sites A and B on the protein of interest. It is possible that only one site is known (i.e., active site) and we test all other potential sites for allosteric communication with the active site; these details will not be discussed

Fig. 4 Overview of rigidity-transmission allostery procedure

here and will appear in forthcoming papers. To predict if A and B are in rigidity-transmission communication we apply the following procedure (*see* Fig. 4 for a schematic overview):

2.2.1 Preparation and Preprocessing

1. Input a PDB structure file (or an ensemble of structures, which can be handled as described in [11]) and add missing hydrogen atoms (typically this is the case with crystal structures; here we use the WHAT IF server (http://swift.cmbi.ru.nl/servers/html/htopo.html) to add hydrogens).
2. Create a corresponding molecular body-bar multigraph G_h and select a hydrogen bond energy cutoff h.
3. Select two sites A and B (in graph theory terminology, A and B are disjoint vertex induced subgraphs in G_h). Sites A and B can be a collection of residues, ligands, or a specified set of atoms.

2.2.2 Task

Check if rigidification of A results in a transmission (reduction) in conformational DOF at B. If there is transmission, calculate the maximum DOF transmission using the following RTA algorithm steps:

1. Calculate the available DOF in site B, call it DOF^B: (DOF^B is the number of independent edges that would need to be added to B that results in its rigidification. We can also run the pebble game algorithm on G_h and count the maximum number of pebbles on B.)
2. Perturb rigidity of A (rigidify A by adding maximum number of independent edges within A).
3. Re-calculate the available DOF in B, call it $DOF^B{}_{Aperturbed}$.

2.2.3 Output

Transmission of DOF from A to B is: $DOF^{AB} = DOF^B - DOF^B{}_{Aperturbed}$.

(i.e., maximum reduction in DOF at B given perturbation/rigidification of A).

When $DOF^{AB} > 0$, then A and B are in allosteric communication.

2.2.4 Remark1

Rigidity-transmission allosteric communication has a symmetric property; in other words the effect of perturbing rigidity of site A on site B and the maximum amount of DOF transmission is identical if we perturb B and observe the effect on site A. That is $DOF^{AB} = DOF^{BA}$.

2.2.5 Remark2

Transmission is possible if A and B have some internal flexibility (i.e., DOF^A and $DOF^B > 0$).

2.2.6 Remark3

Uniqueness and correctness of the DOF^{AB} counts extracted from RTA algorithm, the pebble game extensions that allow fast computations of counts in step 1 and 3, and the relevant region detection algorithm for detection of allosteric are mathematically verified [8].

Extensions of this work (to appear) will show how to accurately map out the pathways that correlate with NMR experimental measures for probing allosteric crosstalk.

2.3 Case Study: GPCR

We will illustrate the RTA method on human adenosine A_{2A} receptor, a G protein coupled receptor (GPCR). GPCRs are the largest class of receptors in the human genome [29–34]. The rhodopsin family of G protein coupled receptors (GPCRs), also known as family A GPCR, represents over 80% of all GPCRs. Humans have over 800 unique GPCRs, which are characterized by the same underlying topology consisting of 7-transmembrane alpha-helices. GPCRs mediate most transmembrane signal transduction by responding to an enormous variety of extracellular stimuli (drugs, hormones, neurotransmitters, ions, proteins, etc.) as well as senses of sight, olfaction, and taste. GPCRs also play an important role in disease and drug discovery with around 50% of all modern medicinal drugs targeting GPCRs [2, 30, 31, 34].

GPCRs are typically regulated by extracellular ligands called agonists. Agonists and inactivating ligands antagonists or inverse agonists usually bind at the similar location at the extracellular region (i.e., orthorsteric pocket) of the receptor and activation can be further increased or decreased through interactions with allosteric modulators that bind at different sites from the orthorsteric site and also residue specific mutations [30, 31]. Agonist binding induces a subtle conformational change within the binding pocket, which causes relative movement of α-helices and a subsequent larger change in conformation at the intracellular side of the receptor [31, 32]. This enables activation of GPCR and binding to its G protein partner, leading to exchange of GDP and GTP, dissociation of the G protein into an α-subunit and a β-γ-subunit and subsequent activation of additional downstream partners.

A significant movement and conformational change at the cytoplasmic end of TM helix 6 is believed to be central in GPCR activation together with smaller rearrangement of TM3, TM5 and TM7 [30–32]. GPCRs are naturally allosteric as orthrosteric site and G protein binding region crosstalk must travers over large distance, spanning the TM region. As is the case with many dynamics proteins, GPCRs do not function through simple on and off switches. GPCRs are highly dynamic and can adopt a multiple of conformational ensemble states which are normally categorized as:

inactive, active-like functional states, fully active state with G-protein and intermediate states linking these states. Agonist binding generally tends to shift the conformational ensemble that closer resembles the active-like states [29, 35]. A key unresolved mystery is how GPCRs transmit the allosteric signal across the TM region leading to activation. In particular, how does binding of certain ligands (agonists, partial agonist, positive allosteric modulators, etc.) trigger an allosteric transmission and the necessary conformational change for activation at the intracellular part of the receptor, while other ligands such as antagonists do not produce this effect. The mechanism that controls ligand binding and GPCR activation is extremely complex, and this puzzle is a major research interest with big implications to design of novel therapeutics [30, 34].

Adenosine receptor is a prototypical family A GPCR and probing how it functions and transmits allosteric signals across the TM region is critical for deepening our overall understanding of GPCR activation mechanism. The A_{2A} receptor plays an important role in regulating myocardial oxygen consumption, coronary blood flow, and is a drug target for multitude of disorders (inflammation, insomnia, Parkinson's disease, cancer, diabetes, infectious diseases, and neuronal defect disorders) [31, 35, 36].

To provide insight into potential allosteric mechanism activation in GPCRs, we have applied the RTA algorithm on several structures of human adenosine A_{2A} receptor (Fig. 5). We defined site A as the orthosteric site (here taken to be all atoms and bonds that are interacting with the agonist (or antagonist)) and site B chosen as residues 230 and 291 at the intracellular side where the receptor interacts with the G protein (Fig. 5b). Starting with four crystal structures of A_{2A} receptor in the presence of different ligands, RTA algorithm was performed for all hydrogen bond energy cutoffs h (*see* Subheading 3.2) in increment steps of 0.01 kcal/mol and DOF transmission (DOF^{AB}) is calculated for each cutoff h. Results are shown in Fig. 5c.

The RTA algorithm predicts that in all three agonist-bound (active-like) structures, perturbation of rigidity at the orthorsteric site will transmit across the receptor, and in turn induce a change in conformational DOF at a remote G protein binding region. The addition of agonist allosterically restricts the overall available DOF at the G-protein binding region, and in terms of conformational selection, agonist binding will bias the receptor to more often sample the conformational state(s) increasing the likelihood for GPCR activation and interaction with G protein [29]. On the other hand, in the inactive structure with a bound antagonist, no DOF transmission occurs; equivalently no allosteric transmission is induced. This analysis suggests that transmissions of rigidity and DOF upon binding of agonist are important for facilitating structural and conformational changes at G-protein binding region, and

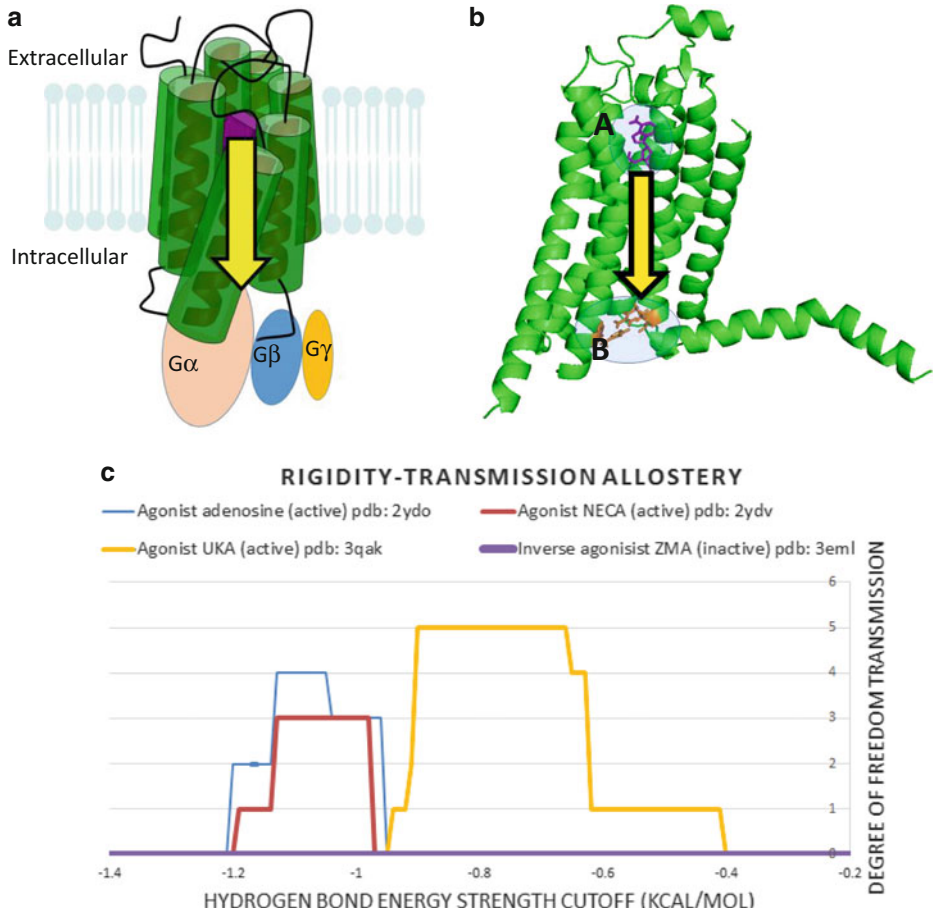

Fig. 5 (**a**) Schematic representation of a GPCR with a G-protein, showing a bound agonist (purple block) at orthosteric site, which leads to conformational change and rearangement of TM helices, transmitting the signal across the TM region, allosterically activating the receptor. (**b**) RTA analysis of adenosine A2A receptor was tested between sites A (orthosteric site) and site B (G-protein binding regions). (**c**) Plot of transmission of DOF as a function of energy cutoff in four different A2A receptor crystal structures. In all three active-like structures bound to agonists, transmission of DOF occurs, in the inactive state transmission of DOF is not seen. When the cutoff is close to 0 kcal/mol no transmission is possible as the whole protein is rigid including sites A and B (*see* also Fig. 2e). As cutoff is lowered, more hydrogen bonds break, the protein becomes less rigid and eventually allosteric transmission starts (in agonist bound structures). Transmission of DOF continues for some range of cutoffs and stops once a significant portion of hydrogen bonds have been diluted

ultimate activation of a GPCR and binding of G protein. On the other hand, antagonist binding prevented the transmission of DOF and subsequent signal propagation across the TM region. This analysis points to the role of rigidity-transmission communication as a mechanistically property of allosteric control of A2A receptor. Similar analysis was performed to detect allosteric sites and describe the role of calcium and magnesium as positive allosteric modulators or A2A receptor [11].

For allosteric communication to be effective and transmit DOF and change in shape across long-range distances across protein networks, the delicate balance between rigidity and flexibility is critical. Some large rigid components (i.e., helices which are great at transmissions over large distances) and connecting flexible regions are needed to observe long-distance propagation. We can see this by observing the range of energy cutoffs where transmission occurs and the corresponding rigid cluster decomposition. We see no allosteric transmission when the protein is either overly rigid or overly flexible (*see* Fig. 2a, d).

The two agonists adenosine and NECA are structurally very similar, and interestingly produce a very similar DOF transmission allostery profiles. Previous studies have shown that the two configurations of adenosine- and NECA-bound crystal structures are in a partially active state [31]. On the other hand, the authors in [32] report that the UKA-bound agonist crystal structure is in active state conformation (fully active state is reached in presence of G-protein). Moreover, UKA is a stronger agonist of the three agonists considered here. This may point to why adenosine- and NECA-bound structures transmit DOF at an almost identical hydrogen bond energy range, whereas UKA-bound receptor is more effective in allosteric transmission as it transmits more DOF and also at an earlier and wider range of energy cutoffs.

3 Notes and Conclusions

The progress over the last 20 years in the field of mathematical rigidity theory has opened up a number of exciting avenues for analyzing the close relationship between protein function, flexibility, and dynamics. A straightforward method that describes how allosteric signals are transmitted across protein structures and describes a mechanistic insight into allosteric propagation has been previously difficult to design and conceptualize. Our novel model of allosteric communication via transmission of degrees of freedom across protein networks and RTA analysis offers a new window to study the allosteric cross-talk between remotes sites in proteins. In this initial methodology expose, we have shown how RTA analysis can be a powerful tool for probing allostery which also provides a strong case for a mechanistic interpretation of mysterious allosteric transmission and regulation. RTA analysis was recently applied on a bacterial homodimeric enzyme fluoracetate dehalogenase where we predicted and accurately demonstrated the presence of physical allosteric pathways between the two protomers as a key functional control of the enzyme catalysis, which is closely supported and validated by experimental data [10] with other applications in detection of allosteric sites in GPCRs and in epitope mapping. Forthcoming work (to appear) will reveal further novel

uses, improvements, and extensions of the methodologies described in this chapter that solidify the RTA analysis and further validation with experimental data. A remarkable strength of the RTA procedure is that we can detect and attain information about complex allosteric communication using only a single 3D snapshot (coordinates) without doing long, complex, and expensive simulations. Due to the speed of the underlying algorithms, RTA procedure is well suited for high throughput allostery analysis. This should eventually allow us to obtain a better understanding of allostery and functionally important features in protein signaling and ultimately have tools to tackle complicated signaling events in the cell.

References

1. Fenton AW (2008) Allostery: an illustrated definition for the "second secret of life". Trends Biochem Sci 33:420–425
2. Nussinov R, Tsai CJ (2013) Allostery in disease and drug discovery. Cell 153(2):293–305
3. Liu J, Nussinov R (2016) Allostery: an overview of its history, concepts, methods, and applications. PLoS Comput Biol 12(6): e1004966
4. Gunasekaran K, Ma M, Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? Proteins 57:433443
5. Changeux JP, Jacob F, Monod J (1963) Allosteric proteins and cellular control systems. J Mol Biol 6:306–329
6. Whiteley W (1996) Some Matroids from discrete applied geometry. In: Bonin J, Oxley J, Servatius B (eds) Matroid theory, volume 197 of Contemp. Math. American Mathematical Society, Providence, pp 171–311
7. Whiteley W (2005) Counting out to the flexibility of molecules. Phys Biol 2:S116–S126
8. Sljoka A (2012) Algorithms in rigidity theory with applications to protein flexibility and mechanical linkages. PhD thesis, York University, Toronto
9. Finbow-Singh W, Whiteley W (2013) Isostatic block and hole frameworks. Siam J Discr Math 27:991–1020
10. Kim TH, Mehrabi P, Ren A, Sljoka A, Ing C, Bezginov A, Ye LB, Pomes R, Prosser RS, Pai EF (2017) The role of dimer asymmetry and protomer dynamics in enzyme catalysis. Science 355:262–U287
11. Ye L, Neale C, Sljoka A, Pichugin D, Tsuchimura N, Sunahara R, Prosser S et al (2018) Bidirectional regulation of the A2A adenosine G protein-coupled receptor by physiological cations. Nat Commun 9:1372
12. Deng B, Zhu S, Macklin AM, Xu J, Lento C, Sljoka A, Wilson D (2017) Suppressing allostery in epitope mapping experiments using millisecond hydrogen/deuterium exchange mass spectrometry. MAbs 1:10
13. Kuhn LA, Rader DJ, Thorpe MF (2001) Protein exibility predictions using graph theory. Proteins 44:150–165
14. Wildman H, Kern KD (2007) Dynamic personalities of proteins. Nature 450:964–972
15. Lewandowski JR, Halse ME, Blackledge M, Emsley L (2015) Direct observation of hierarchical protein dynamics. Science 348 (6234):578–581
16. Sljoka A, Wilson D (2013) Probing protein ensemble rigidity and predictions of hydrogen-deuterium exchange. Phys Biol 10:056013
17. Hartl FU, Hayer-Hartl M (2009) Converging concepts of protein folding in vitro and in vivo. Nat Struct Mol Biol 16:574–581
18. Shaw DE et al (2008) Anton, a special-purpose machine for molecular dynamics. Commun ACM 51(7):9197
19. Katoh N, Tanigawa S (2011) A proof of the molecular conjecture. Discrete Comput Geom 45:647–700
20. Schulze B, Sljoka A, Whiteley W (2014) How does symmetry impact the flexibility of proteins? Philos Transact Royal Soc A 372:20120041
21. Jeliazkov JR, Sljoka A, Kuroda D, Tsuchimura N, Katoh N, Tsumoto K, Gray JJ (2018) Repertoire analysis of antibody CDR-H3 loops suggests affinity maturation does not typically result in rigidification. Front Immunol 9:413
22. Hermans SMA et al (2017) Rigidity theory for biomolecules: concepts, software, and

applications. Wiley Interdiscip Rev Comput Mol Sci 7(4):e1311

23. Wells SA, Menor S, Hespenheide BM, Thorpe MF (2005) Constrained geometric simulation of difusive motion in proteins. Phys Biol 2: S12736
24. Zhu S, Shala A, Bezginov A, Sljoka A, Audette G, Wilson D (2015) Hyperphosphorylation of intrinsically disordered tau protein induces an amyloidogenic shift in its conformational ensemble. PLoS One 10(3):e0120416
25. Sljoka A, Shai O, Whiteley W (2011) Checking mobility and decomposition of linkages via pebble game algorithm. In: Proceedings of the ASME 2011 international design engineering technical conferences and computers and information in engineering conference, IDETC/CIE 2011
26. Tay TS (1984) Rigidity of multigraphs I: linking rigid bodies in n-space. J Combinat Theor Ser B 26:95–112
27. Lee A, Streinu I (2008) Pebble game algorithms and sparse graphs. Discret Math 308 (1425):1437
28. Hespenheide BM, Jacobs DJ, Thorpe MF (2004) Structural rigidity in the capsid assembly of cowpea chlorotic mottle virus. J Phys 16: S5055–S5064
29. Prosser RS, Ye L, Pandey A, Orazietti A (2017) Activation processes in ligand-activated G protein-coupled receptors: a case study of the adenosine A_{2A} receptor. BioEssays 39:1700072
30. Leach K, Sexton PM, Christopoulos A (2007) Allosteric GPCR modulators: taking advantage of permissive receptor pharmacology. Trends Pharmacol Sci 28(8):382–389
31. Lebon G, Warne T, Edwards PC, Bennett K, Langmead CJ, Leslie AGW, Tate CG (2011) Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. Nature 474:521
32. Xu F, Wu H, Katritch V, Han GW, Jacobson KA, Gao ZG, Cherezov V, Stevens RC (2011) Structure of an agonist-bound human A2A adenosine receptor. Science 332:322
33. Staus DP et al (2016) Allosteric nanobodies reveal the dynamic range and diverse mechanisms of G-protein-coupled receptor activation. Nature 535(7612):448–452
34. Tehan BG, Bortolato A, Blaney FE, Weir MP, Mason JS (2014) Unifying family A GPCR theories of activation. Pharmacol Ther 143 (1):51–60
35. Ye L et al (2016) Activation of the A_{2A} adenosine G-protein-coupled receptor by conformational selection. Nature 533(7602):265–268
36. Chen JF (2014) Adenosine receptor control of cognition in normal and disease. In: Mori A (ed) Adenosine receptors in neurology and psychiatry, International review of neurobiology, vol 119. Academic, Cambridge, pp 257–307

Chapter 6

Protein Assembly: Defining the Strength of Protein-Protein Interactions Coupling the Theory with Experiments

Giampiero Mei, Almerinda Di Venere, Luisa Di Paola, and Alessandro Finazzi Agrò

Abstract

In this paper we report a procedure to analyze protein homodimer interfaces.

We approached the problem by means of a topological methodology. In particular, we analyzed the subunits interface of about 50 homodimers and we have defined a few parameters that allow to organize these proteins in six different classes. The main characteristics of each class of homodimers have been discussed also taking into account their stabilization energy, as reported in the literature from the experimental measurements. A paradigmatic example for each class has been reported and a graphical representation proposed in order to better explain the meaning of the parameters chosen.

Key words Dimeric interface, Protein structure, Homodimers, Protein-protein interaction, Protein topology

1 Introduction

The peculiarity of allosteric proteins and enzymes resides in the mechanism of their regulation, a process that requires the propagation through long distances of a mechanical stress produced in a limited region of the macromolecule structure. Such mechanical stress is typically induced by the binding to the polypeptidic chain of molecules that are generally very small in size, as compared to the overall protein dimensions. It is therefore obvious that this "machinery" requires a concerted movement of the protein domains and, in fact, flexibility and cooperativity are the main features that characterize the network of amino acids involved in the fine, complex regulation of allosteric enzymes.

As known, what confers a protein its specific functional properties is its tri-dimensional shape, which is dictated by the sequence of its amino acids and obtained through the so-called folding mechanism. This is even more so if a quality as allostery is needed: the propagation of local changes, produced by the modulator binding

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_6,

process, requires an appropriate scaffolding to produce long-range effects. Studying the connections between folding and functional cooperation is therefore a crucial step that must be faced to unravel the many mysteries of the world of allosteric enzymes [1]. A very common feature of these proteins is their high propensity to form oligomers, and in particular dimers and tetramers. Indeed, despite a quaternary structure is not mandatory [2], most of the enzymes and proteins that play a crucial role in the regulation of metabolic pathways, signaling, and metabolites transportation are allosteric oligomers.

But which is the basic link between allostery and folding? Since the so-called "induced-fit" hypothesis was introduced [3], it was clear that the balance between conformational stability and flexibility is critical for all enzymes and many proteins, too. Obviously, in the case of oligomers the free energy of folding/unfolding also depends on the presence of the subunits interface, whose quaternary interactions are, in several cases, the main driving force for folding and stabilization of these complex proteins.

In oligomeric enzymes characterized by allosteric properties the inter-subunits interactions accomplish another fundamental task: they transmit the mechanical stress produced by ligand binding in one subunit to distal sites belonging to another subunit. Studying what happens at such interfaces is therefore important to characterize how the propagation of signals occurs, from regulatory to active sites. Since the simplest oligomers are those obtained by two identical subunits, discussing the topology of homodimeric interfaces is paradigmatic for more complex kinds of subunits association. Indeed, in a previous of paper [4] we suggested that correlations between the size, the sequence, and the quaternary structure of homodimers might be easily found taking into account a few structural parameters that can be obtained directly from the PDB files deposited in the protein data banks. A further topological analysis through protein contact networks [5] allowed to find a direct correspondence between the experimental folding energy of dimers and the roughness of the subunits interface. It was demonstrated that the topological analysis has two advantages: (1) it drastically reduces the number of descriptors of oligomer stability; (2) it allows predictions on the role played by the interface, independently on the kind of amino acid involved in quaternary interactions.

Here we describe an easy procedure that can be usefully applied to classify any homodimeric structure on the basis of its topological features. In particular, analyzing more than 50 crystallographic structures (Table 1), we have identified six groups of homodimers (Table 2), whose characteristics include the stabilization energy obtained in equilibrium unfolding measurements and the tri-dimensional features of the two chains at the interface (such as the "roughness" of the contact area due to the presence of

Table 1
List of proteins analyzed in this study

PDB code	Protein	$\Delta G_{unfolding}$ kcal/mol	Number of a. a.	Class
1rop	COLE1 ROP protein	17.2	63	1α
1ety	*E. coli* factor for inversion stimulation	13.7	93	1α
1mul	*E. coli* Hu Alpha2 protein	8.3	90	1α
2cpg	Transcriptional repressor COPG	13.4	45	1α
1a7w	Histone HMFB	14.1	68	1α
1puc	Cell-cycle control protein, P13SUC1	20.0	102	1α
2zta	GCN4 leucine zipper	9.0	31	1α
1wrp	DNA-binding domains of TRP repressor	18.8	105	1α
1buo	BTB domain from PLZF	12.8	125	1α
1gta	Glutathione S-transferase	26.0	218	1β
1g6w	Prion protein URE2	49.0	257	1β
1hnb	Glutathione transferase GSTM2–2	42.0	218	1β
1m9a	Glutathione S-transferase with S-hexylglutathione	26.0	216	1β
1cmi	Calmodulin	23.4	89	1β
5cro	CRO repressor protein	11.2	62	2α
1a8g	HIV-1 protease	14.0	99	2α
1siv	SIV protease	13.0	99	2α
1aam	Aspartate aminotransferase R292D	15.9	396	2α
1ohv	4-Aminobutyrate-aminotransferase		461	2α
1xra	S-Adenosylmethionine synthetase		383	2α
1oho	Ketosteroid isomerase Y16F/D40N mutant	22.0	127	2α
1a43	HIV-1 CAPSID protein dimerization domain	21.0	72	2α
1bet	Nerve growth factor	19.3	116	2α
1d1l	CRO-F58W mutant	11.9	61	2α
1lj9	Transcriptional regulator SLYA		145	2α
1b8k	Neurotrophin-3	22.7	122	2α
1qll	Piratoxin-II (PRTX-II) - A K49 PLA2	24.5	121	2β
2gsr	PI glutathione S-transferase	25.2	207	2β
1hti	Human triosephosphate isomerase	19.4	248	2β
1tyd	Tyrosyl-tRNA synthetase	41.7	319	2β

(continued)

Table 1
(continued)

PDB code	Protein	$\Delta G_{unfolding}$ kcal/mol	Number of a. a.	Class
1pkw	GST A1–1	26.8	222	2β
1ypi	Yeast triosephosphate isomerase	24.7	248	2β
2f83	Human coagulation factor XI zymogen	14.7	604	2β
1i5z	CRP-CAMP	26.4	206	2β
1spd	Human CU,ZN superoxide dismutase	28.6	154	3α
1beb	Bovine beta-lactoglobulin	12.0	162	3α
3hzd	Bothropstoxin-I	17.0	133	3α
1dfx	Desulfoferrodoxin	34.6	125	3α
1a7g	E2 DNA-binding	9.8	82	3α
1cp3	Apopain	25.8	277	3α
1psc	Phosphotriesterase	40.4	329	3α
1vqb	Gene V protein	16.3	87	3α
1yai	Bacterial CU,ZN superoxide dismutase	22.7	173	3α
3ssi	Proteinase inhibitor SSI	6.03	113	3α
1cz3	Dihydrofolate reductase	34.0	168	3α
1aoz	Ascorbate oxidase	17.0	282	3β
2fsf	SEC-A	22.5	408	3β
1mt5	Fatty acid amide hydrolase	15.5	299	3β
2tdm	Thymidylate synthase	27.8	250	3β
1run	Activator protein (CAP)	18.6	127	3β

"loops"). A specific dimeric structure has been chosen to represent each class as a practical example and a table summarizing the average values of the descriptors provided, together with confidence tests. General comments are also added in the conclusion section.

2 Methods

2.1 Preparing and Analyzing the PDB Files of Homodimers

Since the crystallographic structure might not correspond to the effective in vivo/ex vivo tri-dimensional assembly, a careful inspection of the PDB file is recommended.

Table 2
Calculation of characteristic parameters of protein dimeric interfaces

	SII[a]	Q/R[a]	IAR[a]	$\Delta G_{unfolding}/N$ (kcal/mol)	Q_ν[a]	N[b]
Class 1α	0.06 ± 0.01	0.82 ± 0.06	0.93 ± 0.04	0.21 ± 0.03	0.70 ± 0.08	75 ± 10
Class 1β	0.05 ± 0.01	0.72 ± 0.04	0.42 ± 0.06	0.18 ± 0.03	0.22 ± 0.02	200 ± 29
Class 2α	0.10 ± 0.01	0.48 ± 0.03	0.90 ± 0.02	0.17 ± 0.02	0.22 ± 0.03	179 ± 42
Class 2β	0.12 ± 0.01	0.57 ± 0.02	0.38 ± 0.05	0.11 ± 0.02	0.12 ± 0.02	272 ± 51
Class 3α	0.25 ± 0.02	0.32 ± 0.03	0.85 ± 0.03	0.14 ± 0.02	0.09 ± 0.02	164 ± 23
Class 3β	0.27 ± 0.08	0.25 ± 0.07	0.58 ± 0.06	0.05 ± 0.01	0.05 ± 0.02	467 ± 91

[a]SII, Q/R and IAR are defined as described in Subheading 2.2
[b]N is the average number of a.a. of proteins belonging to each class

In particular:

1. Control in the literature which is the functional quaternary structure of that specific protein/enzyme;
2. Once the crystallographic file has been uploaded in the protein data bank server (http://www.rcsb.org/pdb/home/home.do), check whether the preferential quaternary structure composition corresponds to that observed for biological function (step 1);
3. Check the number of a.a. included in the PDB file: sometimes crystallization is possible only for a segment and not for the entire, natural polypeptide length;
4. Use the PISA software to analyze the dimer interface (http://www.ebi.ac.uk/pdbe/pisa/). This interactive software (provided in the Protein Data Bank server) checks in the PDB file for the presence of multiple chains, through the "assemblies" option. In the case that more than two chains are present, it provides information on the best combination to form the dimer: for instance, if four homologous chains are present (say A, B, C, and D), and the couple AC scores 1, it means that AC is the most probable (and thus reliable) quaternary structure representing the real protein.
5. Download the file that at **step 4** the PISA routine recognized as the most reliable (in the example mentioned, AC).
6. Enter in "Details" to get the list of a.a. lying at the interface and their position in the primary structure. In this section of the program other interesting parameters are listed, such as the overall area of the protein surface, the area buried upon dimerization, the number of a.a. at the interface, etc.

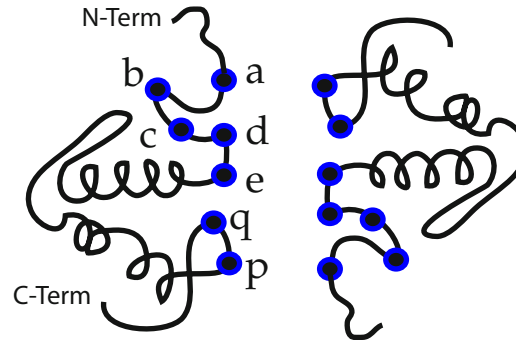

Fig. 1 Schematic representation of a homodimeric protein with interface amino acids represented by circles

2.2 Definition and Calculation of Characteristic Parameters

The main parameters used in the present classification of dimeric interfaces are:

- N, the total number of a.a. in each chain;
- R, the section (in number of a.a.) of the polypeptide chain between the first and last residues involved in quaternary interactions. Thus, for monomers $R = 0$, while in the special case in which the first and last a.a. of each subunits are lying at the interface, $R \equiv N$.
- The Interface A.A. Range, IAR, defined as $\text{IAR} = \frac{R}{N}$, so that $0 < IAR \leq 1$. This parameter indicates the percentage of a.a. directly involved in quaternary interactions.

Others important descriptors are defined in the following protocol based on the example reported in Fig. 1, in which a hypothetic homodimer is sketched. Let us assume that the length of each chain is 100 and that the positions of residues **a** and **q** (i.e., the first and last a.a. lying at the interface) are 5 and 85.

1. Calculate R and IAR. In the case of Fig. 1: $R = 85 - 5 + 1 = 81$ and $\text{IAR} = 81/100 = 0.81$.
2. Calculate the generalized number of a.a. at the interface, Q, defined as the number of consecutive a.a. (less than 5 position far apart from each other) involved in interaction with the other subunit. This parameter is characterized by a slightly larger value with respect to that obtained with the PISA software, since it consider "at interface" also those residues that are not exactly at the interface but that are lying in small loops (less than 5 a.a. in length, *see* **Note 1**) between two residues at the interface. For instance, in the example of Fig. 1 the number of a.a. involved in direct inter-subunit contacts is $n = 5$ (and in particular residues **a**, **d**, **e**, **p**, **q**) while $Q = 7$ because it includes also residues **b** and **c**.

3. Calculate Q/R. This quantity is a raw evaluation of the interface "roughness": the more Q/R is close to 1, the less is the length of the loops formed in each subunit with respect to the dimeric contact surface. In particular when $Q = 1$ all the residues are located at the interface.
4. Calculate Q_n defined as: $Q_n = \frac{Q}{R}\frac{Q}{N}$. This parameter takes simultaneously into account the presence of loops (Q/R) and the percentage of a.a. that are "at" or "close to" the dimeric interface (*see* **Note 2**).
5. Calculate the Squared Loops Length, Sll, defined as the sum of squared distances (in number of a.a. in the sequence) between two consecutive residues involved in quaternary interactions, weighted by the total number of amino acids. In the example of Fig. 1: $\mathrm{Sll} = \frac{\bar{ad}^2+\bar{de}^2+\bar{ep}^2+\bar{pq}^2}{N^2}$

where $\bar{ad}^2 = 9; \bar{de}^2 = 1$, etc.

2.3 Assigning Dimeric Structures to the Different Classes

The different typologies of dimers may be ordered on the basis of four of the main parameters introduced in Subheading 2.2. The assignment of 50 PDB files (*see* **Note 3**) to the six classes of dimers has been performed according to the following procedure.

1. Identify those proteins that do not contain significant loops at the interface: as shown in Table 1, these dimers are characterized by the smallest Sll and the largest Q/R values (*see* **Note 4**). The structural meaning of such values is shown in the examples of Fig. 2.
2. Within the same group identified in **step 1**, two possibilities may exist: (1) all (or a large part of) the a.a. are located at the subunits interface; (2) only a segment of the polypeptide chain is involved in quaternary interactions. These two situations may be easily discriminated examining the IAR value. When IAR > 0.8, the shape of the dimer looks like a flat ellipsoid (class 1α) and a typical example is that represented by the structure of 1rop (Fig. 2, left). On the contrary, a small IAR (< 0.6) indicates a more prolate structure, in which a considerable fraction of a.a. are far away from the interface. Interestingly, this subgroup of proteins (class 1β) is characterized by a larger size with respect to class 1α.
3. Identify subclasses 2α and 2β: in these cases, typically Sll ≥ 0.1 and $Q/R \geq 0.5$ (Table 3 and Fig. 3), diagnostic of loops of intermediate length (30–50 a.a.). As a consequence, class 2 includes dimers with a larger size, especially for low IAR values (class 2β).
4. In the case Sll > 0.25 and $Q/R \leq 0.3$ the dimer must be assigned to class 3, which is characterized by very large loops (Fig. 4). Due to this feature, both subgroups 3α and 3β display a small contact interface with respect to the overall shape and size of the macromolecule.

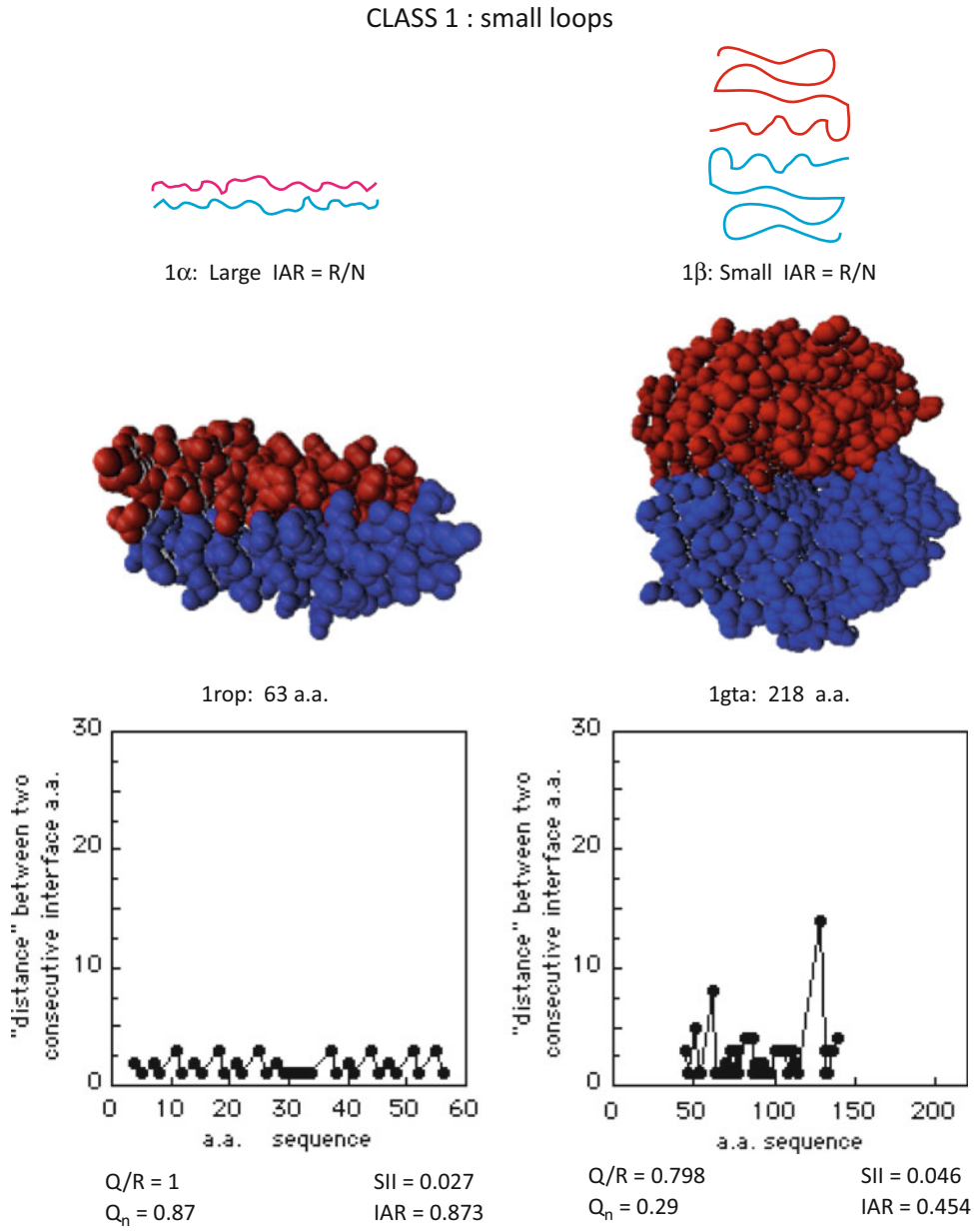

Fig. 2 Characteristics and examples of class 1 dimers

2.4 Insights on the Folding Strategies of Homodimers

A fourth parameter that has not been taken into account so far is Q_n. A protein contact network approach (*see* Chapter 2), performed on a restricted but interesting group of homodimers, has demonstrated that Q_n correlates with the size of the interface and with the intersubunit interface strength, evaluated theoretically as the difference between the graph energy of the dimer and that of the two isolated monomers [5]. Q_n therefore represents a sort of dimeric

Table 3
Calculation of diameric structure and a table summarizing the average values of the descriptors

	SII	Q/R	IAR	$\Delta G_{unfolding}/N$	Q_ν	N
Class 1α	0.06 ± 0.01	0.82 ± 0.06	0.93 ± 0.04	0.21 ± 0.03	0.70 ± 0.08	75 ± 10
Class 1β	0.05 ± 0.01	0.72 ± 0.04	0.42 ± 0.06	0.18 ± 0.03	0.22 ± 0.02	200 ± 29
Class 2α	0.10 ± 0.01	0.48 ± 0.03	0.90 ± 0.02	0.17 ± 0.02	0.22 ± 0.03	179 ± 42
Class 2β	0.12 ± 0.01	0.57 ± 0.02	0.38 ± 0.05	0.11 ± 0.02	0.12 ± 0.02	272 ± 51
Class 3α	0.25 ± 0.02	0.32 ± 0.03	0.85 ± 0.03	0.14 ± 0.02	0.09 ± 0.02	164 ± 23
Class 3β	0.27 ± 0.08	0.25 ± 0.07	0.58 ± 0.06	0.05 ± 0.01	0.05 ± 0.02	467 ± 91

Class 1α: 1ety, 1mul, 1rop, 2cpg, 1a7w, 1puc, 2zta, 1wrp, 1buo
Class 1β: 1g6w, 1gta, 1hnb, 1m9a, 1cmi
Class 2α: 5cro, 1a8g, 1siv, 1aam, 1ohv, 1xra, 1oho, 1a43, 1bet, 1d1l, 1lj9, 1b8k
Class 2β: 1qll, 2gsr, 1hti, 1tyd, 1pkw, 1ypi, 2f83, 1i5z
Class 3α: 1yai, 1spd, 1beb, 1cp3, 1psc, 3hzd, 1dfx, 1a7g, 1wqb, 3ssi, 1cz3
Class 3β: 1mt5, 1aoz, 2fsf, 1run, 2tdm

folding efficiency, i.e., a local descriptor that reflects a compromise between several and somehow opposite requirements. Dimerization, in fact, provides a gain in stability in small size proteins, for which a linear dependence has been found between the free energy of folding and the chains length [4, 6, 7]: therefore it is not surprising that class 1α and class 1β display the highest Q_n values (Table 2). On the other hand, large loops (SII > 0.2; $Q/R \leq 0.3$ in Table 1) mean few contacts at interface, a condition that at the expense of stability (low $\Delta G/N$ and low Q_n, Table 2) provides the dimer with an enhanced flexibility: this is a mandatory requisite for the transmission of signals between the two subunits and thus for cooperativity. While a large part of DNA binding and regulating protein are included in class 1, it can be speculated that allosteric proteins and enzymes mostly belong to class 2 or class 3, the requirement of conformational flexibility being better guarantee by the presence of loops. On the other hand, as recently pointed out by Kim and co-workers [8], the transmission of mechanical signals between the subunits of a dimeric enzyme might be a very general thermodynamic mechanism to enable enzyme catalysis, extending the concept of oligomeric cooperation well beyond the "simple" and "trivial" allosteric regulation. The analysis of protein-protein interfaces combined with experimental data on protein conformational compressibility and flexibility will help to confirm such hypothesis.

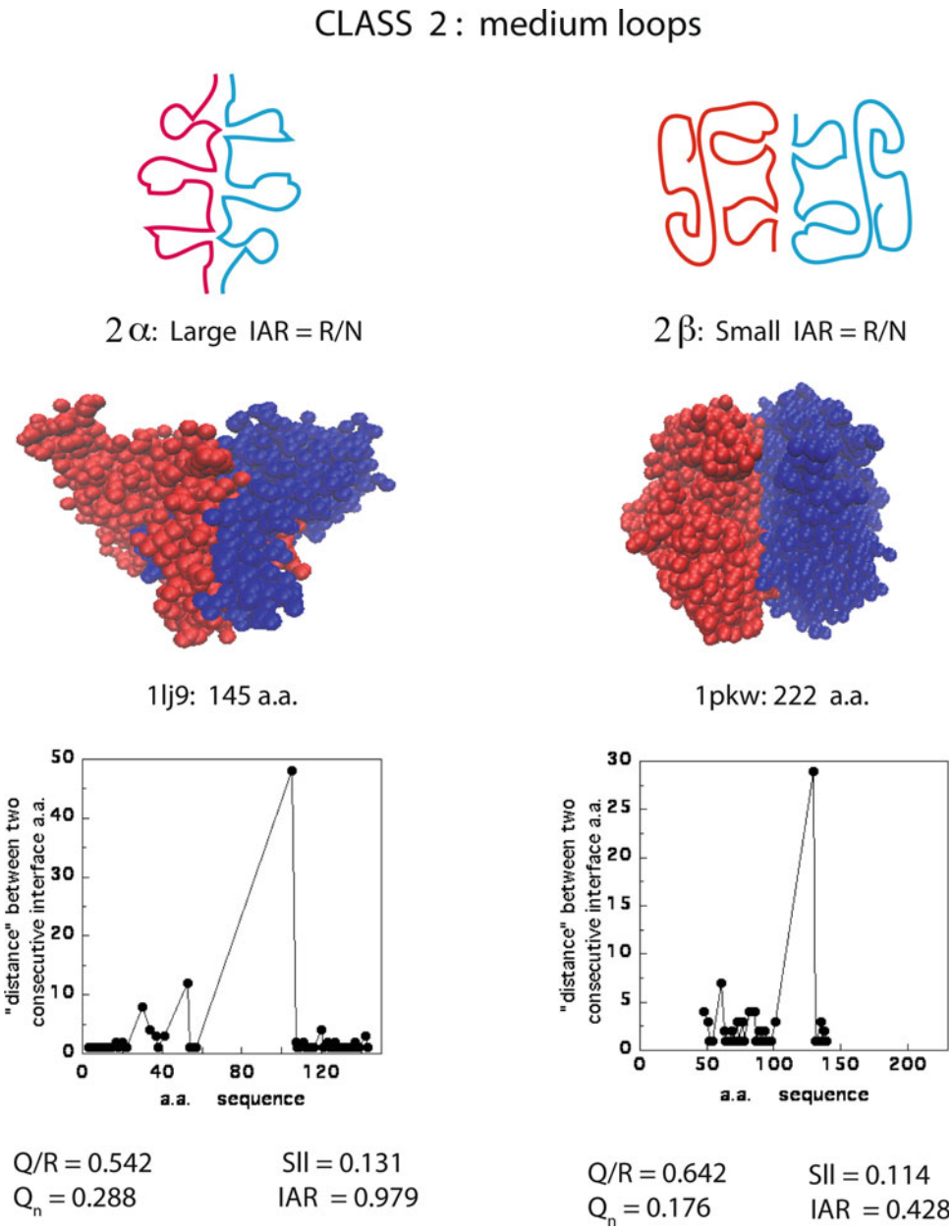

Fig. 3 Characteristics and examples of class 2 dimers

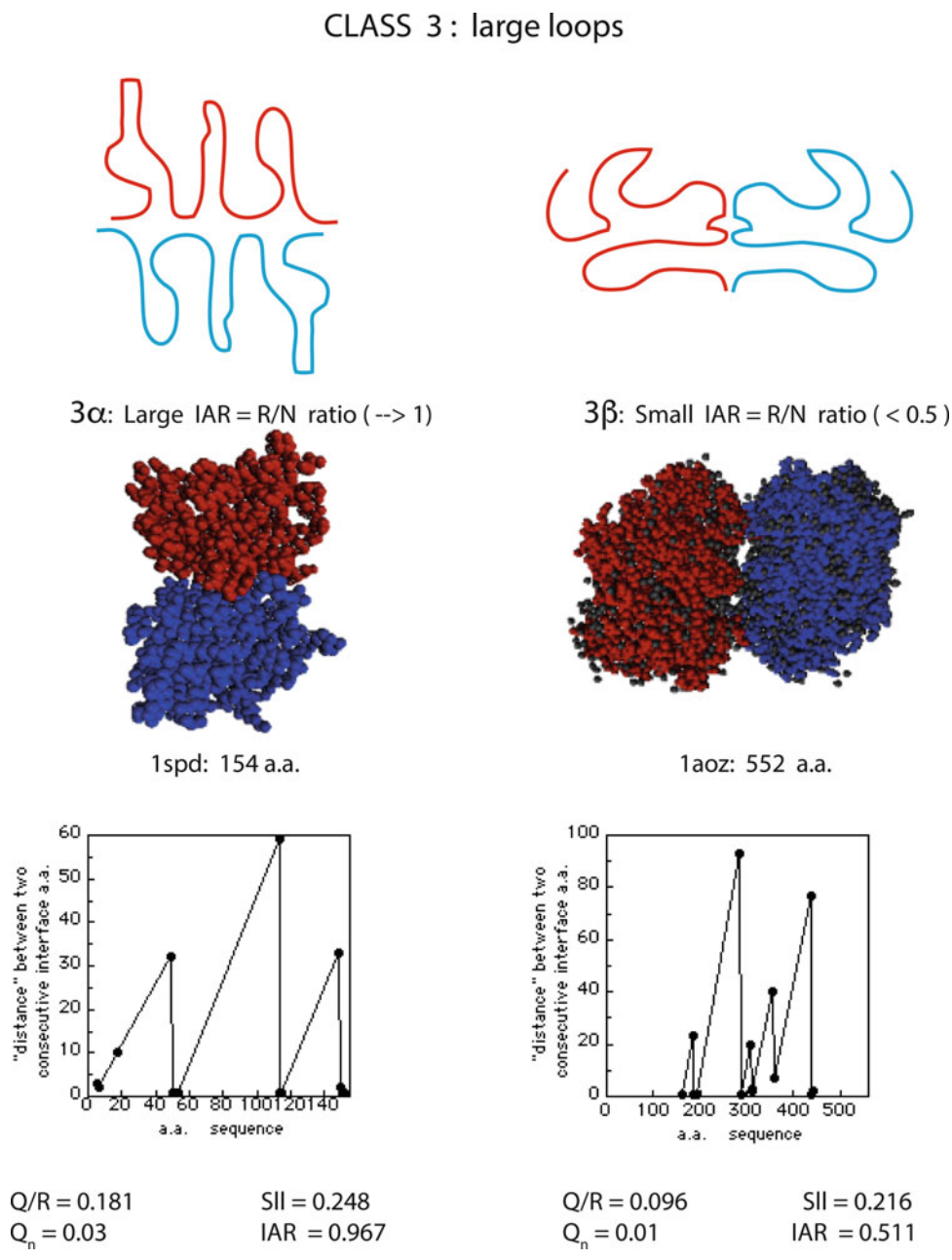

Fig. 4 Characteristics and examples of class 3 dimers

3 Notes

1. Despite being arbitrary, the choice of (at most) five consecutive a.a. to identify small loops is due to the large number of small homodimers that have been found to share such a structural feature (*see* comments on Fig. 2a).
2. Originally [5] Q_n was based exclusively on those (n) a.a. lying at the interface ($Q_n = \frac{n}{R}\frac{n}{N}$), according to the crystallographic data. However, including also residues of small loops (i.e., not necessarily at the interface, but close enough) resulted in a more incisive classification of homodimers.
3. The 50 dimers selected for this study have been preferentially chosen among those whose stabilization energy of folding is reported in literature (from equilibrium and/or kinetic unfolding measurements, *see* Table 1) [4, 7, 9].
4. A classification of the homodimers such as that reported in Table 1 is essentially based on "average" values. The heterogeneity of tri-dimensional structures that protein may assume makes this task hard sometimes because the two parameters chosen to identify the main class of assignation may not go in the same direction. An a posteriori check can be made using a *t*-test (easily performed on line or with most of the graphic packages commercially available) that provides more rigorous criteria of assignment. Such procedure has been used, for instance, in the construction of Table 1.

References

1. Luque I, Leavitt SA, Freire E (2002) The linkage between protein folding and functional cooperativity: two sides of the same coin? Annu Rev Biophys Biomol Struct 31:325–356
2. Cornish-Bowden A (2014) Understanding allosteric and cooperative interactions in enzymes. FEBS J 281:621–632
3. Koshland DE Jr (1959) Enzyme flexibility and enzyme action. J Cell Comp Physiol 54:245–258
4. Mei G, Di Venere A, Rosato N, Finazzi-Agrò A (2005) The importance of being dimeric. FEBS J 272:16–27
5. Di Paola L, Mei G, Di Venere A, Giuliani A (2016) Exploring the stability of dimers through protein structure topology. Curr Protein Pept Sci 17(1):30–36
6. Neet KE, Tim DE (1994) Conformational stability of dimeric proteins: quantitative studies by equilibrium denaturation. Protein Sci 3:2167–2174
7. Rumfeldt JAO, Galvagnion C, Vassall KA, Meiering EM (2008) Conformational stability and folding mechanisms of dimeric proteins. Progress Biophys Mol Biol 98:61–84
8. Kim TH, Mehrabi P, Ren Z, Sljoka A, Ing C, Bezginov A, Ye L, Pomès R, Prosser RS, Pai EF (2017) The role of dimer asymmetry and protomer dynamics in enzyme catalysis. Science 355: eaag2355
9. Marianayagam NJ, Sunde M, Matthews JM (2004) The power of two: protein dimerization in biology. Trends Biochim Sci 29:618–625

Chapter 7

Network Re-Wiring During Allostery and Protein-Protein Interactions: A Graph Spectral Approach

Vasundhara Gadiyaram, Anasuya Dighe, Sambit Ghosh, and Saraswathi Vishveshwara

Abstract

The process of allostery is often guided by subtle changes in the non-covalent interactions between residues of a protein. These changes may be brought about by minor perturbations by natural processes like binding of a ligand or protein-protein interaction. The challenge lies in capturing minute changes at the residue interaction level and following their propagation at local as well as global distances. While macromolecular effects of the phenomenon of allostery are inferred from experiments, a computational microscope can elucidate atomistic-level details leading to such macromolecular effects. Network formalism has served as an attractive means to follow this path and has been pursued further for the past couple of decades. In this chapter some concepts and methods are summarized, and recent advances are discussed. Specifically, the changes in strength of interactions (edge weight) and their repercussion on the overall protein organization (residue clustering) are highlighted. In this review, we adopt a graph spectral method to probe these subtle changes in a quantitative manner. Further, the power of this method is demonstrated for capturing re-ordering of side-chain interactions in response to ligand binding, which culminates into formation of a protein-protein complex in β_2-adrenergic receptors.

Key words Graph theory, Allostery, Protein structure networks, Protein-protein interactions, Side-chain interactions, Weighted networks, Spectral decomposition, Laplacian matrix

1 Introduction

Biology is anchored on the flow of information between macromolecules. These cascades of information exchange are orchestrated at multiple levels and collectively build up to a biological entity. At the heart of this lie protein-protein interactions (PPIs). The overarching factors which characterize PPIs and their interaction interface can be broadly classified into size and shape, surface

Vasundhara Gadiyaram and Anasuya Dighe contributed equally to the work.

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-1-0716-1154-8_7) contains supplementary material, which is available to authorized users.

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_7,

complementarity, interface residue propensity, hydrophobicity of interface residues, and conformational changes on binding [1].

Networks have been used successfully in understanding the complex nature of interactions in various disciplines such as biology, engineering, earth sciences, economics, and social sciences. In this chapter, the primary focus lies on networks of protein structures, and various aspects of network theory related to protein-ligand, protein-protein interactions, and its relevance to allostery are discussed.

To study allostery, the formulation of networks is used extensively. One approach is to build a network of PPIs as done by Szklarczyk et al. [2]. In such a study, the interest is in finding out the binding partner(s) of a protein or family of proteins. The role of evolutionarily conserved residues in PPI and allostery was shown by Süel and co-workers [3]. Tsai et al. [4] explored how hub proteins in a PPI network can bind to multiple partners. A broader approach on the causes behind protein interaction and allostery in co-localization was taken by Kuriyan and Eisenberg [5]. The importance of conformational ensembles in understanding allostery is emphasized by Motlagh et al. [6].

Conformational changes or conservation of residues in allostery can be efficiently studied using *Protein Structure Networks (PSNs)*. Various versions of PSNs such as Protein Contact Networks [7] and Residue Interaction Networks (RINs) [8] are available in literature. Di Paola et al. [9] beautifully capture the utility of these protein interaction networks. While most of these methods focus on backbone topology of proteins, emphasis on side-chain interactions are also given in Protein Side-chain Networks (PScNs) or Protein Energy Networks (PENs) [10]. PSNs give a bottom-up (atom-atom contacts to conformational changes and effects on the binding partners) or top-down (effect of binding partners on atom-atom contact redistribution) approach to hone in on the multi-scale contributions in allostery. A key observation in allostery is that the conformational changes could span the entire protein structure. It was reviewed in [10] that such changes can be deconstructed into allosteric communication pathways. Using known network parameters, such communication pathways can be studied by looking at optimal paths, their sub-optimal or alternate paths and junction nodes.

This chapter deals with network methods to study allostery and PPIs. The basics of network approach and metrics relevant to the studying of PSNs are provided in Subheadings 2 and 3. The details and its capability in investigating allostery have been covered extensively in earlier reviews [10–12], and the salient features are summarized here. Recent advances made in these methods [13–15], weighted networks and their spectral properties capture the influence of small perturbations on the entire system. The technique not only allows us to quantitatively evaluate the differences but also

provides the set of interacting parts undergoing transformation, which can enhance our understanding of allostery at the fundamental level. These developments are presented in Subheading 4. A case study on G-Protein Coupled Receptors (GPCRs) is presented in Subheading 5, illustrating the application of these methods to obtain better insight in allostery in these receptors.

2 Introduction to Networks

For exploring the behavior of a collection of entities, the relationship and dependencies between them should be known. For example, the behavior of a committee of people can be studied by the knowledge of the committee members and the nature of interaction with respect to each other. These entities, along with their inter-relations, can be represented as a *network*, in which the entities are named as *nodes* and the connections between them are called *edges*. The connections between the nodes can be binary (existing or non-existing) or can vary along a scale of values. Depending on the type of connections, the edges can be constructed as *unweighted* (binary) or *weighted*. The range of the weights depends upon the system and the type of problem that is being studied. A network can be represented as a graph or an *adjacency matrix* (row and column arrangement of numbers) of size $n \times n$, where n is the number of nodes in it. Once the network related to the system under scrutiny is constructed, it can be studied through various basic metrics such as *degree, hubs, and clustering coefficient* to gain a better insight about the system. *Cliques and communities* represent higher-order connectivity in a network, and they capture the local geometries in detail, within the framework of global topology of the network. All these parameters of a network can be enlisted and studied using various software like GRAPROSTR [16] and CFinder [17] by giving the adjacency matrix as input.

Though unweighted (binary) networks are easier to comprehend, the connections in most of the real-world networks (like social networks, metabolic networks, and protein structure networks) are non-binary. It is not only whether a connection exists between two nodes that matters, but the strength of the connection is more important here. In this context, weighted networks play an important role in studying real-life situations. Also, apart from the above-mentioned parameters, weighted networks can be analyzed by *ego-net* (sum of weights of all edges) of a node and many other parameters specific to weighted networks. For example, the *shortest path* between specific nodes can be evaluated using methods such as Dijkstra's algorithm [18], and has been adopted to PSNs [19] in elucidating the paths of communication due to allostery. Despite advancement in quantitatively capturing specific parameters, the matrix of a network contains a wealth of information which is

unexplored. One such intricacy is node clustering. The overall connectivity in the network leads to *node clusters*, such that connectivity between nodes of same cluster is higher compared to that across clusters. While the difference in edge weights between same nodes due to minor perturbation in a network speaks about the local variation in connections between them, they may sometimes lead to a change in global clustering among the nodes in the network. The change in node clustering can be identified by comparing Fiedler vectors of both the systems where *Fiedler vector* is defined as the eigen vector corresponding to the second smallest eigen value. Recent developments in network theory include *Network Similarity Score (NSS)* which considers all eigen vectors and eigen values of two networks and compares them at various levels like local edge weight, local clustering change, and global clustering changes [13, 14]. The unique advantage of the method lies in accurate scoring in the case of comparison between large number of extremely similar networks and also in identifying the regions of differences between the networks.

The insight regarding the changes that occur due to allostery and protein-protein interactions is brought by studying the clustering of nodes and changes in it, which forms the primary focus of this chapter.

3 Construction and Analysis of PSNs

Network representation of protein structures has proved to be successful in addressing various problems related to protein folding, dynamical behavior, ligand binding, and interactions between proteins. Network parameters such as hubs, clusters, cliques, communities, and shortest paths are the most common ones that are evaluated for characterization of critical residues, folded and unfolded states, and long-range communication during protein-protein interactions. A detailed overview of analysis of various metrics from network methodology has been discussed in reviews [10, 11] and references therein. A succinct representation is given here, which is a prerequisite to follow the recent developments presented in the following section.

The three-dimensional structure of a polypeptide chain is dictated by an optimal non-covalent interaction between different amino acids in the chain and various definitions are used to represent the non-covalent interactions in a PSN. For example, backbone networks are considered to study gross characteristics such as domain identification and protein folding [20, 21]. Similarly, side-chain networks are considered to study clusters of residues such as those involved in the interactions of the protein with other proteins

or ligands [22]. Different ways of constructing PSNs are described in detail in [10–12]. Here, a brief account of the construction of backbone and side-chain networks is provided.

For a backbone network, C^{α} atom of each amino acid is considered as a node and edges are defined based on a cut-off criterion between C^{α} atom distances. Two nodes are connected by an edge, if their C^{α} atoms are within the distance less than the cut-off. Increase in this cut-off value leads to more number of edges for a given node in the network. Various cut-off values are in usage, but the most common one is 6.5 Å. The radial distribution of non-covalently interacting C^{α} atom distances is maximum around this cut-off value, representing the first shell of interaction [23, 24].

The simplest way to create an unweighted side-chain network is to draw an edge between any two residues i and j in which at least one pair of atoms is within 4.5 Å. Several studies have adopted this procedure. Further, binary networks can also be created, depending on a cut-off value of the interaction strength (I_{min}) between the residues. If the interaction strength between two nodes is equal to or greater than the cut-off value, an edge is placed between the nodes. The interaction strength between the residues used to construct Protein Side-chain Networks (PScN) has been defined based on the equation below:

$$I_{(i,j)} = \left(n_{ij}/sqrt\left(N_i \times N_j\right)\right) \times 100$$

where $I_{(i,j)}$ is the strength of interaction between residues i and j, n_{ij} is the number of atom pairs between residues i and j within a distance cut-off of 4.5 Å, N_i and N_j are normalization values for residues i and j based on the maximum atom contacts the residue can make that are obtained from a statistically significant dataset of proteins [25]. Lower cut-off values in interaction strength (I_{min}) yield networks with higher connectivity and vice versa. The degree of a node in PSN depends on the number of nodes it can interact with and is limited due to steric constraints. Hubs formed in various types of PSNs are identified to correlate with key residues for the structural stability, function, and allosteric communication in proteins. For example, mutations that affect the ligand efficacy, but not the binding affinity, are hub residues or located near hub residues in GPCR allosteric communication pathway networks [26]. The metrics, namely cliques and communities, capture the local geometries in detail within the framework of global topology. They are used to identify rigid regions in the structure and the conformational changes due to ligand binding as shown in the example with Methionyl tRNA synthetase [27, 28]. Network analysis provides an excellent way to identify conformational changes as well as communication paths.

4 Recent Developments in PSNs

Early representations of protein structures as networks are usually binary. (Edges in binary networks are either 0 or 1, treating all edges as similar.) Though these representations capture a wealth of information, more realistic interpretations can be made by weighing the interactions between residues in terms of their strength or energy. This is realized by implementing weighted network approach for studying protein structures. The construction of weighted PSNs and its advantages are presented in detail in Subheading 4.1.

Subtle variations in pair-wise side-chain interactions will not only lead to local changes but can also permeate to global level. In fact, biological functions such as allosteric communication are known to take place at distances away from perturbation sites. Graph theoretical treatment of networks through eigen spectra is ideally suited to capture such global changes. One of the parameters which can get affected by subtle changes in interactions is the grouping of residues also known as clustering of nodes. Recent developments involve comparison of weighted PSNs and capturing changes in residue clustering using graph spectral methods. A brief overview of the graph spectral methods to study residue clustering in proteins is given in Subheading 4.2.

4.1 Weighted PSNs

The weighted PSN is generated by transforming the uniquely folded geometry of the proteins at the side-chain level to a two-dimensional weighted matrix. The edges between two residues in a protein can be weighed in various ways, say, interaction energy obtained by atomistic simulations, surface complementarity, or knowledge-based potentials, to name a few. A consolidated list of the variety of definitions used to create weighted PSNs is outlined in [10]. The simplest way of constructing weighted PSNs is based on interaction energy using geometric coordinates as described in Subheading 3. In the case study presented in Subheading 5 of this chapter, a variation of this method in calculating interaction energy has been used, which is shown below:

$$I_{ij=}n_{ij}/N_{ij}$$

Here N_{ij} is the maximum possible number of contacts that a pair of residues can make (obtained by studying a database of high-resolution protein structures). Such a kind of *weighted* representation elegantly captures the side-chain orientations with respect to each other. For example, a higher edge weight is obtained in case of stacking of aromatic residues. Similarly, in the case of hydrogen bonding between the residues, they are automatically drifted closer to each other leading to more number of atom contacts and hence higher edge weight. Additionally, the normalization of the number

of contacts with respect to the maximum possible contacts between two residues handles the large variation occurring in size and shape of the residues and weighs the interaction between them accordingly. The edge weights in this representation range between 0 and 1. Therefore a change of 0.5 and above is considered as a significant change (more than 50%) in the edge weights. A difference in edge weight between two residues can occur in the case of residues moving apart from each other at the backbone level or a change in their mutual orientations even if there is no change in the distance between their C^{α} atoms. Biological functions such as ligand binding or allosteric communication induce subtle conformational changes in the residues in both local and distal places in the structure of the protein and the change in pair-wise conformations can be studied precisely by considering weighted PSNs.

4.2 Residue Grouping and Spectral Analysis

PSNs are complex in nature and their overall topology contains geometric entities such as hubs, cliques, and communities that are composed of interacting residues and that can be identified from the adjacency matrix as described earlier [10]. What is not easy to comprehend, define, and identify is the subsequent level of organization, i.e., residue grouping. The residues are clustered together in such a way that the residues in one group interact more within themselves than with the remaining residues. In many cases, the function of a protein is attributed to a specific grouping of residues. The changes in residue grouping between two similar proteins (or two states of the same protein) can be obtained broadly by comparing the Fiedler vectors of their networks which can be obtained from graph spectral analysis. Further, they can be visualized by coloring the residues in the three-dimensional protein structure according to their Fiedler vector components.

The graph spectral way of analyzing protein structures has been used earlier to identify entities such as clusters and cluster centers [20]. It has been used to detect domains and domain interfaces in multi-domain proteins [29]. Further, the interface cluster and the hotspots which are responsible for the stability of two alpha-subunits in RNA polymerase were correctly predicted using cluster analysis of PSNs [30]. Cluster analysis is also useful in identifying motifs which are not sequential, but spatially close to quaternary associations in lectins [31]. The formation of non-homogeneous clusters of residues at interface of oligomers elucidates the modular architecture of protein-protein interfaces [32]. The wealth of information from these studies is obtained from graph spectral features of unweighted networks. However, the increase in information will be manifold by adopting a spectral analysis of weighted networks and the results would be more realistic as well as biologically relevant.

The clustering of nodes is uniquely obtained from *graph spectral decomposition* on a network which involves obtaining eigen values and eigen vectors of Laplacian matrix of the network. Nodes of the same cluster have closer numerical values in the Fiedler vector components. Therefore, sorted Fiedler vector components are used to obtain node clustering in the network. A flowchart describing this methodology is depicted in Fig. 1.

A holistic understanding of an allosteric system can be obtained by analyzing edge weights and node clustering using weighted PSNs. This is illustrated below in the case of a protein complex with a well-characterized biological activity.

5 Case Study of β_2 Adrenergic Receptors

We apply the methodology discussed above to understand the mechanism of signal transduction that involves transmission of chemical or physical signals through the cell, triggering a characteristic cellular response. Herein, we focus on the archetypal transmembrane (TM) signaling molecules i.e., G-Protein Coupled Receptors (GPCRs), which possess extraordinary potential to respond to a diverse set of extracellular stimuli like light, ions, hormones, neurotransmitters, and small molecule ligands and thereby mediate cellular signaling by interacting with heterotrimeric GTP-binding proteins (G-proteins) [33]. At the heart of this signaling cascade lies the ligand-dependent activation of GPCRs, followed by G-protein coupling and nucleotide exchange that eventually culminates in regulation of downstream effector proteins [34]. GPCRs function by means of ligand-driven activation followed by conformational changes that mediate interaction with GDP-bound G-protein heterotrimers ($G_{\alpha\beta\gamma}$). Nucleotide exchange and the subsequent dissociation of $G_{\beta\gamma}$ from G_{α} result in regulating the activities of cellular effectors like kinases, ion channels, and other enzymes. GPCRs control a variety of physiological processes that include sense of smell, taste, sight as well as immune response, behavior, autonomous nervous system transmission, and homoeostasis modulation [35, 36]. Their implication in numerous biological phenomena thus warrants an understanding of intricacies in their three-dimensional molecular structure [37].

X-ray crystallography–based experimental studies offer high-resolution atomistic details on molecular structure of GPCRs. Details from the three-dimensional structure can be used to explain effects of GPCRs in response to a diverse array of small molecule ligands. These ligands bind to the GPCRs at conserved or orthosteric binding site located at the core of their seven-helical transmembrane (7TM) region. Discrete classes of ligands include those that maximally activate the receptor (agonists), induce sub-optimal activity (partial agonists), inhibit basal activity (inverse agonists),

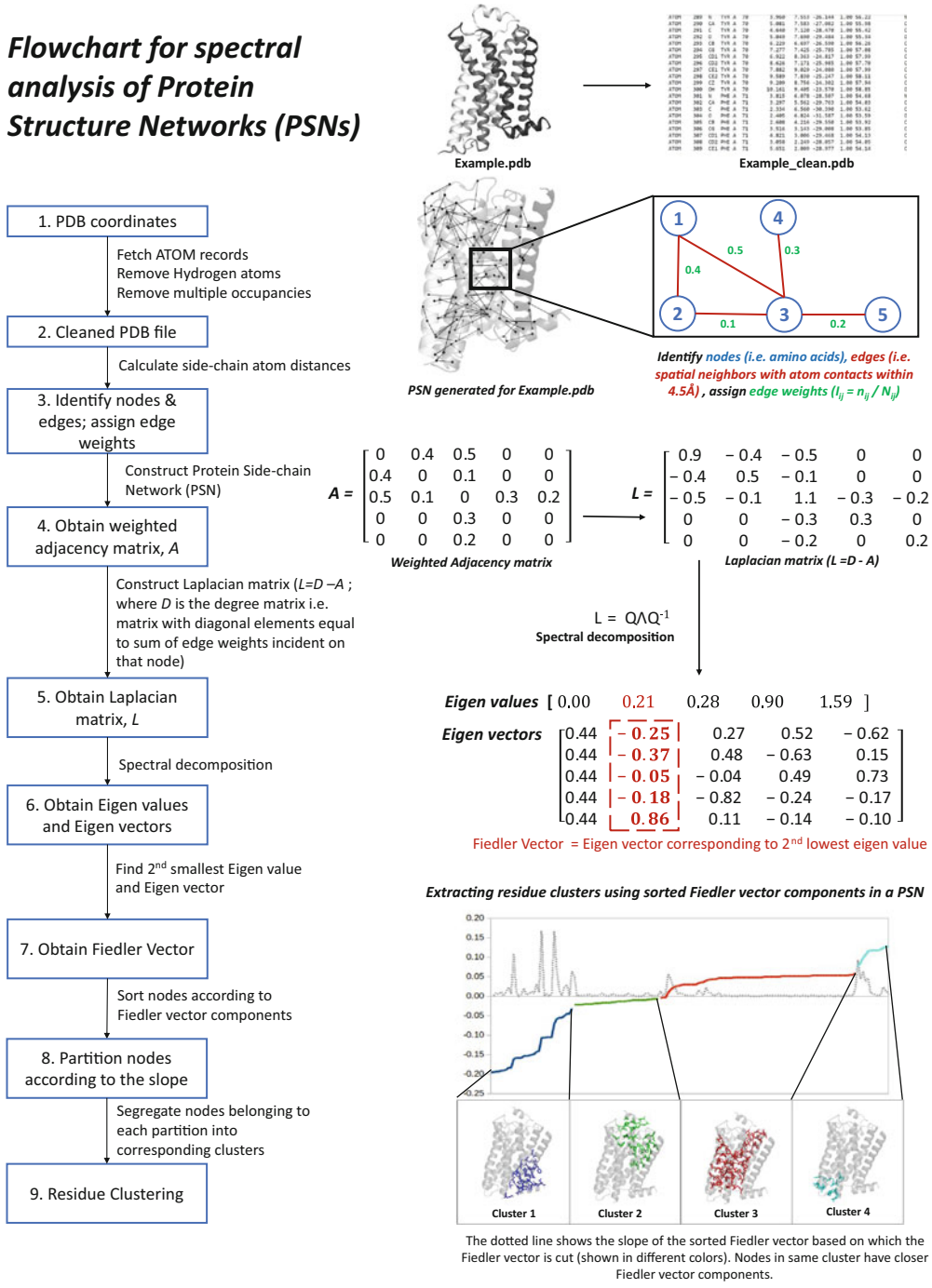

Fig. 1 Flowchart illustrating the procedure for spectral analysis of Protein Structure Networks (PSNs). A step-by-step illustration of carrying out spectral analysis of PSNs. Left panel depicts steps involved, whereas right panel contains a pictorial representation of each step

selectively activate specific signaling cascades (biased agonists), and abolish activity of other ligands (antagonists). Co-crystallization of the GPCR with such ligands reveals distinct structural changes that occur upon GPCR activation.

The structure of the agonist-occupied (ligand P0G, i.e., 8-[(1R)-2-{[1,1-dimethyl-2-(2-methylphenyl)ethyl]amino}-1-hydroxyethyl]-5-hydroxy-2H-1,4-benzoxazin-3(4H)-one), active state of β_2-Adrenergic Receptor (β_2AR) bound to a nucleotide-free G_s-protein heterotrimer provides the first high-resolution view of the ternary complex of GPCR activation (PDB ID: 3SN6, hereafter referred to as β_2AR-G_s). Here we revisit this classical complex [38] and analyze it in terms of weighted PSNs. Additionally, we also consider an antagonist alprenolol-bound inactive state β_2AR [39] (PDB ID: 3NYA, hereafter referred to as β_2AR-anta) for elucidating ligand-induced conformational changes at side-chain levels. Being one of the first GPCRs to be biophysically characterized, cloned [40] and structurally determined by means of X-ray crystallography [41], the β_2-Adrenergic Receptor serves as a classical system for understanding cell signaling. β_2AR functions by binding to hormone and neurotransmitter adrenaline (also known as epinephrine) and inducing physiological responses like smooth muscle relaxation and bronchodilation via the agency of L-type Calcium channels [38]. A schematic representation of the ternary complex of GPCR, G-protein and ligand is depicted in Fig. 2.

With a view to understand ligand-induced complex formation between GPCR and G-protein, we investigate agonist and antagonist-bound GPCR complexes in terms of weighted side-chain edges and bring to light distinct clustering patterns that delineate ligand-induced conformational changes. We adopt the weighted PSN methodology to understand the propagation of information across the 7TM architecture in terms of non-covalent side-chain interactions. In this section we investigate the following: (1) ligand-protein contacts in β_2AR-G_s and β_2AR-anta; (2) consequence of ligand binding on redistribution of interactions in GPCR-G-protein bound systems, and (3) manifestation of these perturbations on the local and global re-wiring of GPCRs in terms of residue (node) clustering.

5.1 Binding Site Comparison of β_2AR-G_s and β_2AR-Anta Complexes

Comparison of the ligand binding pocket in β_2AR-G_s and β_2AR-anta complexes reveals crucial information about the role of TM helices in maintaining ligand-protein contacts. In particular, the high-affinity agonist P0G (or BI-167107) in β_2AR-G_s is housed in the binding pocket by means of strong hydrogen bonds mediated by residues Asp113$^{3.32}$, Asn312$^{7.39}$, Ser203$^{5.42}$, and Ser207$^{5.46}$ (Fig. 3 panel a, superscripts indicate Ballesteros-Weinstein numbering Scheme [42] in which the first digit in superscript indicates helix number and digits in superscript following decimal point indicate residue number relative to the most conserved residue in that particular TM helix which is numbered as

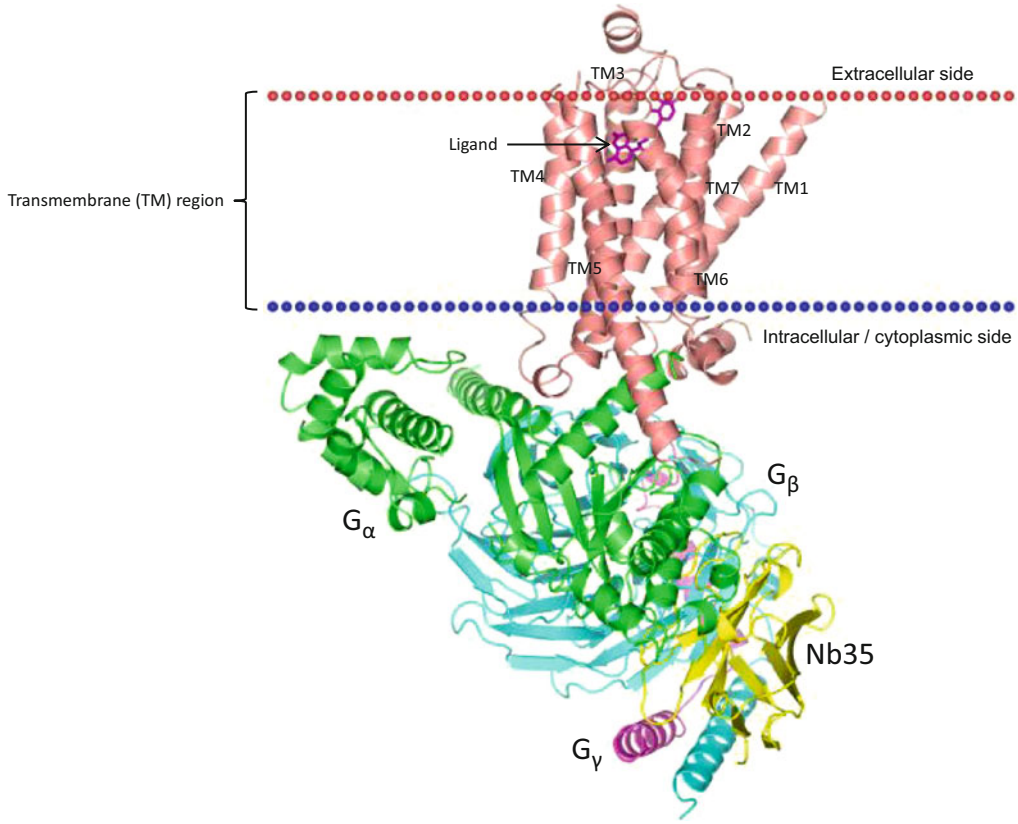

Fig. 2 A schematic representation of a typical GPCR. A cartoon representation of a typical G-protein Coupled Receptor (GPCR, orange) showing its association with a ligand (pink, sticks) and G_s-protein (G_α: green, G_β: cyan, G_γ: magenta), along with a stabilizing nanobody (yellow). Membrane boundaries depicted as red dots (extracellular side) and blue dots (intracellular/cytoplasmic side)

50 in the given helix). Residues belonging to TM6 and ECL2 also participate in maintaining this complex network of interactions that nestles the agonist in the binding pocket in an optimal fashion.

The ligand binding pocket in the β_2AR-anta complex (Fig. 3, panel b) shows a shift in these ligand-protein contact preferences. Contacts with Ser203$^{5.42}$ and Ser207$^{5.46}$ are lost, though Asp113$^{3.32}$ and Asn312$^{7.39}$ continue to interact with the antagonist through four hydrogen bonds formed with β-OH and NAP atoms of antagonist JTZ. Formation of short hydrogen bonds with Tyr316$^{7.43}$ recruits TM7 into the role of locking the antagonist into the binding pocket.

The loss of hydrogen bonds with TM5 (mediated by Ser203$^{5.42}$ and Ser207$^{5.46}$) results in a cascade of subtle ligand-induced conformational changes which render TM helices 5 and 6 immobile in the β_2AR-anta structure. In GPCRs, the motion of these helices is crucial for the formation of a cleft on the intracellular side of the receptor that gets occupied by the α5 helix of the

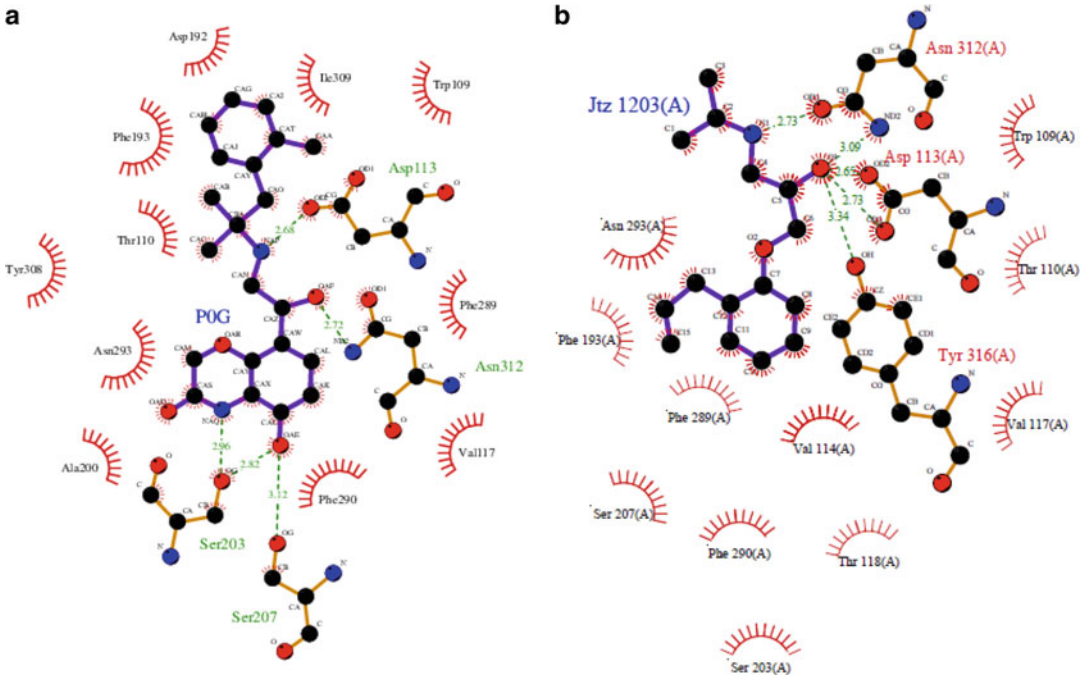

Fig. 3 Comparison of ligand binding sites of β_2AR-G_s and β_2AR-anta. A schematic diagram depicting protein-ligand interactions at the ligand binding site of (**a**) β_2AR-G_s, and (**b**) β_2AR-anta. Green dotted lines represent hydrogen bonds. Binding site representations are obtained through LigPlot [43]

G-protein. Hence, it is interesting to note how changes in ligand binding pocket translate into structural alterations in TM5, TM6 and TM7 which facilitate the binding of G_s. Agonist interactions with Ser203$^{5.42}$ and Ser207$^{5.46}$ give rise to a 2 Å inward motion of TM5 at Ser207$^{5.46}$ and 1.4 Å inward movement at Pro211$^{5.50}$ [38]. This pull disturbs a highly conserved network of contacts between Pro211$^{5.50}$, Ile121$^{3.40}$, Phe282$^{6.44}$ and Asn318$^{7.45}$. As a result, Ile121$^{3.40}$ and Phe282$^{6.44}$ are repositioned, along with a rotation about Phe282$^{6.44}$, leading to an outward displacement of TM6 at the cytoplasmic end. In the β_2AR-anta complex, the antagonist JTZ (Alprenolol) is unable to form hydrogen bond with Ser203$^{5.42}$ and Ser207$^{5.46}$ because of the absence of an electronegative atom in JTZ in the vicinity. Additionally, Pro211$^{5.50}$ stacks with Trp122$^{3.41}$. This freezes Pro211$^{5.50}$ in its position and hence it is incapable to disturb the network of contacts between TMs 3, 5 and 6. Consequently, Phe282$^{6.44}$ is unable to undergo local twisting to undergo displacement. Hence, the mobility of TM6 is compromised owing to re-wiring of contacts in the ligand binding pocket.

Another scenario brought to light is the disengagement of individual helices because of agonist-induced conformational

changes. In other words, the presence of an agonist in the ligand binding pocket enables individual helices to extricate themselves and gain flexibility, thus ultimately facilitating binding of the G_s-protein. On the other hand, presence of an antagonist compels the helices to associate with each other (at the cytoplasmic region) and thus makes binding of G_s-protein impossible. Overall, we observe that local changes in the ligand binding site translate to a re-wiring of contacts at the cytoplasmic region, thus exhibiting a characteristic allosteric response. Further, the strength of these interactions examined in terms of side-chain edge weights also highlights the mechanism of information flow across the TM architecture, which is discussed below.

5.2 Differences in Edge Weights Between the Complexes β_2AR-G_s and β_2AR-Anta

Analyzing non-covalent side-chain interactions in terms of edge weights highlights subtle differences in the overall architecture of the transmembrane region. Herein, we compared the agonist and antagonist-bound β_2AR complexes and obtained a list of interactions which differ among the compared systems by an edge weight of 0.5 (Table 1).

5.2.1 Differences Between β_2AR-G_s and β_2AR-Anta Complexes

Disengagement of individual helices in the agonist bound case has a dramatic effect on β_2AR to interact with G_s. A closer look at edge weight differences between β_2AR-G_s and β_2AR-anta reveals an interesting trend among residues that are present in the cytoplasmic side of TMs 5 and 6 as well as the ICL2 (Fig. 4). Ideally, when the receptor coordinates with its cognate G-protein, these residues are recruited for establishing contacts with the α5 and αN helical regions of G_s, leading to stabilization of the complex. Here, we highlight certain examples from our findings that illustrate a change in the behavior of these bridging residues. First, Ser143 (ICL2) forms strong residue contacts with Asp130$^{3.49}$ and Val67$^{2.38}$ in β_2AR-anta (edge weights 0.75 and 1.00 respectively). In the G-protein bound complex the same residue, i.e., Ser143 (ICL2), coordinates via its OG atom and forms intermolecular interaction with the Ala139 (CB) in the αN helix. Second, Gln229$^{5.68}$ interacts with a high edge weight with Lys140 (ICL2) in the β_2AR-anta (edge weight 0.75). In the G-protein bound complex β_2AR-G_s, it establishes contacts with a network of residues (Asp381, Gln384, Arg385 and Leu388) present in the α5 helical region. Third, in β_2AR-anta, His269$^{6.31}$ forms an interaction with Ala226$^{5.65}$ (edge weight 0.67). But Ala226$^{5.65}$ from the receptor interacts with Leu388 and Leu393 from the α5 region of G-protein in the G-protein bound complex. From these cases, we postulate that residues from the interface region are coerced into forming intramolecular interactions rather than getting involved in inter-

Table 1
Pairwise comparison of differences in side-chain edge weights

Edge weight differences between β2AR-Gs and β2AR-anta							
Serial no.	**Residue name**	**Residue no.**	**Helix**	**Residue name**	**Residue no.**	**Helix**	**Edge weight difference**
Inter-helical interactions:							
1	**HIS**	**269**	**TM6**	**ALA**	**226**	**TM5**	**−0.67**
2	**PRO**	**211**	**TM5**	**TRP**	**122**	**TM3**	**−0.62**
3	**CYS**	**116**	**TM3**	**ALA**	**78**	**TM2**	**−0.5**
4	**SER**	**120**	**TM3**	**ALA**	**78**	**TM2**	**−0.5**
5	**MET**	**279**	**TM6**	**ALA**	**128**	**TM3**	**−0.5**
6	CYS	327	TM7	VAL	54	TM1	0.5
7	SER	161	TM4	ALA	119	TM3	0.5
8	ALA	335	TM8	PHE	61	TM1	0.71
9	PHE	332	TM8	CYS	327	TM7	0.77
10	ASN	322	TM7	LEU	75	TM2	1
Intra-helical Interactions:							
11	**VAL**	**86**	**TM2**	**MET**	**82**	**TM2**	**−0.5**
12	**SER**	**165**	**TM4**	**GLY**	**162**	**TM4**	**−0.5**
13	**SER**	**207**	**TM5**	**SER**	**203**	**TM5**	**−0.5**
14	ASP	79	TM2	LEU	75	TM2	0.5
Helix-loop:							
15	**VAL**	**67**	**TM2**	**SER**	**143**	**ICL2**	**−1**
16	**GLN**	**229**	**TM5**	**LYS**	**140**	**ICL2**	**−0.75**
17	**LYS**	**273**	**TM6**	**PHE**	**264**	**IL3**	**−0.64**
18	ASN	103	TM3	GLU	188	ECL2	0.5
19	ALA	134	TM3	TYR	141	ICL2	0.5
20	PHE	332	TM8	LEU	64	ICL1	0.53
21	GLU	268	TM6	CYS	265	ICL3	0.6
Loop-Loop:							
22	**CYS**	**191**	**ECL2**	**TRP**	**99**	**ECL1**	**−0.88**
23	**SER**	**143**	**ICL2**	**ASP**	**130**	**TM3**	**−0.75**
24	**TYR**	**185**	**ECL2**	**ARG**	**175**	**ECL2**	**−0.61**
25	**THR**	**189**	**ECL2**	**GLU**	**187**	**ECL2**	**−0.57**

(continued)

Table 1
(continued)

Edge weight differences between β2AR-Gs and β2AR-anta							
Serial no.	**Residue name**	**Residue no.**	**Helix**	**Residue name**	**Residue no.**	**Helix**	**Edge weight difference**
26	**LEU**	**144**	**ICL2**	**SER**	**137**	**ICL2**	**−0.5**
27	LYS	140	ICL2	SER	137	ICL2	0.67

A tabular representation of differences in side-chain edge weights between β_2AR-G_s and β_2AR-anta. Interactions in bold font indicate those that are stronger in β_2AR-anta and regular font indicate interactions stronger in β_2AR-G_s. Interactions are grouped according to secondary structures involved for ease of understanding

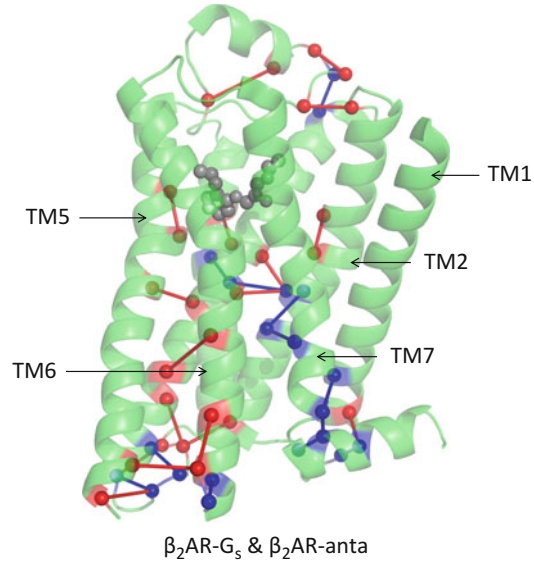

Fig. 4 Comparison of edge weight differences across β_2AR-G_s and β_2AR-anta. Comparison of side-chain edge weight differences across systems, i.e., β_2AR-G_s and β_2AR-anta. GPCR is depicted as green cartoons while the ligand is shown as grey spheres. Red and blue spheres represent C^{α} atoms of amino acid residues. Red edges denote edges that are stronger in β_2AR-anta, whereas blue edges signify those that are stronger in β_2AR-G_s

molecular interactions with its cognate G-protein in the antagonist bound β_2AR (Fig. 4). The same behavior can be observed among other residues present in the ICL2 region of β_2AR. Specifically, residues like Leu144, Ser137, and Tyr141 form interactions within

the receptor rather than entering into stabilizing interactions with the G-protein.

Another striking feature observed is weakening of interaction between Ser203$^{5.42}$ and Ser207$^{5.46}$. This interaction suffers a loss in its edge weight (0.75 in β_2AR-anta and 0.25 in β_2AR-G_s). Due to this weakening, the Pro211$^{5.50}$ is unable to get inward into the TM core and as a resulting cascade of interactions; it leads to the immobility of TM6. Also interesting is that this inability of Pro211$^{5.50}$ to move inward may be owing to strong stacking interactions with Trp122$^{3.41}$ signified by a high edge weight (0.69 in β_2AR-anta and 0.07 in β_2AR-G_s).

5.3 Spectral Decomposition-based Side-Chain Clustering Patterns

The Fiedler vectors of the side-chain networks reveal the difference in clustering profiles of the systems analyzed (Fig. 5). These are obtained by spectral decomposition of the Laplacian of weighted side-chain networks. The residues depicted in same color are present in the same clusters. Figure 5, panel a shows β_2AR-G_s with the spread of its clusters across the protein structure such that the core of the TM helices forms one dominant cluster, as depicted in red color. Additionally, residues in the intracellular regions of TM 5 and 6 segregate into a smaller, yet significantly sized, cluster (dark blue color). Interestingly these residues are also implicated in interacting

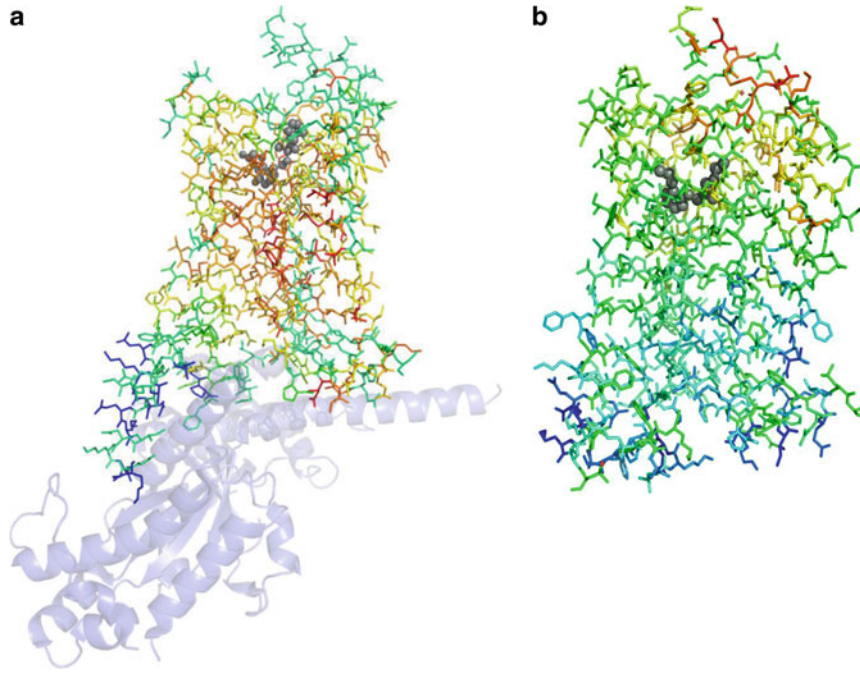

Fig. 5 Spectral decomposition-based residue clustering Node clustering based on Fiedler vector of side-chain networks in (**a**) β_2AR-G_s, and (**b**) β_2AR-anta. The side-chains are shown in stick representation, and ligand in grey spheres. The binding partners, i.e., G_s (panel **a**), are shown as purple cartoons. For better clarity, residues 88–202 from G_α subunit are omitted in panel **a**

with the G-protein. Clusters are seen to be spread over extracellular regions encompassing residues of ECL2 and ECL1 regions (cyan color). These clusters can also be seen to propagate to the intracellular region, hence establishing a possible link between these areas. In the case of β_2-anta (Fig. 5, panel b), the limited mobility of the helices because of antagonist binding leads to a more compact and discrete intracellular region. Hence the lateral clustering becomes more dominant and discrete clusters can be observed in intracellular region.

To explore the spread of clusters in a more elegant way, the sorted Fiedler vectors of the agonist and antagonist cases are compared in Figs. 6 and 7. A tabular representation of a representative cluster and its constituent residues is given in Tables 2 and 3. Supplementary Table 1 contains a comprehensive list of residues belonging to all clusters in β_2AR-G_s and β_2AR-anta. In Figs. 6 and 7, the top panel depicts the sorted Fiedler vectors of β_2AR-G_s (3SN6) and β_2AR-anta (3NYA). It is interesting to note that the clustering is clearer and the clusters are more distinguishable in β_2AR-anta (3NYA) compared to that of β_2AR-G_s (3SN6). The clusters are shown in different colors, and the coloring scheme is

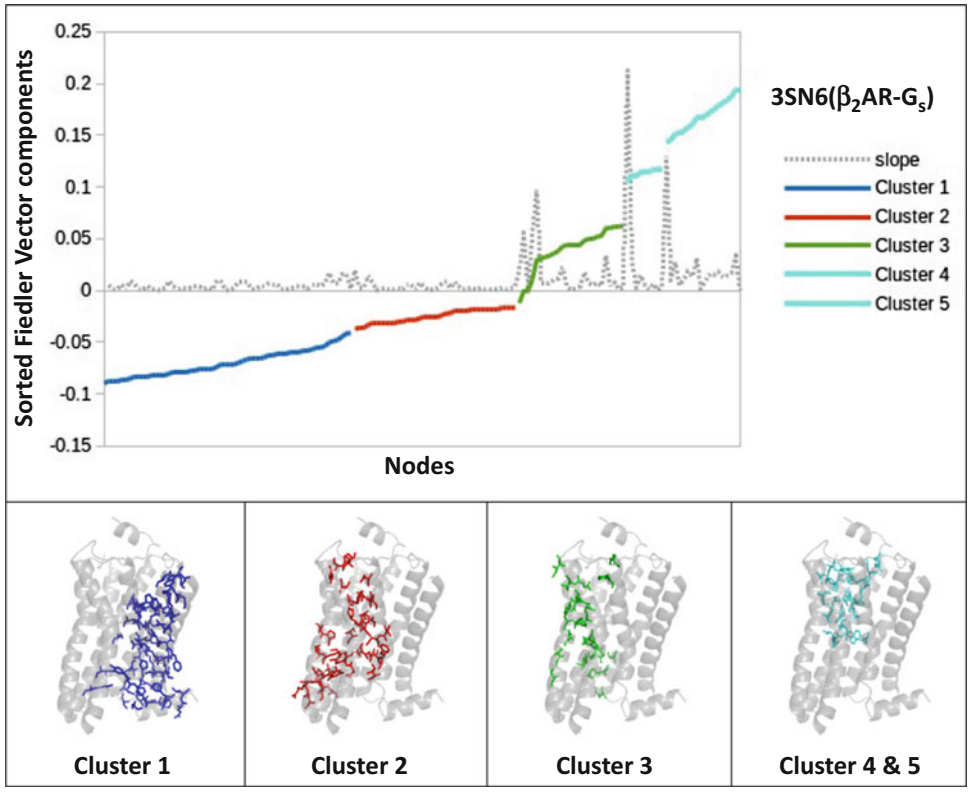

Fig. 6 Sorted Fiedler vectors and clustering representations of β_2AR-G_s (3SN6). Top: Profile of sorted Fiedler vectors in β_2AR-G_s (3SN6); Bottom: Location of individual clusters mapped onto protein structure

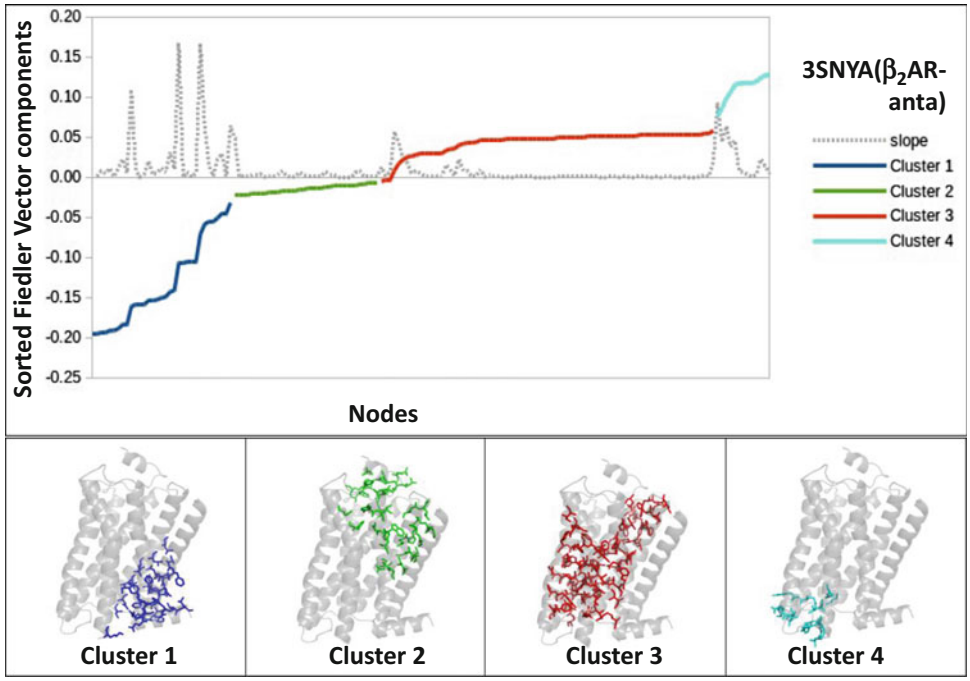

Fig. 7 Sorted Fiedler vectors and clustering representations of β_2AR-anta (3NYA). Top: Profile of sorted Fiedler vectors in β_2AR-anta (3NYA); Bottom: Location of individual clusters mapped onto protein structure

used to represent the position of residues in three-dimensional structures in the bottom panel. The global effect of the disengagement of the TM helices in the agonist bound case, i.e., β_2AR-G_s (3SN6), and the proximity of helices in the antagonist bound case, i.e., β_2AR-anta (3NYA), at the side-chain interaction level are strikingly obvious from the clustering patterns adopted by these two systems. In the agonist bound case, i.e., β_2AR-G_s (Fig. 6, 3SN6), the residue clustering has happened perpendicular to the membrane plane, i.e., in a vertical fashion, reflecting the disengagement of helices in the agonist bound case, thus facilitating the binding of G_s protein. On the other hand, in antagonist-bound case β_2AR-anta (Fig. 7, 3NYA), the clusters traverse parallel to the membrane plane, i.e., in a horizontal fashion. This way of clustering has even more dramatic impact in terms of keeping the receptor amenable for communication across the membrane boundaries.

In summary, we have reviewed the methods available to extract network parameters from side-chain interactions for the study of protein structures. Some of these have already been reviewed in literature and are summarized here. The recent development of extracting information from weighted network is elaborated in this chapter. Subtle changes in the conformations during allostery and protein-protein interactions are captured by difference in the edge weight and the manifestation of its effect at global levels. We

Table 2
Residue clustering in β_2AR-G_s. A tabular representation containing list of residues which belong to cluster2 in β_2AR-G_s shown as red sticks. Individual columns represent node, residue number, TM helix number, sorted Fiedler vector component and slope of the sorted Fiedler vector respectively

Residues constituting cluster2 in 3SN6 (β2AR-Gs)				
Node	ResNum	TM	LapFV	Slope
224	296	TM6	-0.03741	0.01968
223	295	TM6	-0.03691	0.00248
231	303	TM7	-0.03643	0.00242
220	292	TM6	-0.03352	0.01452
235	307	TM7	-0.03203	0.00747
240	312	TM7	-0.03201	0.00007
216	288	TM6	-0.03178	0.00119
245	317	TM7	-0.03132	0.00228
239	311	TM7	-0.03128	0.00019
253	325	TM7	-0.03108	0.00103
249	321	TM7	-0.03063	0.00225
209	281	TM6	-0.03058	0.00024
250	322	TM7	-0.02938	0.00600
213	285	TM6	-0.02836	0.00512
246	318	TM7	-0.02822	0.00069
243	315	TM7	-0.02672	0.00748
93	124	TM3	-0.02605	0.00335
86	117	TM3	-0.02604	0.00007
217	289	TM6	-0.02598	0.00027
214	286	TM6	-0.02551	0.00239
44	75	TM2	-0.02425	0.00630
114	145	TM4	-0.02249	0.00880
102	133	TM3	-0.02207	0.00207
98	129	TM3	-0.02043	0.00822
187	221	TM5	-0.01981	0.00311
105	136	TM3	-0.01973	0.00039
101	132	TM3	-0.01947	0.00128
176	210	TM5	-0.01896	0.00257
184	218	TM5	-0.01885	0.00052
180	214	TM5	-0.01846	0.00196
177	211	TM5	-0.01834	0.00058
90	121	TM3	-0.01825	0.00049
94	125	TM3	-0.01816	0.00045
206	278	TM6	-0.01791	0.00121
185	219	TM5	-0.01755	0.00183
97	128	TM3	-0.01749	0.00028
207	279	TM6	-0.01709	0.00201
181	215	TM5	-0.01633	0.00382

Table 3
Residue clustering in β_2AR-anta. A tabular representation containing list of residues which belong to cluster3 in β_2AR-anta shown as red sticks. Individual columns represent node, residue number, TM helix number, sorted Fiedler vector component and slope of the sorted Fiedler vector respectively

Residues constituting cluster3 in 3NYA (β2AR-anta)				
Node	ResNum	TM	LapFV	Slope
82	113	TM3	**-0.00478**	0.01276
55	86	TM2	**-0.00445**	0.00167
51	82	TM2	**-0.00317**	0.00637
250	322	TM7	**0.00823**	0.05703
85	116	TM3	**0.01699**	0.04379
50	81	TM2	**0.02174**	0.02378
47	78	TM2	**0.02527**	0.01762
54	85	TM2	**0.02636**	0.00547
206	278	TM6	**0.02794**	0.00787
253	325	TM7	**0.02899**	0.00528
81	112	TM3	**0.02933**	0.00168
205	277	TM6	**0.02963**	0.00150
201	273	TM6	**0.02995**	0.00162
256	328	TM7	**0.03017**	0.00107
202	274	TM6	**0.03030**	0.00068
89	120	TM3	**0.03344**	0.01568
77	108	TM3	**0.03529**	0.00928
73	104	TM3	**0.03567**	0.00190
44	75	TM2	**0.04019**	0.02257
93	124	TM3	**0.04184**	0.00825
74	105	TM3	**0.04346**	0.00810
210	282	TM6	**0.04353**	0.00034
92	123	TM3	**0.04388**	0.00176
43	74	TM2	**0.04591**	0.01014
58	89	TM2	**0.04598**	0.00038
246	318	TM7	**0.04675**	0.00383
214	286	TM6	**0.04687**	0.00060
243	315	TM7	**0.04705**	0.00089
86	117	TM3	**0.04713**	0.00043
213	285	TM6	**0.04728**	0.00074
209	281	TM6	**0.04729**	0.00006
123	154	TM4	**0.04737**	0.00040
90	121	TM3	**0.04774**	0.00187
242	314	TM7	**0.04800**	0.00126
181	215	TM5	**0.04831**	0.00159
177	211	TM5	**0.04841**	0.00046
70	101	TM3	**0.04850**	0.00047
238	310	TM7	**0.04851**	0.00006

Table 3
(continued)

91	122	TM3	**0.04856**	0.00021
249	321	TM7	**0.04863**	0.00039
94	125	TM3	**0.04872**	0.00043
180	214	TM5	**0.04908**	0.00183
245	317	TM7	**0.04930**	0.00106
61	92	TM2	**0.04962**	0.00164
218	290	TM6	**0.04996**	0.00167
174	208	TM5	**0.05011**	0.00076
207	279	TM6	**0.05032**	0.00105
65	96	TM2	**0.05048**	0.00081
69	100	TM2	**0.05073**	0.00121
68	99	TM2	**0.05096**	0.00118
157	191	ECL2	**0.05118**	0.00108
75	106	TM3	**0.05120**	0.00010
96	127	TM3	**0.05120**	0.00001
178	212	TM5	**0.05145**	0.00127
62	93	TM2	**0.05158**	0.00063
40	71	TM2	**0.05165**	0.00037
78	109	TM3	**0.05183**	0.00088
59	90	TM2	**0.05202**	0.00094
185	219	TM5	**0.05204**	0.00012
170	204	TM5	**0.05213**	0.00046
188	222	TM5	**0.05244**	0.00155
204	276	TM6	**0.05249**	0.00022
222	294	TM6	**0.05261**	0.00062
175	209	TM5	**0.05265**	0.00021
189	223	TM5	**0.05279**	0.00068
119	150	TM4	**0.05281**	0.00011
211	283	TM6	**0.05297**	0.00082
95	126	TM3	**0.05309**	0.00057
36	67	TM2	**0.05314**	0.00025
208	280	TM6	**0.05314**	0.00003
112	143	ICL2	**0.05331**	0.00084
99	130	TM3	**0.05336**	0.00024
114	145	TM4	**0.05337**	0.00006
97	128	TM3	**0.05340**	0.00014
122	153	TM4	**0.05360**	0.00101
215	287	TM6	**0.05413**	0.00263
219	291	TM6	**0.05436**	0.00116
184	218	TM5	**0.05821**	0.01926

have utilized specialized algorithms of spectral analysis of weighted networks to obtain re-grouping of residues as clusters in response to binding of different ligands or proteins. The information at such a level obtained on protein structures has been presented in the context of allostery and protein-ligand complex formation. The protocol of interpreting the metrics like Fiedler vector has been demonstrated, and the biological relevance of the methods discussed here has been highlighted with the example of β_2 adrenergic receptor, a member of GPCR family.

6 Additional Notes

1. While comparing PSNs using edge weight difference, sufficient care is to be taken such that comparison is being made between nodes corresponding to same residues. A Needleman-Wunsch alignment [44] can be used with default settings for this node correspondence. Any missing residues are either to be compensated by adding additional dummy rows and columns in the network with missing ones or to be removed from the other network.
2. The spectral decomposition of the networks using Laplacian matrix requires that there are no isolated nodes (nodes with zero edges) in the network. Hence any such nodes are to be made non-isolated by adding an edge to those nodes with a logically relevant node. For example, for backbone network, the node with nearest C^{α} distance can be selected. For side-chain networks, the node can be connected to that of a residue with atom contacts nearest to it. The weight of the edge added should be of very low value (for example, 10^{-32}) so that it does not interfere with the spectral information.
3. Clusters in a network are identified by cutting the Fiedler vector of the Laplacian matrix of the PSN. If the clusters are well separated as in the case of antagonist in GPCR case-study (Fig. 5, panel b), they are also well separated in their Fiedler vector components, nodes in the same cluster possessing very close values. In the case of clusters well interspersed with each other like that of agonist in case study (Fig. 5, panel a) making the sorted Fiedler vector a continuous increasing one, the cut in the Fiedler vector to obtain clusters is to be done judiciously.

Acknowledgments

S.V. thanks National Academy of Sciences (NASI), Allahabad, India, for Senior Scientist Fellowship. V.G. thanks IISc for incentive research support grant and CSIR for Research Associate fellowship.

We thank SERC and MBU of the Indian Institute of Science for computational facilities.

References

1. Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci U S A 93(1):13–20
2. Szklarczyk D et al (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43(Database issue):D447–D452
3. Suel GM et al (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 10 (1):59–69
4. Tsai CJ, Ma B, Nussinov R (2009) Protein-protein interaction networks: how can a hub protein bind so many different partners? Trends Biochem Sci 34(12):594–600
5. Kuriyan J, Eisenberg D (2007) The origin of protein interactions and allostery in colocalization. Nature 450(7172):983–990
6. Motlagh HN et al (2014) The ensemble nature of allostery. Nature 508(7496):331–339
7. Bagler G, Sinha S (2007) Assortative mixing in Protein Contact Networks and protein folding kinetics. Bioinformatics 23(14):1760–1767
8. Piovesan D, Minervini G, Tosatto SC (2016) The RING 2.0 web server for high quality residue interaction networks. Nucleic Acids Res 44(W1):W367–W374
9. Di Paola L et al (2013) Protein contact networks: an emerging paradigm in chemistry. Chem Rev 113(3):1598–1613
10. Bhattacharyya M, Ghosh S, Vishveshwara S (2016) Protein structure and function: looking through the network of side-chain Interactions. Curr Protein Pept Sci 17(1):4–25
11. Vishveshwara S, Ghosh A, Hansia P (2009) Intra and inter-molecular communications through protein structure network. Curr Protein Pept Sci 10(2):146–160
12. Vishveshwara S, Brinda KV, Kannan N (2002) Protein structure: insights from graph theory. J Theor Comput Chem 01(01):187–211
13. Gadiyaram V, Ghosh S, Vishveshwara S (2017) A graph spectral-based scoring scheme for network comparison. J Complex Networks 5 (2):219–244
14. Ghosh S, Gadiyaram V, Vishveshwara S (2017) Validation of protein structure models using network similarity score. Proteins 85 (9):1759–1776
15. Gadiyaram V, Dighe A, Vishveshwara S (2019) Identification of crucial elements for network integrity: a perturbation approach through graph spectral method. Int J Adv Eng Sci Appl Math 11:91–104
16. Vijayabaskar MS, Vidya N, Saraswathi V (2011) GraProStr—graphs of protein structures: a tool for constructing graphs and generating graph parameters for protein structures
17. Adamcsek B et al (2006) CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 22(8):1021–1023
18. Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1 (1):269–271
19. Bhattacharyya M, Bhat CR, Vishveshwara S (2013) An automated approach to network features of protein structure ensembles. Protein Sci 22(10):1399–1416
20. Patra SM, Vishveshwara S (2000) Backbone cluster identification in proteins by a graph theoretical method. Biophys Chem 84 (1):13–25
21. Vendruscolo M et al (2002) Small-world view of the amino acids that play a key role in protein folding. Phys Rev E Stat Nonlinear Soft Matter Phys 65(6 Pt 1):061910
22. Brinda KV, Kannan N, Vishveshwara S (2002) Analysis of homodimeric protein interfaces by graph-spectral methods. Protein Eng 15 (4):265–277
23. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18 (3):534–552
24. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 256(3):623–644
25. Kannan N, Vishveshwara S (1999) Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol 292 (2):441–464
26. Bhattacharya S, Vaidehi N (2014) Differences in allosteric communication pipelines in the inactive and active states of a GPCR. Biophys J 107(2):422–434
27. Ghosh A, Vishveshwara S (2008) Variations in clique and community patterns in protein

structures during allosteric communication: investigation of dynamically equilibrated structures of methionyl tRNA synthetase complexes. Biochemistry 47(44):11398–11407

28. Bhattacharyya M et al (2010) Allostery and conformational free energy changes in human tryptophanyl-tRNA synthetase from essential dynamics and structure networks. Proteins 78 (3):506–517
29. Sistla RK, Brinda KV, Vishveshwara S (2005) Identification of domains and domain interface residues in multidomain proteins from graph spectral method. Proteins 59(3):616–626
30. Kannan N et al (2001) Stabilizing interactions in the dimer interface of alpha-subunit in Escherichia coli RNA polymerase: a graph spectral and point mutation study. Protein Sci 10 (1):46–54
31. Brinda KV, Surolia A, Vishveshwara S (2005) Insights into the quaternary association of proteins through structure graphs: a case study of lectins. Biochem J 391(Pt 1):1–15
32. Reichmann D et al (2005) The modular architecture of protein-protein binding interfaces. Proc Natl Acad Sci U S A 102(1):57–62
33. Trzaskowski B et al (2012) Action of molecular switches in GPCRs—theoretical and experimental studies. Curr Med Chem 19 (8):1090–1109
34. Wettschureck N, Offermanns S (2005) Mammalian G proteins and their cell type specific functions. Physiol Rev 85(4):1159–1204
35. Hazell GG et al (2012) G protein-coupled receptors in the hypothalamic paraventricular and supraoptic nuclei—serpentine gateways to neuroendocrine homeostasis. Front Neuroendocrinol 33(1):45–66
36. Sharma N, Akhade AS, Qadri A (2013) Sphingosine-1-phosphate suppresses TLR-induced CXCL8 secretion from human T cells. J Leukoc Biol 93(4):521–528
37. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5(12):993–996
38. Rasmussen SG et al (2011) Crystal structure of the beta2 adrenergic receptor-Gs protein complex. Nature 477(7366):549–555
39. Wacker D et al (2010) Conserved binding mode of human beta2 adrenergic receptor inverse agonists and antagonist revealed by X-ray crystallography. J Am Chem Soc 132 (33):11443–11445
40. Dixon RA et al (1986) Cloning of the gene and cDNA for mammalian beta-adrenergic receptor and homology with rhodopsin. Nature 321 (6065):75–79
41. Rasmussen SG et al (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature 450(7168):383–387
42. Ballesteros JA, Weinstein H (1995) Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In: Stuart CS (ed) Methods in neurosciences. Academic, New York, pp 366–428
43. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 8(2):127–134
44. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453

Chapter 8

Topology Results on Adjacent Amino Acid Networks of Oligomeric Proteins

Claire Lesieur and Laurent Vuillon

Abstract

In this chapter, we focus on topology measurements of the adjacent amino acid networks for a data set of oligomeric proteins and some of its subnetworks. The aim is to present many mathematical tools in order to understand the structures of proteins implicitly coded in such networks and subnetworks. We mainly investigate four important networks by computing the number of connected components, the degree distribution, and assortativity measures. We compare each result in order to prove that the four networks have quite independent topologies.

Key words Adjacent amino acid network, Topology of graphs associated with proteins, Subnetworks, Protein contact network, Long range network, Hot spot network, Induced hot spot network, Connected components, Degree distribution, Assortativity measures

1 Introduction

The main goal of this chapter is to present mathematical tools to investigate the structure of protein using network topology. It is well known that proteins, oligomeric proteins in our case study, have geometrical control over their shapes (folding) and functions (dynamics) based on four structural levels, so-called 1D, 2D, 3D, and 4D structural levels [1, 2]. Proteins have a 1D structure which is the sequence of amino acids involved in each chain; this sequence is by definition one-dimensional and has an intrinsic ordering coming from the addition upon protein synthesis, of the amino acids one by one via peptide bonds. The next remarkable geometrical structure comes from the 2D structure of the protein that is the ability of constructing local geometrical structures like alpha helix or beta sheet; these structures are rather local and involve nearest amino acids of the 1D structure. The folding mechanism is also crucial to construct the 3D (tridimensional) structure by bringing into contacts amino acids, which are far in the 1D structure but become close in space. By this step, the 2D structures are connected

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_8,

to one another to form a 3D shape. The last important structuration is related to the construction of interfaces between chains in order to construct a protein oligomer; the 4D structure is the number of copies of a chain, which composed the oligomer (see the tiling theory for proteins in [3]). Not all proteins have a 4D structure.

We propose to investigate these geometrical structures by considering adjacent amino acid properties and networks. This transformation from geometrical structure to networks using a cutoff allows us to ignore geometrical positions and focus only on the topology of the proteins, that is, the connectivity of amino acids given by the network of adjacent amino acids, distant by a cutoff of 5 Å. In network theory, many measurements exist to investigate topology, among which we will use the following:

- The Connected Components of the network or any sub-networks relevant to the structural hierarchy considering the Stoichiometry (4D structure) (*see* [4]),
- The Degree Distribution of the nodes of the network (*see* [4]),
- The Assortativity gives a global measurement for the local topology of the networks and is linked to folding rates in protein networks and more generally is involved in the spreading rates of error propagation or of diseases (*see* [4, 5, 6, 7, 8]).

The different structural levels that compose a protein can be respectively modeled by different networks of adjacent amino acids. The most studied network is the Adjacent Amino Acid Network (also called in the literature Protein Contact Network or protein structure network) (*see* DPB, [2, 9, 10]) where the nodes are amino acids and the links between two amino acids are when amino acid are at distant less or equal to a cutoff value (here we use 5 Å) [11]. This network gives information on the whole structure of the protein and we use PCN to denote the Adjacent Amino Acid Network.

In addition, we investigate three other networks: the HSN, the IHSN, and the LRN. The Hot Spot Network (HSN) represents the amino acids involved in the interface between chains in oligomeric proteins codes for the 4D structure (*see* [2, 12] and references within (we call a Hot Spot, an amino acid, which is linked to at least one amino acid from another chain). The Induced Hot Spot Network (IHSN) constructed from the Hot Spots and all amino acids in contact with hot spots in the chain (with distant less or equal to 5 Å). The IHSN gives the contacts of the hot spots with other hot spots (contacts between chains) and with amino acids within a chain. Thus HSN provides only the 4D contacts of the interface while IHSN provides all contacts information of the hot spots. The Long Range Network (LRN), which is constructed for each chain considering two amino acids in spatial contact in the

protein and far in the chain according to the sequence ordering (that is with at least with 7 amino acids in the sequence between the two considered amino acids). The Long Range Network codes for the 3D structure.

The HSN, IHSN, and LRN are sub-networks of the PCN modeling the contacts involved in the different structural levels of a protein. We investigate if these networks have different topological properties, considering the connected components, the degree distribution, and the assortativity, which suggest the different structural levels have some specificity, coding, and structural independency.

The chapter is divided in five sections. The first describes the proteins of the dataset and the networks; the second defines the four networks; the connected components, the degree distribution, and the assortativity are provided in the third, fourth, and fifth sections.

2 Proteins and Networks

In this chapter, we focus on oligomeric proteins, which are composed of k times the same chain of amino acids (k is called the quaternary structure of a protein). The amino acid residues of different chains, called Hot Spots, are connected together by chemical and physical interactions to form the atomic interfaces that build the oligomer. Thus usually we name the k chains by distinct letters in an alphabetic order (for a pentamer the names of the five chains could be A, B, C, D, E or D, E, F, G, H or other combinations of consecutive letters in an alphabetic order). Each chain is formed by a sequence of amino acids, which corresponds to the order in which the amino acids are covalently attached to one another (1D structure). The ordering from the sequence which is encoded at the DNA level gives a natural ordering of the amino acids in each chain; this ordering is given by a sequence of consecutive integers (thus we could speak of the amino acid number 50 in the chain B and for short we use the notation 50:B [or B50 in this chapter] for this amino acid with the concatenation of the name of the chain with the position of the amino acid in the chain). The 20 amino acids with special chemical and geometrical properties are constituted by a set of covalent atoms. In order to understand the construction of oligomeric proteins, we propose to define formally several networks and investigate their topology. Thus, we use mathematical tools and theory of networks to study biological constructions and coding.

The main mathematical tool for our study is a network (also called graph) $G = (N, L)$ with nodes (also called vertices) noted N and links (also called edges) noted L. This graph is constructed by a Protein Data Base file associated with a given protein (Fig. 1) from

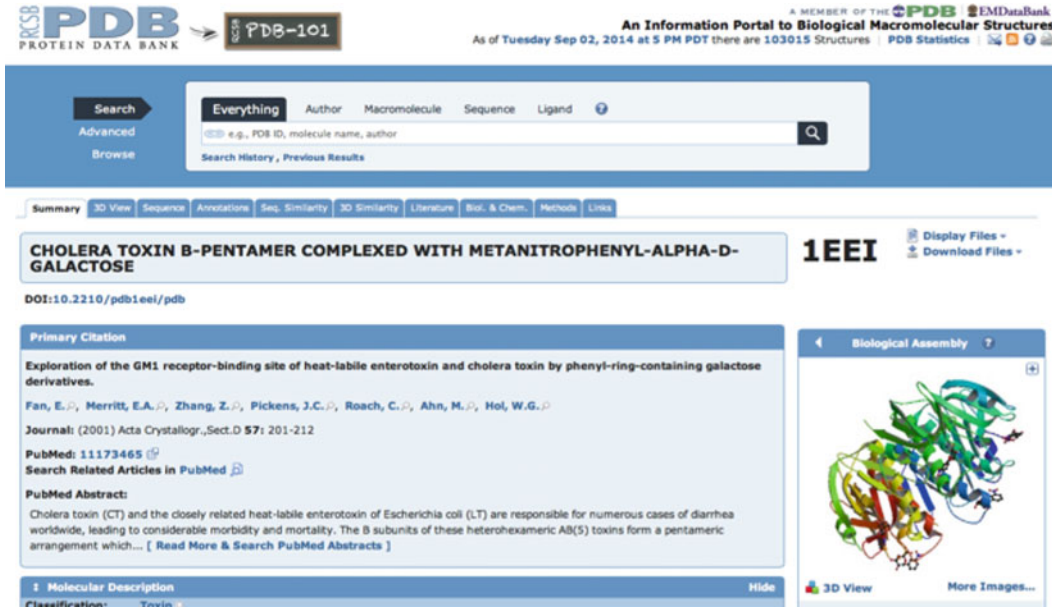

Fig. 1 Screenshot of one page of the Protein Data Bank

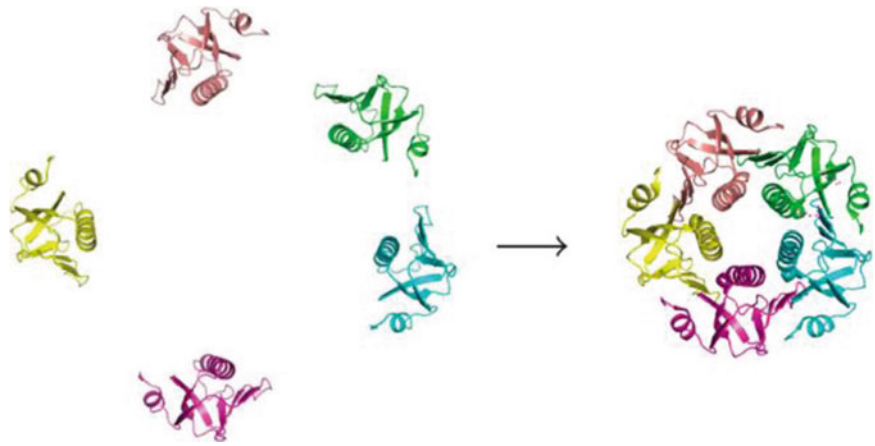

Fig. 2 Cyclic shape of the cholera toxin

atomic coordinates according to the X-ray structure of the protein. To illustrate the notion, we use the 1EEI.pdb file, which is the crystallographic version of the Cholera Toxin (*see* [13]).

The Cholera Toxin is composed by 5 identical chains and the whole protein has a cyclic shape (Fig. 2).

Nevertheless, the whole study is done using a data set of 750 oligomeric proteins from 2 up to 20 chains (Fig. 3).

For each protein, we take the associated PDB file and we consider all the atoms and their spatial positions in the three dimensional space given in Angstroms. The Adjacent Amino Acid

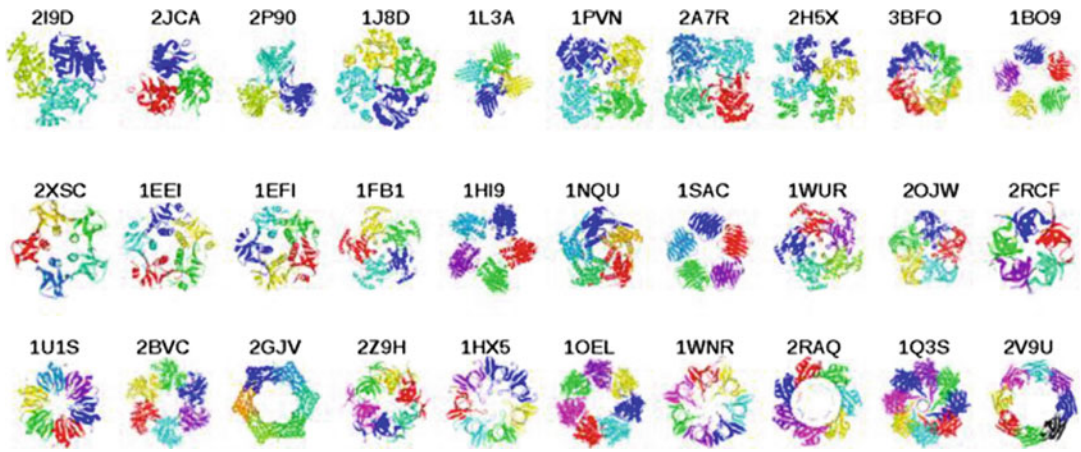

Fig. 3 Part of the Data Set (*see* [12])

Network PCN(N,L) is constructed considering the amino acids of the protein as set of nodes N and, for every two nodes in N there exists a link in L between these two amino acids (two nodes) if the two amino acids have at least one atom each in proximity, namely at a distance less than 5 Å. In fact, a node in the network represents a union of atoms of a given amino acid and a link between nodes means that two amino acids are at spatial distance less than 5 A (*see* [2, 10]).

In Fig. 4, we find the representation of the Cholera Toxin by plotting all atoms.

The PDB file contains many information and in particular the atomic (x, y, z) coordinates of each atom and this is useful to compute the distance between amino acids (Fig. 5).

By computing the Euclidian distance between two amino acids and for all of them using a cut off at 5 Å, we construct the Adjacent Amino Acid Network of each protein (*see* Fig. 6 for the Adjacent Amino Acid Network of the Cholera Toxin).

We are able to represent graphically the network, and investigate its geometry and topology. Thus the amino acids could be near in a chain (called intramolecular residues) or between at least two chains (called intermolecular residues) (Fig. 7). We remark, in Fig. 7, that D39 is a Hot Spot because it is linked to the amino acid 2, 3, 4, and 6 of the H chain.

3 The Four Networks

This chapter is based on the Adjacent Amino Acid Network PCN (N,L) defined in Subheading 2. And we would like to investigate sub-networks or induced networks defined formally in the next paragraphs.

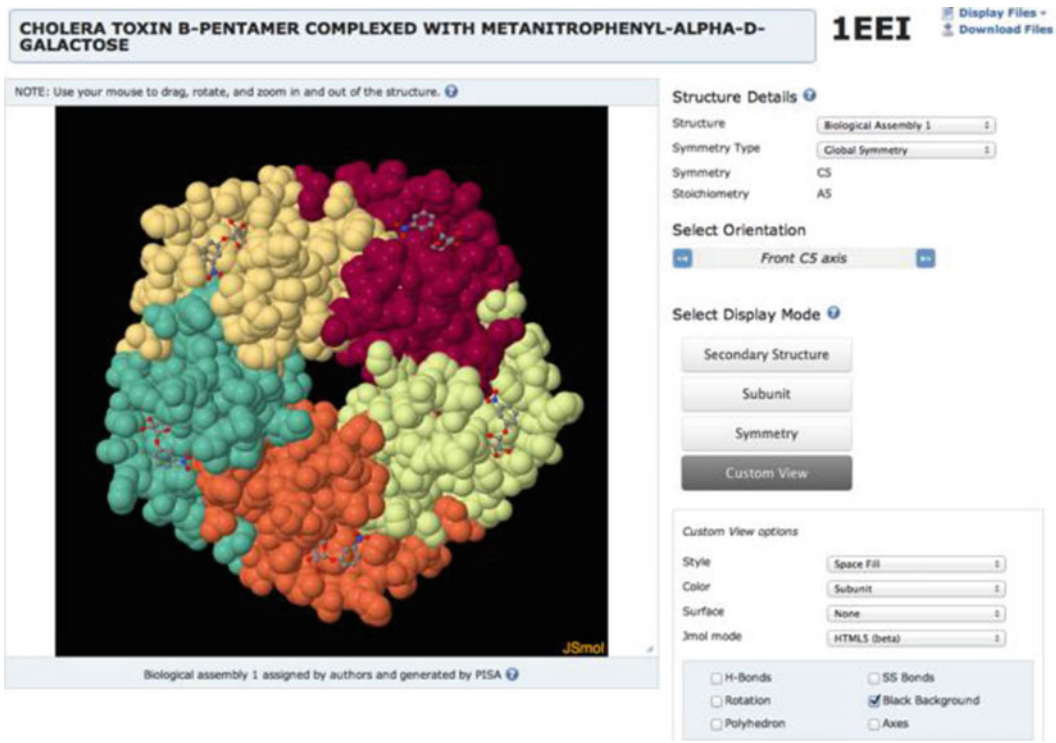

Fig. 4 Atomic representation of the cholera toxin in spacefill (van der Waals atomic radius is represented by spheres)

```
ATOM      1  N    THR D   1      38.037  31.348  29.862  1.00 33.00           N
ATOM      2  CA   THR D   1      36.589  31.127  30.121  1.00 35.55           C
ATOM      3  C    THR D   1      35.742  31.862  29.096  1.00 33.96           C
ATOM      4  O    THR D   1      35.831  33.092  28.952  1.00 32.50           O
ATOM      5  CB   THR D   1      36.193  31.558  31.545  1.00 35.17           C
ATOM      6  OG1  THR D   1      36.841  30.686  32.469  1.00 39.57           O
ATOM      7  CG2  THR D   1      34.687  31.426  31.763  1.00 33.50           C
ATOM      8  N    PRO D   2      34.892  31.113  28.372  1.00 33.37           N
ATOM      9  CA   PRO D   2      34.031  31.737  27.351  1.00 34.14           C
ATOM     10  C    PRO D   2      33.093  32.795  27.950  1.00 31.05           C
ATOM     11  O    PRO D   2      32.735  32.731  29.127  1.00 32.31           O
ATOM     12  CB   PRO D   2      33.273  30.536  26.746  1.00 34.27           C
ATOM     13  CG   PRO D   2      33.247  29.524  27.908  1.00 35.08           C
ATOM     14  CD   PRO D   2      34.641  29.662  28.498  1.00 29.79           C
ATOM     15  N    GLN D   3      32.710  33.770  27.141  1.00 29.08           N
ATOM     16  CA   GLN D   3      31.830  34.806  27.619  1.00 30.28           C
ATOM     17  C    GLN D   3      30.464  34.666  27.011  1.00 27.08           C
ATOM     18  O    GLN D   3      29.554  35.366  27.398  1.00 25.73           O
ATOM     19  CB   GLN D   3      32.414  36.182  27.330  1.00 39.02           C
ATOM     20  CG   GLN D   3      33.767  36.404  28.018  1.00 48.24           C
ATOM     21  CD   GLN D   3      33.898  37.791  28.618  1.00 51.86           C
ATOM     22  OE1  GLN D   3      32.950  38.339  29.188  1.00 54.42           O
ATOM     23  NE2  GLN D   3      35.083  38.370  28.499  1.00 53.78           N
ATOM     24  N    ASN D   4      30.283  33.673  26.156  1.00 24.18           N
ATOM     25  CA   ASN D   4      29.007  33.476  25.504  1.00 23.35           C
ATOM     26  C    ASN D   4      28.898  32.053  24.991  1.00 24.80           C
ATOM     27  O    ASN D   4      29.885  31.318  24.971  1.00 25.80           O
ATOM     28  CB   ASN D   4      28.881  34.449  24.343  1.00 25.31           C
ATOM     29  CG   ASN D   4      30.015  34.307  23.351  1.00 26.76           C
ATOM     30  OD1  ASN D   4      30.162  33.270  22.716  1.00 28.05           O
ATOM     31  ND2  ASN D   4      30.851  35.332  23.235  1.00 28.85           N
```

Fig. 5 List of atoms and atomic (*x, y, z*) coordinates

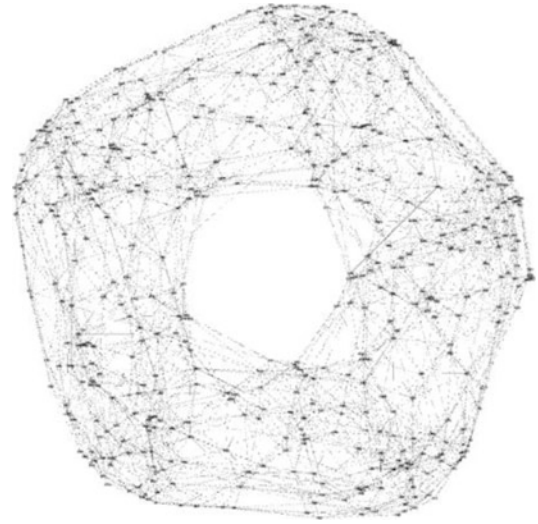

Fig. 6 PCN for 1EEI (representation of the network by Gephi)

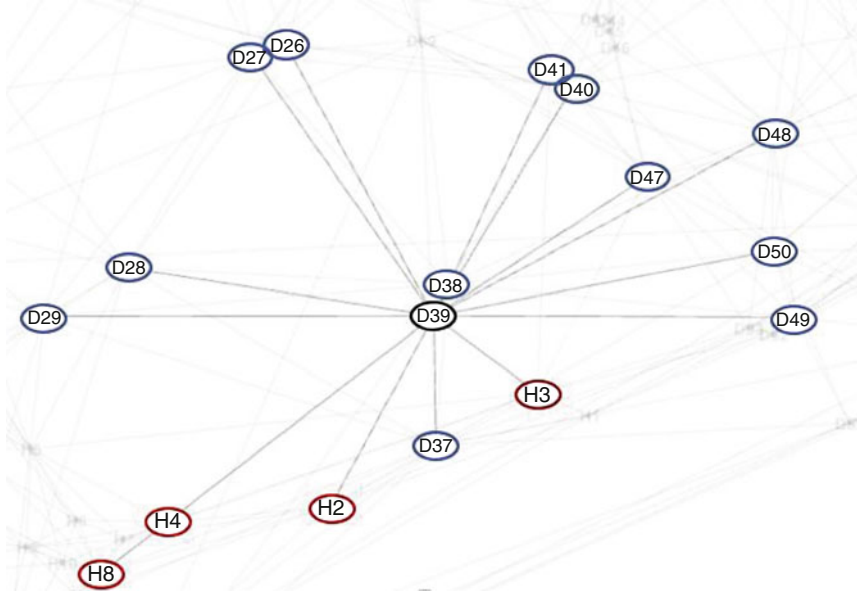

Fig. 7 Zoom in the 1EEI PCN. First neighbors of D39 (Black circle) with amino acids in the D chain (Blue circle, intramolecular residues) and in the H chain (Red circle, intermolecular residues); D39, H2, H3, H4 and H6 are Hot Spots (Amino Acids in the interface)

The general definition of an induced network IN(N′,L′) of PCN(N,L) is a subset N′ of nodes in N, and there is a link in the induced network between n,m ∈ N′ if there exists a link between the nodes n and m in PCN(N,L) (see for example the textbook [14]). Thus, the induced graph is a subgraph of nodes N′ and all links in L but only between nodes in N′.

- PCN: Adjacent Amino Acid Network: the whole adjacent amino acid network.
- HSN Hot Spot Network: the nodes are Hot Spots and the links are between two nodes in different chains (each amino acid involves in the interfaces). This network models the 4D structure.
- IHSN Induced Hot Spot Network: the induced graph of amino acids involved in the interfaces. The nodes are Hot Spots (subset N′ of the N nodes in PCN) and the links of the induced network are all links between Hot Spot in the PCN network. This network is given by the Hot Spots and all links between them; thus, the links between hotspots within the same chain are present while in the HSN they are not.
- LRN Long Range Network: For each amino acid i in a single chain, we take all amino acids j on the same chain with distant position in the sequence of at least 8 amino acids (this is with the distance on the sequence $|i - j| > 7$) that are linked in PCN and thus we add in the network all links in PCN from i to j and the amino acids i and j such that $|i - j| > 7$. For example, if we consider the chain A and the amino acid in position 10 then we add in the network the amino acids of the chain A in position 1, 2, and the 18, 19, and so on that are linked with the amino acid 10 in PCN. This network models the 3D structure of the protein.

All these networks have many distinct topological properties and may be involved in some important structural mechanisms of the proteins.

The Adjacent Amino Acid Network gives information on the whole structure of the protein and it is based the sub-networks that control the 2D, 3D, and the 4D structures. Comparing the four different networks PCN, HSN, IHSN. and LRN aims at investigating the independency of the structural level coded by each of the sub-network, that is, to what extent the structural levels have their own topological characteristics.

We are particularly interested in comparing the topological properties of the interface with the rest of the protein because the Hot Spot Network is made of only 10% or less of the total amino acids of the protein, making the interface potentially fragile. Since protein oligomers are not more susceptible to mutations than monomeric proteins, it suggests the existence of a mechanism to increase the interface robustness and protect it from harm. IHSN can be compared with the Adjacent Amino Acid Network (PCN) and with the Hot Spot Networks to study how the protein masks the position of the interface or to investigate the backup of the interface, respectively [15]. Backups are additional links that make the interface more robust to mutation and to special perturbation

of atomic motion. More generally, we could compare IHSN with PCN to test the hypothesis that there is a mechanism (topological mechanism) that protects the interface: prevent perturbations from reaching it (mutations of non-hotspot amino acids) or allow correcting errors introduced by mutations of hotspots.

The Long Range Network codes for the 3D structure and could be compared to the HSN to see if the 4D and the 3D topologies are independent. We could test different assembly mechanisms such as the fly casting mechanism where interface formation and folding happen concomitantly while the induce-fit mechanism folding occurs first ([16] and references within, [17]). The Long Range Network topology is known to have a strong correlation with the folding rate of the protein (*see* [5]) and remark that in the paper they consider amino acids in the sequence with distance greater than 11 while in our construction we use a distance greater than 7).

4 Stoichiometry and Connected Components

First of all, we investigate the stoichiometry and the number of connected components of the four networks in order to understand the topology of proteins in the data set. In fact, the stoichiometry is an information on the number of chains used for the protein construction. A connected component of a network is the set of nodes that are connected together by paths following the links. By definition PCN has only one connected component because an oligomeric protein is composed of a single protein. HSN could have only one connected component; that means that all the interfaces are connected together by paths following the links. More interestingly if there are 2 or more connected components in HSN this means that some interfaces are relatively far from each other in the protein.

In our data set, we study 750 oligomeric proteins with stoichiometry (the number of chains) from 2 up to 20 (*see* [15]). In fact, in the PDB sometimes the protein appears with copy of itself, and thus the number of chains and connected components increases numerically; thus, we just discard these cases and reduce the number of cases. Only 714 PDB files appear with 1 connected component for PCN; this means with a single protein in the PDB, 30 PDB files with two copies of itself and 2 PDB files with three copies of itself. If we restrict to the 714 PDB files with a description of a single protein, we notice that 218 proteins have 4 chains, 168 proteins have 2 chains, 116 proteins have 3 chains, 70 proteins have 6 chains, 46 proteins have 6 chains, 30 proteins have 12 chains, 23 proteins have 5 chains, and so on (Fig. 8). Remember that the number of divisors of the stoichiometry k is important because if k is composed by many divisors the combinatorics give more proteins than

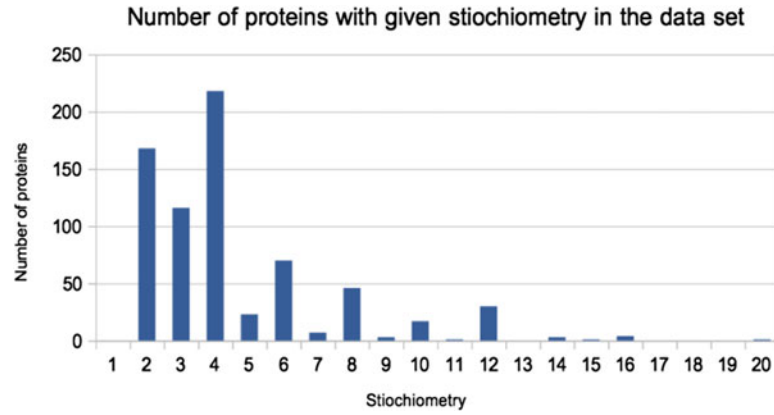

Fig. 8 The different stoichiometries in the Data Set

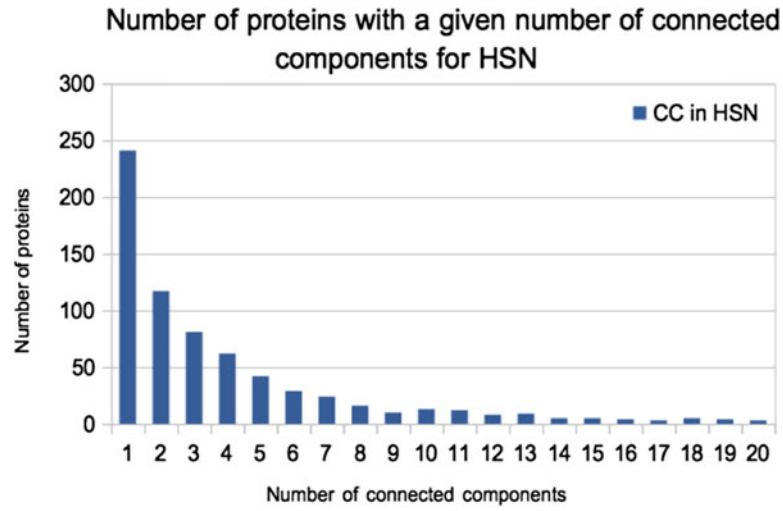

Fig. 9 The number of connected components for HSN

prime numbers (*see* [15]). We have only 1 proteins with 11 chains and in contrast 30 proteins with 12 chains because 12 could be written $12 = 2 \times 6 = 3 \times 4 = 3 \times 2 \times 2$; this means a cyclic protein with 12 chains or a dimer of 6 chains or a trimer of 4 chains and so one.

Interestingly if we look carefully at interfaces between chains in the data set that is for the Hot Spot Network, we find that 241 proteins have a single connected component (that is all existing interfaces between chains are connected by links between hot spots (*see* Fig. 14 to see this property of single connected component for the Cholera Toxin), 117 proteins with 2 connected components, 81 proteins with 3 connected components, 62 proteins with 4 connected components, and so on (*see* Fig. 9). In this study, we

give no insight on the topology of the connected components of the Hot Spot Network and for example the single connected component could be as a ring or not.

Thus we find a nice power law for the number of connected components for HSN (by computing directly on the graphic we fit the function $f(x) = 388x^{-1.56}$ with $R^2 = 0.963$).

We notice that proteins (1EEI, 1LTR, 1EFI, the latter with unknown assembly mechanism) which share similar structures but have different assembly mechanisms (*see* [18]) nevertheless have the same number of CC in the four networks: 1 connected component for PCN, HSN, IHSN and 10 connected components for LRN. This means that the number of connected components is a rather global measurement, which does not account for the way the protein is built. Globally the interfaces between chains of the three "versions" of the cholera toxin are linked together to form a whole global interface network. On the other hand, the three toxins 1EEI, 1LTR, and 1EFI, all AB_5 toxins, have their 3D structures organized in two domains, indicated by the ten CC.

For the Induced Hot Spot Network, we find 568 proteins with 1 connected component, 55 with 2 connected components, 38 with 3 connected components, 28 with 4 connected components, and 6 with 5 connected components (Fig. 10). Thus, if the IHSN network has mostly a single connected component this means that the Induced Hot Spot Network globally controls the spatial position of the interfaces in the whole protein. And to be robust to mutation it is better to link all the k interfaces in a single connected component. While in article [12], we argued that too many connections lead to fragility to perturbations, here we have no measurement on the density of the connections in the network.

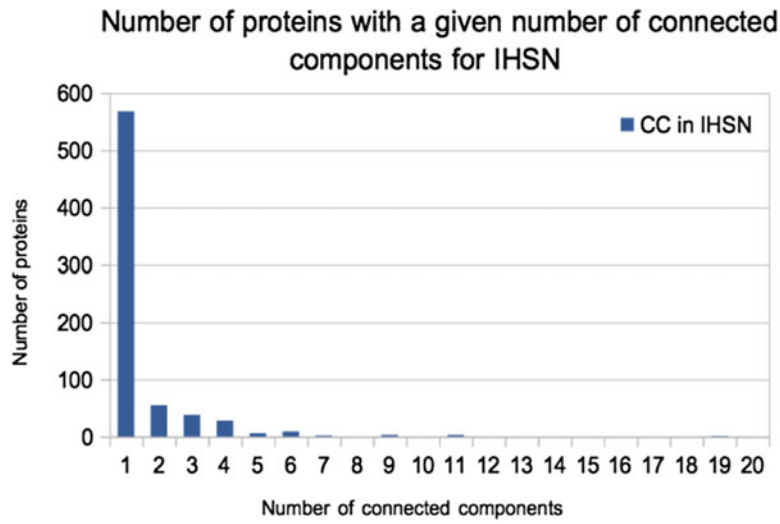

Fig. 10 The number of connected components for IHSN

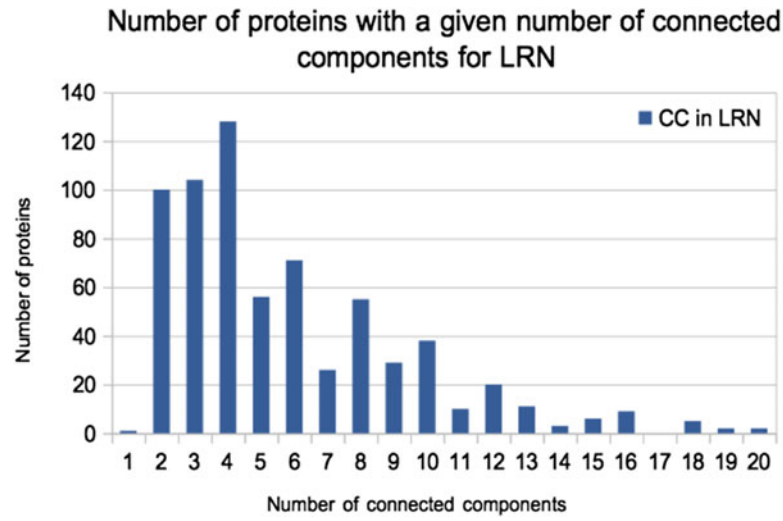

Fig. 11 The number of connected components for LRN

We just conclude by a statistical argument that $568/715 = 79\%$ of the data set has a single connected component for IHSN, and this property implies that proteins prefer to adopt a single CC in order to control the spatial position of the interfaces.

For the Long Range Network, we find equal or more connected components than the stoichiometry, and this means that in a same chain the network of the 3D contacts could be split in different subnetworks and these subnetworks are independent. This means that we find with the connected components of LRN some independent 3D structures and we must make mutations in order to prove that these connected components control the folding domains of the protein (Fig. 11).

5 Degree Distribution

The degree distribution of the nodes gives information on the global structure of each network. They also reflect the laws of construction and of the network emergence or of subnetwork emergence that is the way of building each part of the Adjacent Amino Acid Network.

5.1 Degree Distribution in Adjacent Amino Acid Network

We already constructed the whole adjacent amino-acid networks (Fig. 6) and we would like to investigate the degree distribution of the nodes. We find for the cholera toxin 1EEI the following distribution of degrees for PCN (Fig. 12). While the degree distribution of the whole data set is near a Gaussian (Fig. 13) to construct the plot, the occurrences of each degree are summed over the entire dataset for all proteins; for example, we have almost 80,000 nodes

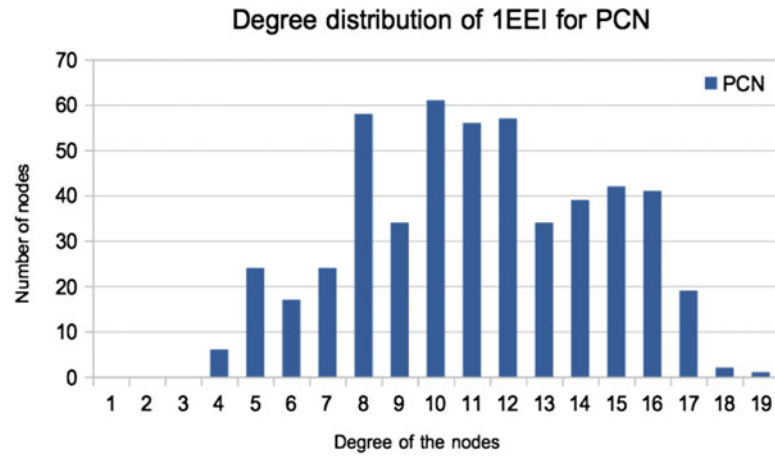

Fig. 12 Degree distribution of PCN for the Cholera Toxin

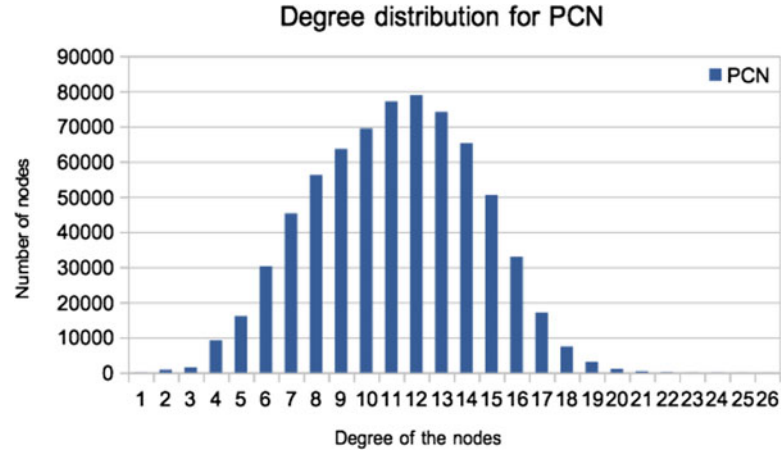

Fig. 13 Degree distribution of PCN on the whole Data Set

in the Data Set with degree 12. This global distribution means that the degrees are driven by a normal distribution to have a majority of moderate degrees and not too many high or low degrees. While each node could control the 2D, 3D, and/or 4D structures of the protein, the normal distribution of degrees masks the role of each node (that is, we are not able to predict the structural role of each amino just by looking at its degree).

5.2 Degree Distribution in Hot Spot Network

The Hot Spot Network is constructed by adjacent amino acids in two different chains and compose the interfaces of the protein. We remark that this network between two chains is bipartite: that is between two chains the first neighbors of a given Hot Spot are Hot Spots of the other chain (Fig. 14). Indeed, two nodes of a link in this network fall by construction in two distinct chains. We first

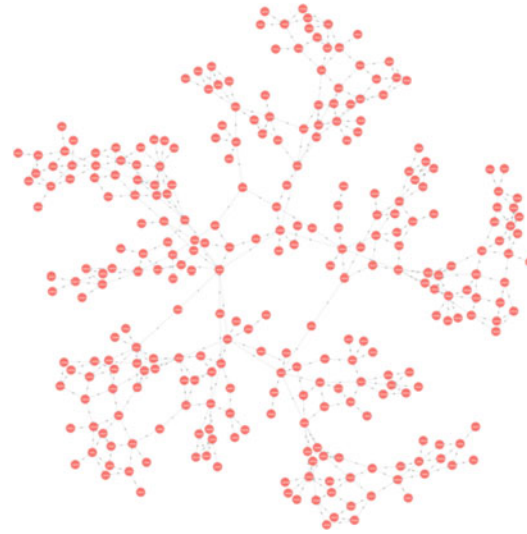

Fig. 14 Hot Spot Network for the Cholera Toxin

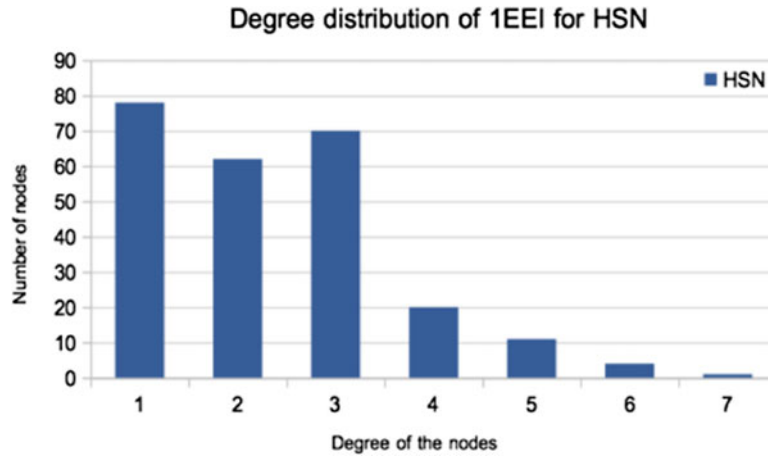

Fig. 15 Degree distribution of HSN for the Cholera Toxin

investigate the cholera toxin 1EEI and we find in Fig. 15 the distribution of degrees for HSN. The degree distribution of the HSN for the whole data set is an exponential decrease law (Fig. 16) (by computing directly on the graphic we fit the function $f(x) = 491568e^{-0.9264x}$ with $R^2 = 0.97$). Notice that the distribution is not anymore Gaussian-like in the whole proteins, and the exponential decrease reflects the very special structure of the Hot Spot Network. We already notice in [12, 19] that for interfaces between two chains the degree distribution is exponential; this means no hub in the interface but quite low degrees. For the PCN, we find a Gaussian distribution, so this means considering

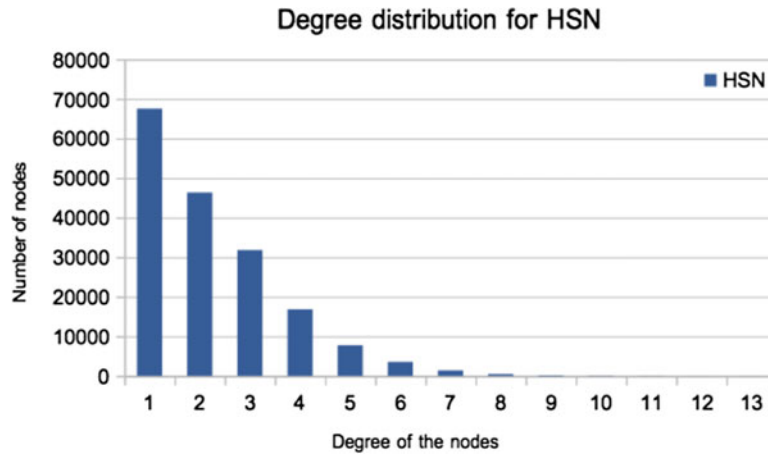

Fig. 16 Degree distribution of HSN for the Data Set

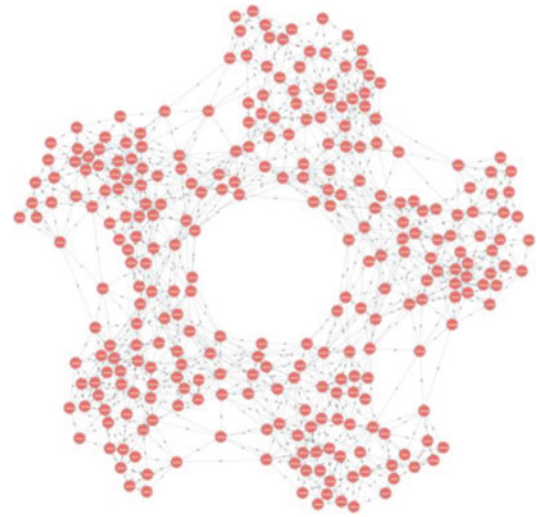

Fig. 17 Induced Hot Spot Network for the Cholera Toxin

all contacts, the amino acids adopt a characteristic moderate degree right with a mode (degree with highest frequency in the distribution) and a mean around 12. This means that many links are added to the Hot Spots outside of the Hot Spot Network to reach the PCN degree distribution. This mechanism is crucial to mask the position of the Hot Spots.

5.3 Degree Distribution in Induced Hot Spot Network

The Induced Hot Spot Network of amino acids involves in the interfaces is represented in Fig. 17. We find for the cholera toxin 1EEI the following distribution of degrees for IHSN (Fig. 18). We notice that the degree distribution of the Induced Hot Spot Network is also near a Gaussian (Fig. 19). Remark that the mode for

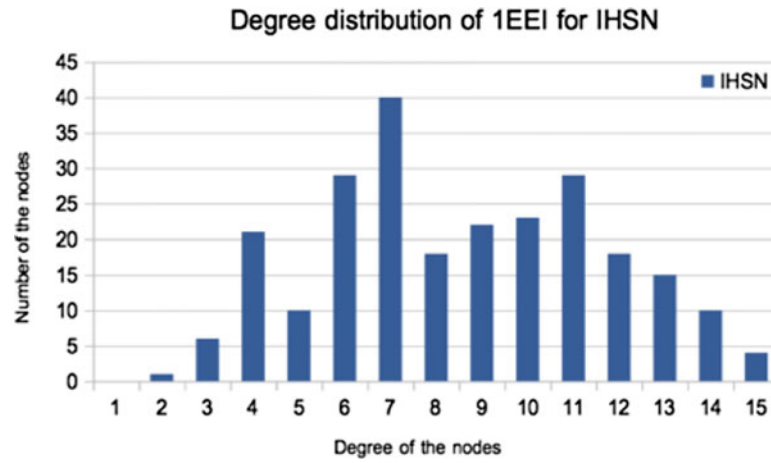

Fig. 18 Degree distribution of IHSN for the Cholera Toxin

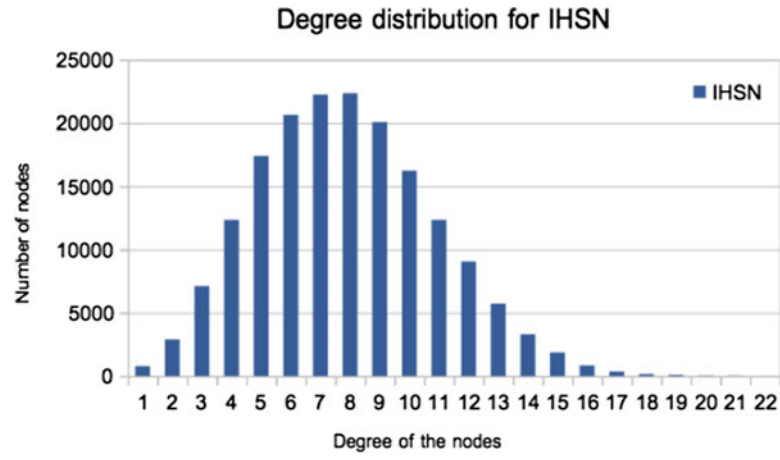

Fig. 19 Degree distribution of IHSN for the whole Data Set

IHSN is in 8 while the mode for PCN is 12. Thus, we find a shift of 4 degrees in the degree values between the Gaussian of PCN to the Gaussian of IHSN. This means that in means 8 links are devoted to control the interfaces and in means 4 more links control the neighborhood to mask the position of the interfaces and to plunge the sub-network IHSN in the whole PCN network.

5.4 Degree Distribution in Long Range Network

For the Long Range Network, we take for all pairs of amino acids in the same chain at positions i and j in the sequence with $|i - j| > 7$ (the distance in the sequence is greater or equal to 8) and such that the amino acid i and j are also linked in PCN and thus we construct the Long Range Network (Fig. 20). For the cholera toxin 1EEI we find in Fig. 21 the distribution of degrees for LRN. The degree distribution of the whole data set is not any more a Gaussian and

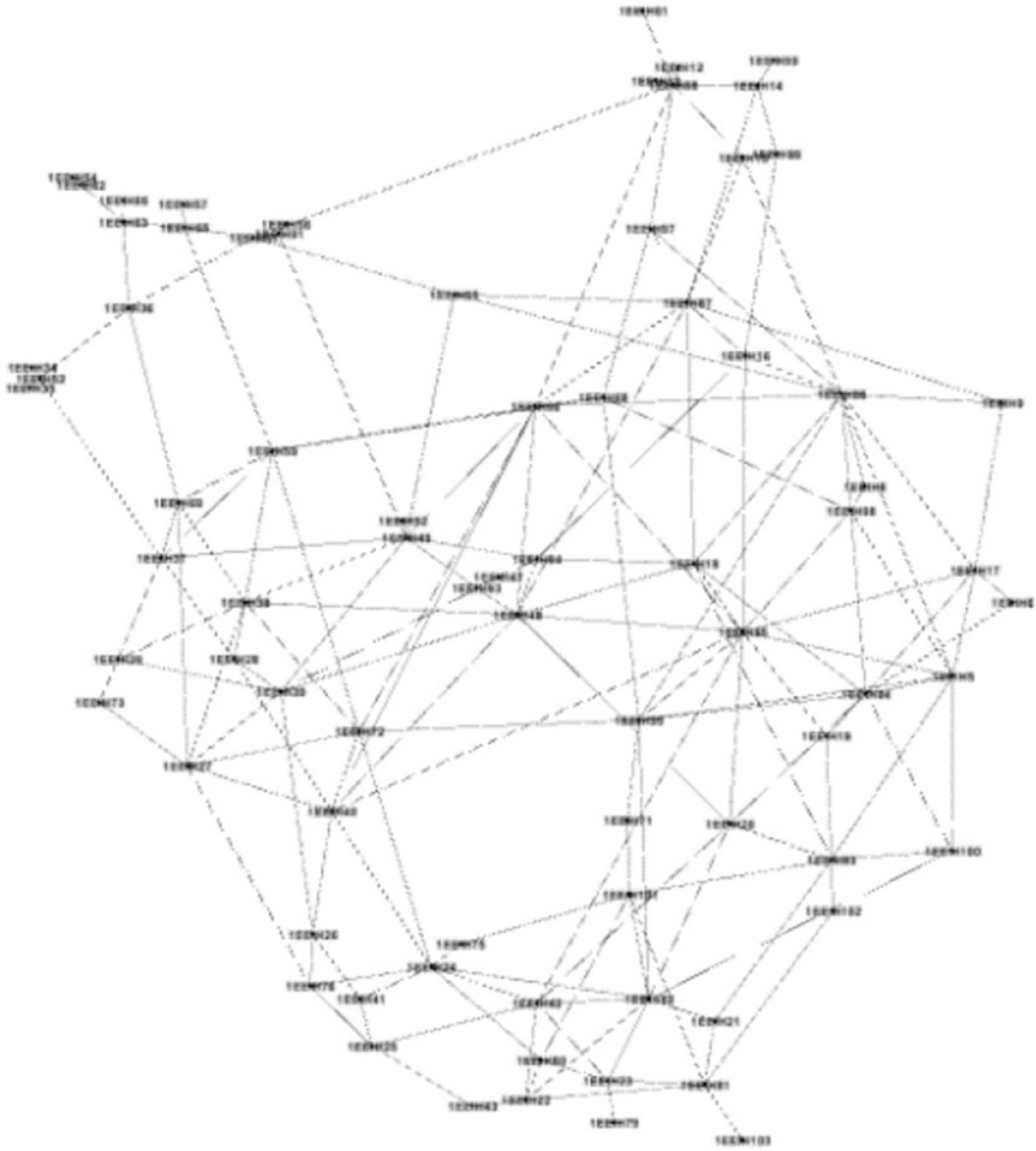

Fig. 20 Long Range Network for the Cholera Toxin

the mode is now 4 (*see* Fig. 22). In fact, we notice that the LRN is constructed with less links than the PCN and we have to add in mean 8 links to reach the PCN degree distribution. This implies that in order to mask the position of the LRN and to reach a Gaussian distribution we add in mean 8 links.

To conclude these parts on the topology of each network and subnetwork, we have seen that the subnetworks have really distinct behaviors and are embedded on the PCN with a mask property of each subnetwork. We now want to describe more precisely the

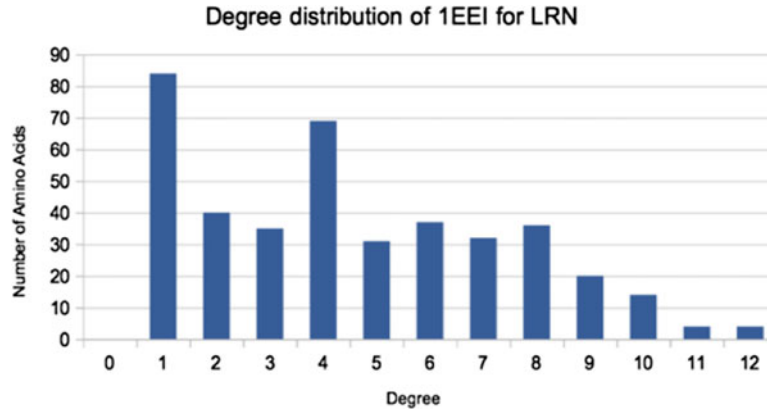

Fig. 21 Degree distribution of LRN for the Cholera Toxin

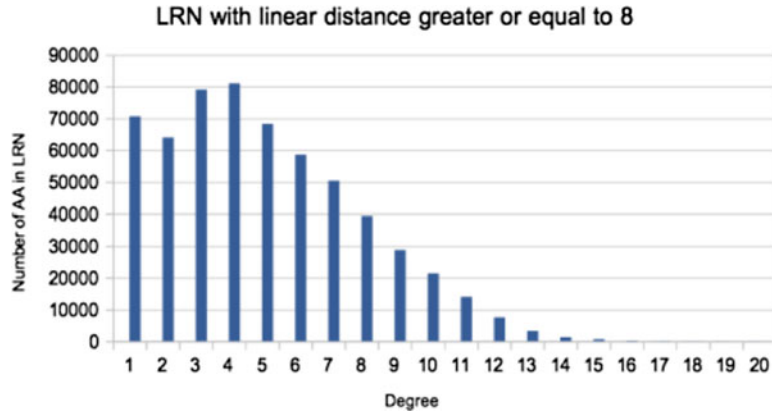

Fig. 22 Degree distribution of LRN for the whole data set

structure of each network in terms of assortativity, and we will focus on the relative independence of each network.

6 Assortativity Measures

We investigate the topology of various networks by considering the correlation of degree defined by Newman [7, 8]. The idea is to compare globally the difference between the degree of each node to the mean degree of its first neighbors. To do that we compute for each protein the degree correlation coefficient:

$$r = \sum_{i,j} \frac{ij\left(e_{i,j} - p_i p_j\right)}{\sigma^2}$$

where $e_{i,j}$ is the probability of finding a node with degrees i and j at two ends of a randomly chosen link and p_i is the probability to have a degree i node at the end of a randomly chosen link and

$$\sigma = \sum_i i^2 p_i - \left[\sum_i i p_i\right]^2.$$

In fact, r is the usual Pearson regression coefficient and its variation is between -1 and 1. If r is around 0 then the network is neutral that is $e_{i,j} = p_i p_j$ and this means that the average degree of the neighbors is independent of the degree of the node. If r is greater than 0, then the network is assortative and high degree nodes are preferentially linked to high degree nodes and low degree nodes are preferentially linked to low degree nodes. If r is lower than 0 then the network is disassortative and high degree nodes are preferentially linked to low degree nodes and low degree nodes are preferentially linked to high degree nodes. In biology, gene regulation networks are disassortative and protein-protein interaction network is also disassortative [7]. Here we focus on assortativity results of adjacent amino acid networks. We compute the assortativity measures for the four networks defined previously for each protein of our data set. For example, for the cholera toxin (PDB 1EEI) we have the values of r for PCN Adjacent Amino Acid Network: 0.2455, for LRN Long Range Network: 0.1259, for IHSN Induced Hot Spot Network: 0.2426, for HSN Hot Spot Network: -0.2667 (*see* Fig. 23). We remark that the first three networks for the cholera toxin are assortative and the Hot Spot Network is disassortative. And the computation of the assortativity values on the whole data set gives the following results (*see* Fig. 24).

To summarize the results for the 746 proteins of the data set (because for assortativity it is not important to discard the copy of proteins in the PDB), we sort the result in three classes: the neutral class with assortativity measure beginning by 0.00 or -0.00 (for example a network with assortativity measure 0.0012 or -0.0099 are considered in the neutral class), the assortative class with values above 0.001, and the disassortative class with values below -0.001.

The results are summarized in Table 1 and we have for the Adjacent Amino Acid Network 745 assortative networks and 1 disassortative network confirming the fact that the adjacency amino acid networks are assortative (*see* [5]). For the Long Range Network, we have 615 assortative networks, 31 neutral networks, and 99 disassortative networks. This means that the network that controls the 3D structure is most often assortative. Remark that in the 31 neutral networks we have 5 proteins with an empty graph for LRN (for the following PDB numbers: 3nve, 2omp, 3nhc, 2omq, 2ona) because these proteins are constructed by juxtaposition of short peptides and thus no long-range interactions are available. Remember also that 99 proteins, that is only 13% of the Data Set,

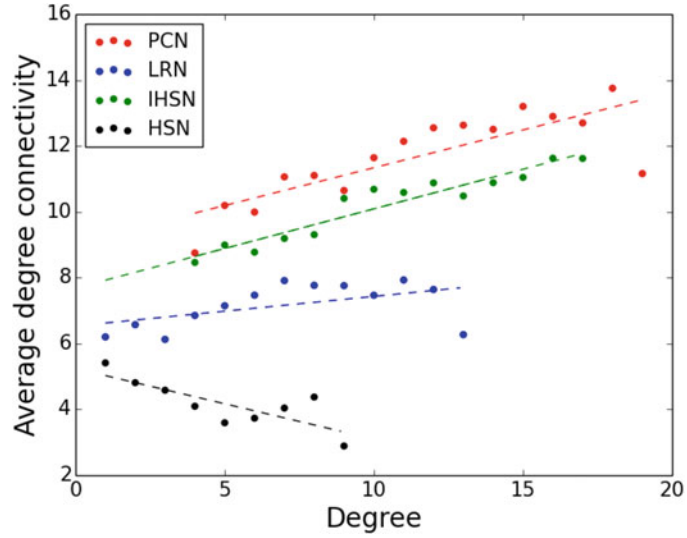

Fig. 23 Assortativity values of 1EEI

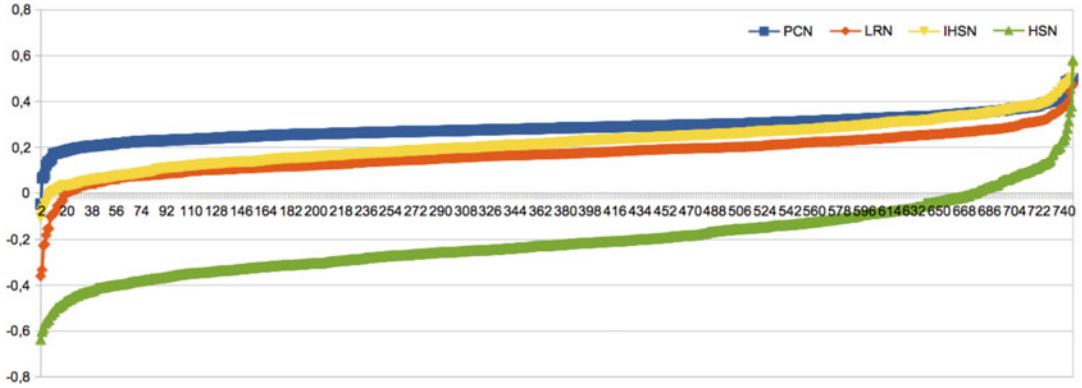

Fig. 24 Assortativity values of the Data Set (in abscissa we have the proteins in the data set from 1 to 746 and in ordinate the assortativity values with range from −1 to 1)

Table 1
Assortativity results for the 4 networks

	Assortative	Neutral	Disassortative
PCN	745	0	1
LRN	615	31	99
IHSN	736	4	6
HSN	68	9	669

have a disassortative Long Range Network and this is in accordance with the results on the article of Bagler and Sinha [5] about the global assortativity of LRN. For the Induced Hot Spot Network, we have 736 assortative networks, 4 neutral networks, and 6 disassortative networks. This could mean that the IHSN mimics the whole protein network property in order to mask the position of HSN. In contrast, for HSN that is the network of interfaces between two chains of the protein, we find 68 assortative networks, 9 neutral networks, and 669 disassortative networks.

The first direct interpretation comes simply by the fact that HSN is by construction a bipartite graph. Indeed, for a complete bipartite graph $K_{m,n}$ we have $r = -1$ if $m \neq n$ and $r = 1$ if $m = n$. Thus, a complete bipartite graph is either fully assortative or fully disassortative. Nevertheless, HSN is bipartite but far from being complete. This is why we think that the disassortative property of the interface between two chains could be the signature of a disassortative barrier. In fact, we have seen already that in the interface a high degree node is protected from damages by placing a low degree node near, as perturbation propagates to the neighbors of the high degree nodes instead of impacting the low degree node connections [ASVSL].

Now we would like to study the independence of the topology of network two by two by using information of the assortativity and proving that there are only few correlations between the assortativity measures of each network. Indeed, if we plot for each protein the assortativity values of PCN versus the assortativity values of HSN, all values are almost on a circle. This is the signature of noncorrelated variables and in particular the R^2 is near zero (the value is around 0.0076) (*see* Fig. 25). In conclusion the assortativity property of the whole protein is independent of the assortativity property of the interface.

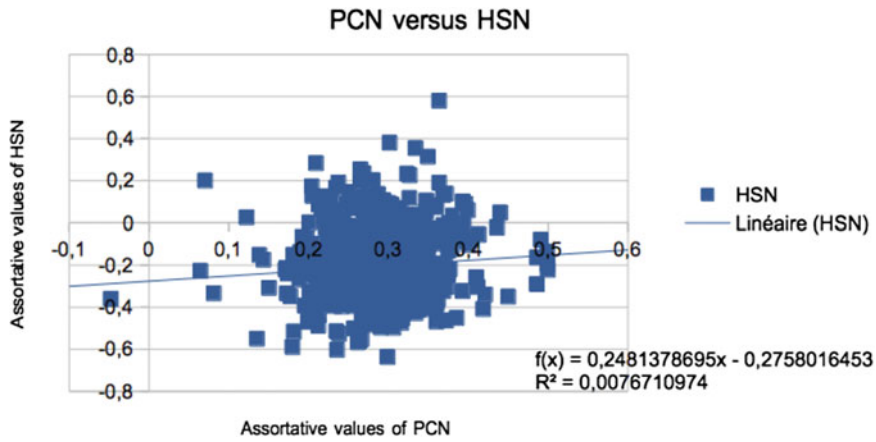

Fig. 25 Assortativity of PCN versus HSN

In the same spirit, if we plot the assortativity values of PCN versus LRN we find a signature of very slightly correlated variables with $R^2 = 0.02733$ and thus in the whole protein network the assortativity property is independent of the Long Range Network. If we plot the assortativity values of PCN versus IHSN we find $R^2 = 0.0405$; thus, the protein assortativity property is independent of the Induced Hot Spot Network. One explanation of these nonintuitive topological results comes from the fact that the protein structure masks the position of crucial networks. In particular the Long Range Network is important in order to maintain the 3D structure and to fold the protein. And looking at the degree correlation of amino acids is not sufficient to guess which one is in the LRN. We find the same result between LRN and IHSN ($R^2 = 0.0058$) and between LRN and HSN ($R^2 = 0.0406$) and thus few correlations between the 3D network (LRN) and the interface networks (HSN or IHSN). Thus PCN vs HSN, PCN vs LRN, and LRN vs IHSN are topologically independent, and this could be a mechanism to prevent propagation of modifications from one network to another. It is also the reflection of a non-hierarchical structure of the whole proteins and a relative independence of each network.

In the opposite if we plot for each protein the assortativity values of IHSN versus the assortativity values of HSN, the values are elongated around a segment of line, and this is the signature of correlated variables. Indeed, the R^2 is around 0.18 and implies a slight correlation between variables that contrast with the noncorrelation property between HSN and PCN (*see* Fig. 26). Of course, a tight control of the interface is embedded in the HSN and that means that many links of PCN are not involved in this control. This could be also a defense property because the evolution masks the essential links for the control of the structure by random mutations.

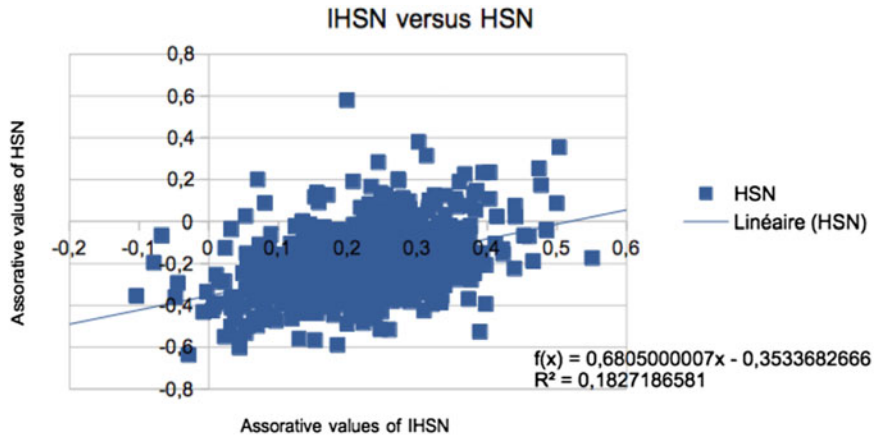

Fig. 26 Assortativity of IHSN versus HSN

In conclusion the structural property and the assortativity property of the interfaces influence the assortativity property of the induced hot spot network.

Acknowledgement

The authors would like to thank the CNRS and the MITI for the support within the project 80|Prime.

References

1. Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41(2):133–180
2. Lesieur C (2014) The assembly of protein oligomers—old stories and new perspectives with graph theory. In: Lesieur C (ed) Oligomerization of chemical and biological compounds. InTech
3. Lesieur C, Vuillon L (2014) From tilings to fibers: bio-mathematical aspects of fold plasticity. In: Lesieur C (ed) Oligomerization of chemical and biological compounds. InTech. ISBN: 978-953-51-1617-2 https://doi.org/10.5772/58577
4. Barabási A-L, Pósfai M (2016) Network science. Cambridge University Press, Cambridge
5. Bagler G, Sinha S (2007) Assortative mixing in protein contact networks and protein folding kinetics. Bioinformatics 23(14):1760–1767
6. Karsai M, Kivelä M, Pan RK, Kaski K, Kertész J, Barabási A-L, Saramäki J (2011) Small but slow world: how network topology and burstiness slow down spreading. Phys Rev E Stat Nonlinear Soft Matter Phys 83:025102
7. Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89:208701
8. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67:026126
9. Di Paola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48
10. Vuillon L, Lesieur C (2015) From local to global changes in proteins: a network view. Curr Opin Struct Biol 31:1–8
11. Viloria JS, Allega MF, Lambrughi M, Papaleo E (2017) An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass. Sci Rep 7(1):2838
12. Feverati G, Achoch M, Vuillon L, Lesieur C (2014) Intermolecular β-strand networks avoid hub residues and favor low interconnectedness: a potential protection mechanism against chain dissociation upon mutation. PLoS One 9(4):e94745
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235
14. Reinhard D (2000) Graph theory, graduate texts in mathematics, vol 173. Springer-Verlag, Berlin
15. Achoch M, Dorantes-Gilardi R, Wymant C, Feverati G, Salamatian K, Vuillon L, Lesieur C (2016) Protein structural robustness to mutations: an in silico investigation. Phys Chem Chem Phys
16. Lesieur C, Cliff MJ, Carter R, James FL, Clarke AR, Hirst TR (2002) A kinetic model of intermediate formation during assembly of cholera toxin b-subunit pentamers. J Biol Chem 277(19):16697–16704
17. Shoemaker BA, Portman JJ, Wolynes PG (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proc Natl Acad Sci 97(16):8868–8873
18. Zrimi J, Ng Ling A, Arifin EGR, Feverati G, Lesieur C (2010) Cholera toxin b subunits assemble into pentamers-proposition of a fly-casting mechanism. PLoS One 5(12):e15347
19. Achoch M, Feverati G, Vuillon L, Salamatian K, Lesieur C (2014) Protein subunit association: NOT a social network. In: TABIS. Institute of Physics, Belgrade, Belgrade

Chapter 9

Community Network Analysis of Allosteric Proteins

Ivan Rivalta and Victor S. Batista

Abstract

Community network analysis (CNA) of correlated protein motions allows modeling of signals propagation in allosteric proteic systems. From standard classical molecular dynamics (MD) simulations, protein motions can be analysed by means of mutual information between pairs of amino acid residues, providing dynamical weighted networks that contains fundamental information of the communication among amino acids. The CNA method has been successfully applied to a variety of allosteric systems including an enzyme, a nuclear receptor and a bacterial adaptive immune system, providing characterization of the allosteric pathways. This method is complementary to network analyses based on different metrics and it is particularly powerful for studying large proteic systems, as it provides a coarse-grained view of the communication flows within large and complex networks.

Key words Protein correlated motions, Allosteric enzyme, Protein graph, Community network analysis, Allosteric pathways

1 Introduction

Understanding protein allostery is hindered by the complex and elusive nature of the allosteric mechanisms in proteic systems [1]. In fact, despite pioneering allostery models date more than half a century back [2] and the distinct views developed so far have been recently unified [3, 4], many aspects of allosteric phenomena remain poorly understood [5]. In particular, considering the ubiquitous role of allostery in biological systems, it is highly desirable to fully exploit its potential for rational drug design and protein engineering [6–8]. For instance, allosteric enzymes could be manipulated to inhibit/enhance their enzymatic activity by point mutagenesis or by binding of exogenous allosteric ligands (effectors) and/or by disclosing unknown allosteric sites [9–12]. To his aim, computer simulations represent an exceptional tool that can provide the necessary system-specific information for rational design, avoiding extended (and costly) trial-and-error investigations.

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_9,

In allosteric enzymes, the binding of a (effector) ligand at the allosteric site, i.e., a site distant at least 1 nm from the functional active site, regulates the enzymatic function, thus involving communication of a chemical signal from the allosteric to the active site. This allosteric signal is expected to involve physico-chemical interactions between, generally conserved amino acid residues important for allostery [13–15], encompassing secondary structure elements that define the "allosteric pathways" of the enzymatic system [16]. Experimental characterization of allosteric pathways is extremely challenging, and computational chemistry techniques, such classical molecular dynamics (MD) simulations, could provide unique information on protein dynamics at atomistic resolution that significantly contributes to the elucidation of allosteric mechanisms [17–19]. Standard MD simulations, in fact, are routinely used to monitor protein motions up to the μs timescale, yielding MD trajectories that comprise the protein dynamics underpinning the allosteric mechanisms [20]. On the other hand, the atomistic resolution of MD simulations and the large size of proteins (in the atomistic scale) make the recognition of allosteric pathways (within the large network of physico-chemical interactions typical of proteic systems) a real challenge for standard analysis of MD trajectories. Graph theory encompasses the appropriate tools for modeling complex dynamical networks of chemical interactions resulting from atomistic MD simulations. Various graph theory approaches could be exploited to represent a protein and to decipher its allosteric mechanism, including the contact and the elastic network models reported in details in this book series [21, 22]. Here, we describe a method that combines the information on the correlated protein motions resulting from atomistic MD simulations with a network analysis based on graph partitioning into mutually exclusive groups, named communities. The community network analysis (CNA) has been applied to several biological systems, including allosteric enzymes, nuclear receptors, and bacterial adaptive immune systems [23–30], providing elucidation of allosteric pathways and supporting rational discovery of synthetic allosteric modulators [31, 32]. As exemplifying case, we report here the CNA analysis of the imidazole glycerol phosphate synthase (IGPS) enzyme from the thermophile *Thermotoga maritima*, an allosteric enzyme that represents a potential target for allosteric drugs development [23, 33–35].

2 Materials

2.1 Initial Conditions

The proposed CNA method is based on the measure of protein motion correlations and thus relies on the protein dynamics resulting from classical MD simulations. In order to set up MD simulations, initial conditions need to be defined. First, basic structural

information on the target protein is required, with initial structure usually extracted from the reference database for protein structures, i.e., the Protein Data Bank [36]. The protein dynamics is preferentially simulated in realistic environmental conditions, i.e., in presence of the explicit solvent (generally water) and eventually, if the protein is particularly small, accounting for appropriate salt concentration. The whole system is generally comprised within a cubic box, which is then periodically replicated using periodic boundary conditions.

2.2 Molecular Dynamics Simulations

Several freely distributed coded are available for performing classical MD simulations, including for instance AMBER [37], GROMACS [38], and NAMD [39], which also an extended set of useful tutorials for beginners [40–42]. In our applications of the CNA method, we have employed the NAMD software, choosing the AMBER (all-atom) force fields [37, 43] to describe the interactions between atoms. Anyway, the choice of a different all-atom force field does not exclude the possibility of performing the CNA analysis, as explained in the following section. A time window for the CNA analysis has to be selected, in order to perform MD simulations of opportune lengths (*see* **Note 1**). Standard MD simulations generally provide trajectories for not more than few micorseconds and thus enough statistics could be collected for time windows of few hundreds of nanoseconds (*see* **Note 2**), a timescale in which several protein motions (such as unhindered surface side chain loop motions, collective motions, partial folding/unfolding, helix-coil transitions, etc.), eventually related to allostery, already occur.

2.3 Protein Motion Correlations

MD simulations produce trajectories that are subsequently analyzed to obtain protein motion correlations. Pair-correlations between motions of two amino acid residues can be obtained from MD trajectories by calculating the normalized covariance matrix, $\mathbf{r}[x_i,x_j]$, of atomic fluctuations (x_i and x_j for atoms i and j, respectively), generally using alpha carbon atoms (C_α) to represent each residue, defined analogously to the Pearson correlation coefficient as

$$r[x_i, x_j] = \langle x_i \cdot x_j \rangle / \left(\langle x_i^2 \rangle \langle x_j^2 \rangle \right)^{1/2} \quad (1)$$

with values close to 0 for uncorrelated motions and $0 < r[x_i, x_j] \leq 1$ for correlated motions and $-1 < r[x_i, x_j] \leq 0$ for anti-correlated ones. This measure of pair-correlation is however limited to linear correlations and is strongly dependent on the relative orientation of atomic fluctuations, i.e., correlated motions that are orthogonal result as uncorrelated due to zeroing of the dot product in the definition of $\mathbf{r}[x_i,x_j]$. Since the protein graph in the CNA method is based on the measure of such pair-correlations, using a more accurate estimate of protein motions is highly desired [23]. We have

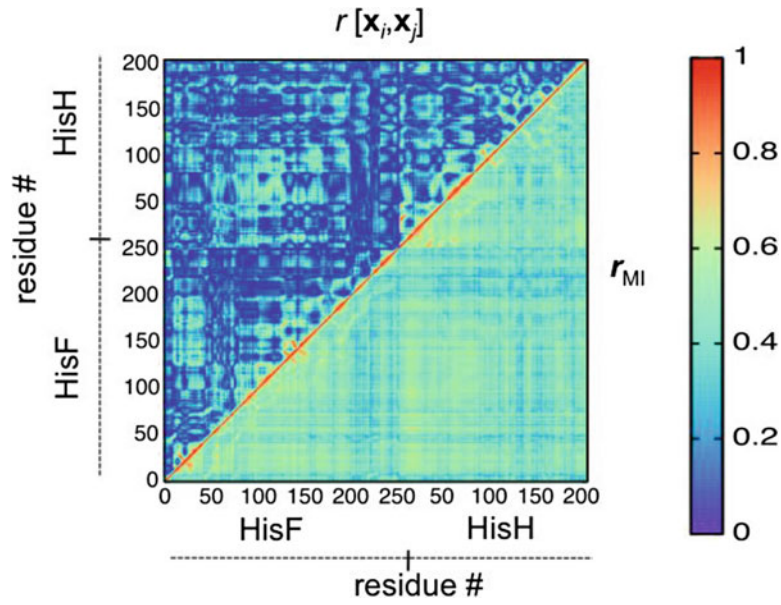

Fig. 1 Comparison of the generalized (r_{MI}, lower right triangle matrix) and the Pearson ($r[x_i,x_j]$, upper left triangle matrix) correlation coefficients, with Pearson coefficient absolute values reported. The data refer to 50 ns representative MD simulation of the apo-IGPS enzyme, as reported in and reprinted with permission from ref. [23]

opted for the generalized correlation coefficient $\mathbf{r}_{MI}$ proposed by Lange et al. [44] and defined as

$$r_{MI}[x_i, x_j] = \{1 - \exp(-2I[x_i, x_j]/d)\} \quad (2)$$

where d is the dimensionality of the variables x_i and x_j, and $\mathbf{I}[x_i,x_j]$ is the mutual information (MI) between the two variables (*see* **Note 3**), which is associated to the expected value of the information content of a discrete random variable, i.e., to its Shannon entropy [43]. The information content is calculated using the C_α atomic positions fluctuations derived from the MD simulations, upon removal of translational and rotational motions by alignment of C_α atoms along an MD trajectory, by means of the k-nearest neighbor distances algorithm [45] as implemented in the "g_correlation" code [44] of the GROMACS software [38], which can be also used to compute the Pearson correlation coefficient. The $r_{MI}[\mathbf{x}_i,\mathbf{x}_j]$ coefficient will be zero for fully uncorrelated motions and it will assume values up to 1 for fully correlated motions, so it can be compared with the absolute value of the $r[x_i,x_j]$ Pearson correlation coefficient. As shown in Fig. 1, the $r_{MI}[\mathbf{x}_i,\mathbf{x}_j]$ coefficient captures more correlations with respect to the $r[x_i,x_j]$ Pearson coefficient as it accounts for nonlinear correlations and it does not vanish for correlated motions with orthogonal orientations.

In order to improve the estimates of the protein motion correlations a statistical analysis over multiple MD simulations should be performed. The correlation coefficients can be computed for each independent MD trajectory (or as a single computation of concatenated independent trajectories) and then averaged out. As mentioned above, the correlation coefficients (including $\mathbf{r}[x_i,x_j]$) can be computed using the GROMACS software, which requires the topology file (containing all the necessary information to compute the forces and velocities for the MD) to be in the GROMACS format. This topology file format can be obtained from other topology formats (e.g., from AMBER topology file) using free software, such as ParmEd [46].

3 Methods

3.1 Dynamical Weighted Network

The pair-correlations coefficients defined above measure how correlated are the motion of two residues recorded during a MD simulation. When employing the mutual information measure, i.e., the generalized $r_{\mathbf{MI}}[\mathbf{x}_i,\mathbf{x}_j]$ coefficients, the pair-correlations are directly related to the exchange of information between two residues and can then be used to weight a graph that models the information exchange within a protein according to its dynamics, i.e., to build a dynamical weighted network. First, a protein graph needs to be defined. Since the correlation coefficients are computed using the C_α atomic positions fluctuations, each node of the graph is associated to each C_α of the protein representing each amino acid residue in the primary sequence, *see* Fig. 2a. To complete the graph, one needs to build an adjacency matrix, A, i.e., a matrix that defines the existence of connections (edges) between the nodes, thus with entries A_{ij} equal to zero if the nodes i and j are not linked by an edge, or is different from zero if the nodes are connected. For a graph weighted by w_{ij} coefficients, the adjacency matrix reads

$$A_{ij} = \begin{cases} w_{ij} & \text{if edge for nodes } i \text{ and } j \text{ exists} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The zeros of the adjacency matrix should thus represent the pairs of residues that do not exchange information (through correlated motions) and thus are not linked in the dynamical weighted network. Sethi et al. [27] proposed to use a distance cutoff to discriminate between residues that are linked or not in dynamical weighted networks, i.e., suggested to include only edges that represent pairs of residues that are found at chemical distance (at least one distance between heavy atoms is below the cutoff, varying in the range of 3.5–5.5 Å) during the MD simulations, thus excluding long-range correlations from the network of communications, *see*

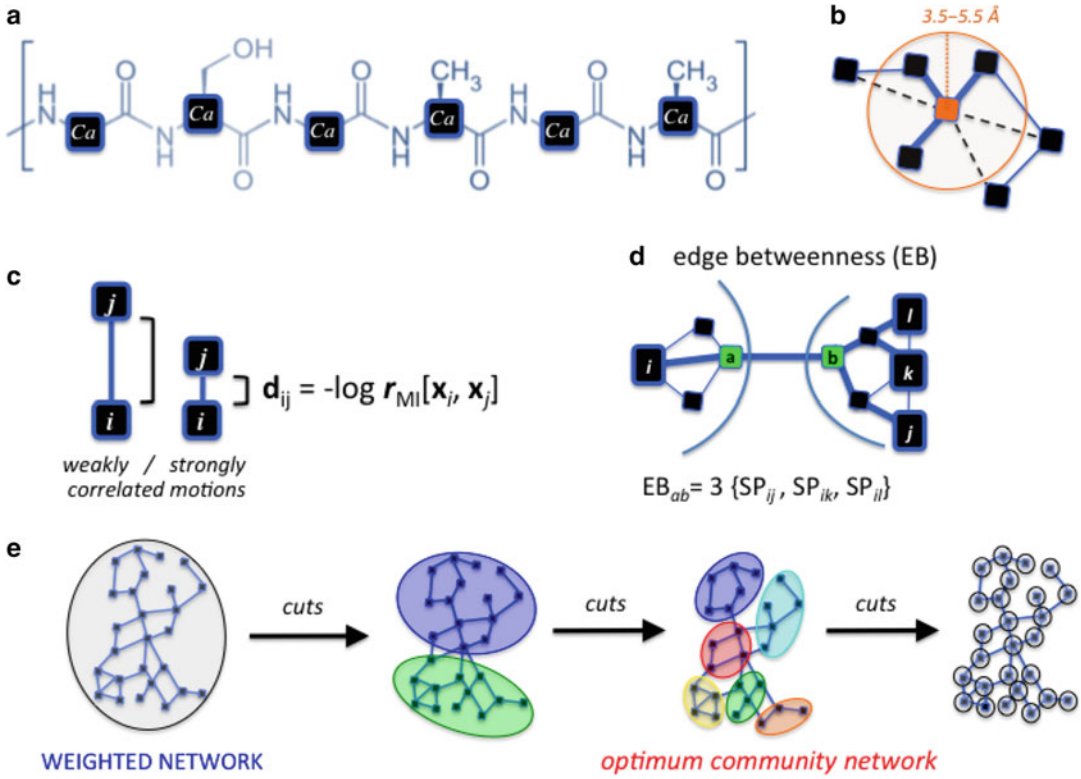

Fig. 2 Schematic representation of the graph theory methodology employed. (**a**) The nodes in the graph are associated to the Cα of the amino acid residue in the protein primary sequence. (**b**) The edge between two nodes exists if a contact distance cutoff (varying between 3.5–5.5 Å) is satisfied. (**c**) For all pairs of residues *i* and *j* that are in contact, the generalized correlation coefficient $r_{MI}[x_i,x_j]$ is converted into communication "distance" d_{ij} that is used to weight each edge in the network. (**d**) The edge betweenness (EB) defined as the number of shortest pathways (SPs) that cross a given edge is used as partitioning criterion for the weighted network. In the example, the SPs for the three pairs of residues *i-j, i-k, i-l* all cross the edge between residues *a* and *b*, yielding an EB_{ab} equal to three. (**e**) The Girvan-Newman algorithm removes (or cuts) edges with the highest EBs, partitioning progressively the weighted network into communities

Fig. 2b. The distance cutoff thus defines the presence of contacts between residues and it is complemented by a statistical "percentage" cutoff, which is related to how long two residues are in contact along the MD trajectories: for a fixed cutoff distance, the contacts are evaluated along the trajectories and their existence is determined as function of the percentage of frames (from 65% to 85%) in which these contacts are effectively present. The CNA results are checked against these two cutoffs (distance and percentage) in order to provide robust and converged outcome.

Finally, the edges weights w_{ij} that fill the nonzero entries of the adjacency matrix, as defined in Eq. 3, are obtained from the generalized correlation coefficient $r_{MI}[\mathbf{x}_i,\mathbf{x}_j]$ (*see* Eq. 2) by converting them into communication "distances" $\mathbf{d}_{ij}$ by taking the $-log$ $[r_{MI}(\mathbf{x}_i,\mathbf{x}_j)]$ for all pairs of residues *i* and *j* that are in contact, *see*

Fig. 2c. In this way, the dynamical weighted network represents a communication network where the residues physically in contact in the protein structure (for most of the time along the MD simulation) exchanging more information (i.e., have larger pair-correlation coefficients) and are found closer in distance within the graph with respect to those that, despite being in physical contact, have lower correlation coefficients and thus communicate less among each other.

3.2 Community Network Analysis

The dynamical weighted network, as defined in the above section, it contains information paths and critical nodes that are important for communication within the proteic system under study. Such network already contains relevant information for allostery and it can be analyzed by determining the shortest pathways (SPs) between all pairs of (not directly linked) nodes/residues. The SPs within the protein communication network can be calculated using the Floyd-Warshall algorithm [47], representing the best communication pathways among pairs of residues. The communication pathways between physically distant amino acid residues (such as those in the active and allosteric sites) are of extreme relevance for the elucidation of allosteric mechanisms. Still, it is not straightforward to identify the allosterically relevant residues among which to calculate the SPs, since both the allosteric and active sites are generally characterized by sized domains, possibly involving a relatively large number of amino acid residues. Thus, it becomes quite useful, especially for large proteic systems, to use the information embodied in the SPs for partitioning the whole dynamical network, providing a coarse-grained view of the communication flows within the protein network that facilitates the understanding of the allosteric mechanisms.

The Girvan-Newman algorithm [48] can be used to split the dynamical weighted network into "communities," i.e., local substructures involving groups of nodes within which the connections are dense but between which they are sparser. This algorithm exploits as partitioning criterion for the weighted network the *edge betweenness*, EB, defined as the number of SPs that cross a given edge, *see* Fig. 2d [49], measuring of how much this edge is responsible for connecting the other pairs of nodes in the network. Therefore, the edges with the highest between-nesses are those carrying on the highest amount of information exchange within the protein network. An iterative procedure that starts from the whole protein network as a single big community and removes/cuts the edges with highest between-nesses would, then, isolate the nodes progressively creating smaller and smaller communities, up to generation of a number of communities that corresponds to the number of (isolated) nodes, *see* Fig. 2e. The iterative procedure could, however, be interrupted at given step and produce a specific partition of the dynamical weighted network, namely a community

structure. If one considers the community structure as a network of communities, then each node in such graph is a community and the edge connecting two different communities is the sum of all (protein network) edges (with highest EBs) that have been removed during the Girvan-Newman procedure. Thus, the community network represents a coarse-grained, where linked communities are connected by edges that are weighted by the inter-communities EB (IEB), i.e., the sum of the EBs associated to the pairs of residues connecting the two communities. The IEBs indicate the strength of the communication flow between pairs of communities and thus represent a simplified way to represent the communication associated to correlated protein motions within a protein.

The optimal partition of the protein network into communities is obviously not that where the number of communities equals the number of nodes and, thus, the quality of the network partition has to be evaluated in order to determine the best distribution of nodes in the communities, i.e., the optimum community structure. The modularity of a given network division, i.e., the difference in probability of intra- and inter-community connections [49], is a very useful quantity to measure the quality (or strength) of a community structure. The modularity, Q, is defined as

$$Q = \sum_i \left(e_{ii} - a_i^2\right) \tag{4}$$

where e_{ij} is the fractions of edges that link nodes in community i to nodes in community j, and $a_i = \Sigma_j e_{ij}$ is the fraction of edges that connect to nodes in community i. The modularity values range from 0 to 1 (*see* Fig. 3), the higher the values, the higher the quality of the community structure, with typical values for community networks originated from 3D structures being >0.4 [49]. By selecting the community structure with the maximal modularity among those generated by the iterative Girvan-Newman algorithm, the optimum community structure is chosen in such a way that each community contains nodes that are highly intra-connected while different communities are poorly inter-connected but through few critical edges, representing the pairs of nodes crucial for communications among communities.

3.3 Assessment of Cutoff Parameters Choice

As mentioned in the previous sections, the adjacency matrix defining the dynamical weighted network is bound to the contact criterion associated to the distance and percentage cutoff parameters. As a consequence, the generation of the optimum community structure for the dynamical weighted network, described in the above section, is carried out for a given set of these two cutoff parameters, i.e., one contact distance and one percentage of MD frames. In order to assess the reliability of an optimum community structure representing the correlated protein motions in a given MD simulation, thus, is necessary to analyze the effect of the cutoff

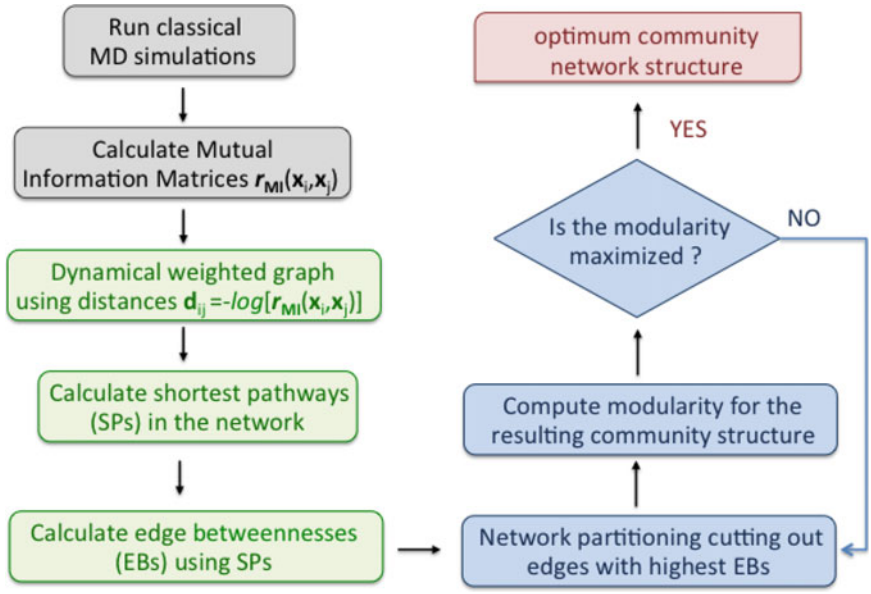

Fig. 3 Schematic workflow of the community network analysis. MD simulations and estimate of correlated motions (gray boxes) are followed by dynamical weighted graph construction and edge betweenness computations (green boxes), allowing application of the iterative Girvan-Newman algorithm with evaluation of the community network modularity (blue boxes) in order to define the optimum community network (red box)

parameters. To this aim, the community repartition difference (CRD) [27] between two different community structures (e.g., c_1 and c_2) can be computed as

$$\mathrm{CRD}_{(c_1,c_2)} = 1 - \frac{\sum_{n_i,n_j} z(n_i,n_j,c_1)z(n_i,n_j,\ c_2)}{\sum_{n_i,n_j} z(n_i,n_j,c_1)} \qquad (5)$$

where $z(n_i,\ n_j,\ c_i)$ is 1 if nodes n_i and n_j belong to the same community in a given community structure (c_i) and 0 otherwise. Thus, CRD represents a normalized count of node pairs that are grouped together in both community structures (c_1 and c_2), going to 0 if the two community structures are identical or to 1 if they are totally different. The CRD is then a good estimate of the similarities between two optimum community structures obtained with different values of the distance and/or percentage cutoff parameters (*see* Fig. 5).

4 Applications

The CNA method (*see* **Note 4**) described in the above section has been applied to various proteic systems featuring allosteric regulation, including allosteric enzyme [23], nuclear receptor [24], and

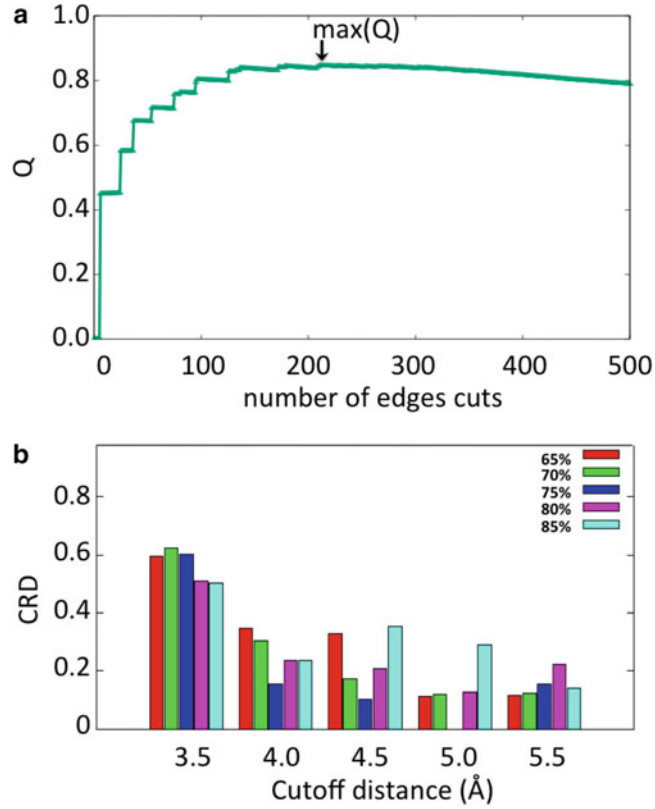

Fig. 4 (**a**) Typical evolution of modularity (*Q*) values during the iterative Girvan-Newman procedure, with maximum value, max(*Q*), determining the optimum community structure. (**b**) Typical plot of community repartition difference (CRD) values obtained comparing community structures with different distance (3.5–5.5 Å) and percentage (65–85%) cutoff parameters, with reference set to community structure at 5.0 Å and 75%. Reprinted with permission from ref. [23]

bacterial adaptive immune system [25]. In particular, the imidazole glycerol phosphate synthase (IGPS) enzyme from the thermophile *Thermotoga maritima*, *see* Fig. 4, has been the target system for assessing the community network analysis based on mutual information of correlated motions. This prototypical enzyme is, indeed, a V-type allosteric enzyme (i.e., allosteric regulation affect enzymatic kinetics but not substrate binding affinity) featuring a tight connection between protein dynamics and allosteric regulation as previously demonstrated by nuclear magnetic resonance (NMR) relaxation dispersion experiments and calorimetry measurements [50, 51]. The IGPS enzyme is involved in essential biochemical pathways (purines and histidine synthesis) of pathogens but it is absent in mammals, thus representing a potential target for antibiotic and antifungal development [33–35]. The HisH glutamine amidotransferase, which catalyzes the hydrolysis of the substrate (glutamine), and the HisF cyclase, where the effector PRFAR, i.e.,

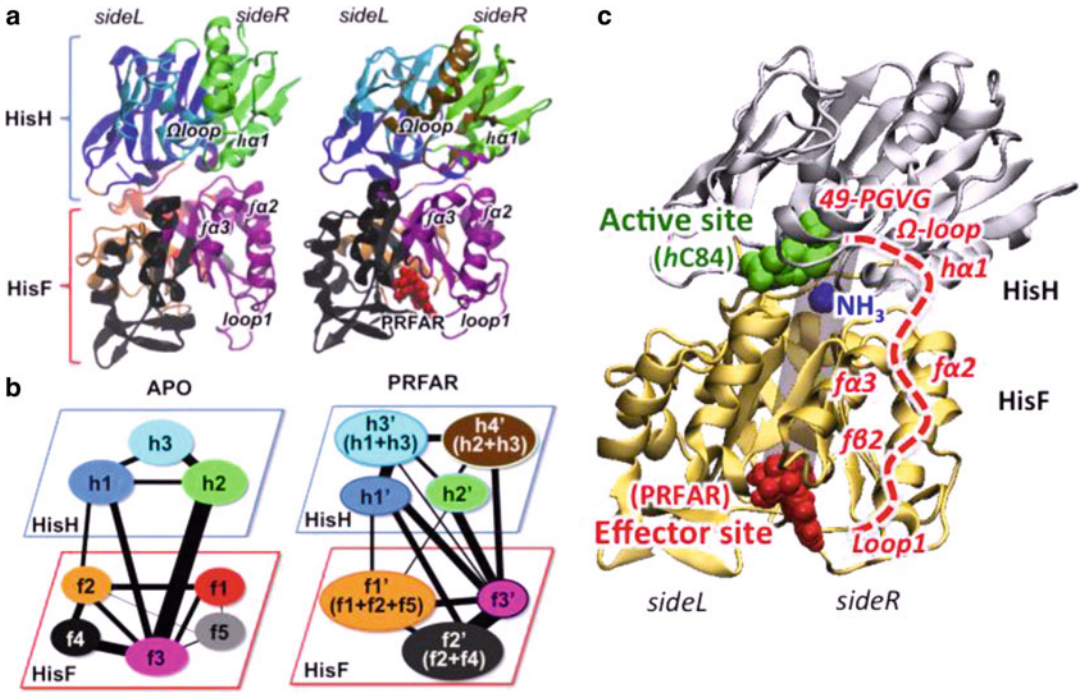

Fig. 5 (**a**) 3D representation of the community network structures for apo and effector-bound IGPS enzymes, showing how communities are related to groups of secondary structure elements within the HisH and HisF proteins. (**b**) Schematic representation of the community network structures, showing how the links between communities, whose widths are proportional to the IEB values, can readily describe the changes in communication flow induced by the PRFAR effector. (**c**) IGPS allosteric pathways connecting effector and active sites, as suggested by the CNA method. Reprinted with permission from ref. [23]

N′-[(5′-phosphoribulosyl)formimino]-5-aminoimidazole-4-carboxamide-ribonucleotide binds, are the two tightly associated proteins constituting the IGPS allosteric enzyme, as shown in Fig. 4. PRFAR effector binding accelerates glutamine hydrolysis by ca. 5000-fold, with respect to the apo-IGPS enzyme [52].

As shown in Fig. 4, the outcome of the CNA method is a coarse-grained picture of the division of the allosteric proteic system, which has a straightforward structural interpretation but it also contains information on the communication flow within the complex network of amino acid residues. In the case of IGPS, the CNA method showed to be quite sensitive to the changes in communication network induced by the allosteric regulator, allowing detection of suggested secondary structure elements and key residues involved in the allosteric signal propagation [23]. In particular, the IGPS allostery involves a specific sequence of interactions at one side of the IGPS complex (*sideR*, *see* Fig. 4) that alters the protein dynamics, with i) hydrophobic interactions in the *fβ*2 strand and hydrogen bonds in the flexible *loop1* at the HisF allosteric site; ionic interactions between *fα*2, *fα*3, and *hα*1 helices at the

HisF/HisH interface; and hydrogen-bonding between the *Ω-loop* and the conserved 49-PGVG sequence (i.e., the oxyanion strand) adjacent to the HisH active site. The CNA outcome showed to be of particular help for further manipulation of the allosteric regulation in IGPS, allowing rational design of allosteric inhibitors that could interfere with the suggested allosteric pathways [32] and promoting experimental mutagenesis studies [53] that granted knockout of IGPS allosteric signal propagation. Moreover, the CNA proved to be a transferable approach that we have successfully employed to other allosteric systems [24, 25], in conjunction with other graph approaches involving the eigenvector centrality metric to account for long-range correlated motions [26] and to dynamical perturbation networks based on inter-residues physical contacts along MD simulations [54].

5 Notes

1. The CNA method is based on the outcome of standard MD simulations, and the outcome is strictly related to the simulation time and the window the user decides to analyze. Generally, time windows around 50–150 ns are reasonable, while this depends on the system size: larger protein usually requires longer MD simulation time to sample motions possibly relevant to allostery.
2. To get the best possible statistical analysis of correlated motions, once a time window is selected (e.g., 100 ns), several "running" time windows of that length should be extracted from the MD trajectories. As mentioned above, running more than one independent simulation is strongly suggested. The correlation coefficients computed for each of these time windows (and for multiple trajectories) can then be averaged out to provide a single correlation matrix for each system. A preliminary investigation of the CNA outcome as function of the time window length chosen (e.g., comparing results with 50, 100, 150 ns time windows) is also suggested.
3. For very large proteic systems, the computation of the correlation coefficients might be quite demanding. To overcome this limitation, a suggested solution is to use linearized MI coefficients, as suggested by Lange et al. and implemented in the "g_correlation" code [44] of the GROMACS software [38].
4. The code for the CNA method is available under request to the authors of this chapter. The code provides as output pictures like those in Fig. 4b and text files that can be readily used to

produce 3D representation of the community network structures by employing a visualization software, such as the Visual Molecular Dynamics tool [55].

References

1. Fenton AW (2008) Allostery: an illustrated definition for the 'second secret of life'. Trends Biochem Sci 33(9):420–425
2. Changeux JP (2013) 50 years of allosteric interactions: the twists and turns of the models. Nat Rev Mol Cell Biol 14(12):819–829
3. Hilser VJ, Wrabl JO, Motlagh HN (2012) Structural and energetic basis of allostery. Annu Rev Biophys 41:585–609
4. Tsai CJ, Nussinov R (2014) A unified view of "how allostery works". PLoS Comput Biol 10 (2):e1003394
5. Wodak SJ, Paci E, Dokholyan NV, Berezovsky IN, Horovitz A, Li J, Hilser VJ, Bahar I, Karanicolas J, Stock G, Hamm P, Stote RH, Eberhardt J, Chebaro Y, Dejaegere A, Cecchini M, Changeux J-P, Bolhuis PG, Vreede J, Faccioli P, Orioli S, Ravasio R, Yan L, Brito C, Wyart M, Gkeka P, Rivalta I, Palermo G, McCammon JA, Panecka-Hofman J, Wade RC, Di Pizio A, Niv MY, Nussinov R, Tsai C-J, Jang H, Padhorny D, Kozakov D, McLeish T (2019) Allostery in its many disguises: from theory to applications. Structure 27(4):566–578
6. Wootten D, Christopoulos A, Sexton PM (2013) Emerging paradigms in gpcr allostery: implications for drug discovery. Nat Rev Drug Discov 12(8):630–644
7. Taly A, Corringer PJ, Guedin D, Lestage P, Changeux JP (2009) Nicotinic receptors: allosteric transitions and therapeutic targets in the nervous system. Nat Rev Drug Discov 8 (9):733–750
8. Nussinov R, Tsai CJ (2013) Allostery in disease and in drug discovery. Cell 153(2):293–305
9. Christopoulos A (2002) Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. Nat Rev Drug Discov 1 (3):198–210
10. Gohara DW, Di Cera E (2011) Allostery in trypsin-like proteases suggests new therapeutic strategies. Trends Biotechnol 29(11):577–585
11. Makhlynets OV, Raymond EA, Korendovych IV (2015) Design of allosterically regulated protein catalysts. Biochemistry 54 (7):1444–1456
12. Lisi GP, Manley GA, Hendrickson H, Rivalta I, Batista VS, Loria JP (2016) Dissecting dynamic allosteric pathways using chemically related small-molecule activators. Structure 24 (7):1155–1166
13. Suel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 10(1):59–69
14. Amaro RE, Sethi A, Myers RS, Davisson VJ, Luthey-Schulten ZA (2007) A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. Biochemistry 46(8):2156–2173
15. Bruschweiler S, Schanda P, Kloiber K, Brutscher B, Kontaxis G, Konrat R, Tollinger M (2009) Direct observation of the dynamic process underlying allosteric signal transmission. J Am Chem Soc 131(8):3063–3068
16. del Sol A, Tsai CJ, Ma B, Nussinov R (2009) The origin of allosteric functional modulation: multiple pre-existing pathways. Structure 17 (8):1042–1050
17. Feher VA, Durrant JD, Van Wart AT, Amaro RE (2014) Computational approaches to mapping allosteric pathways. Curr Opin Struct Biol 25:98–103
18. Martin NE, Malik S, Calimet N, Changeux JP, Cecchini M (2017) Un-gating and allosteric modulation of a pentameric ligand-gated ion channel captured by molecular dynamics. PLoS Comput Biol 13(10):e1005784
19. Markwick PR, McCammon JA (2011) Studying functional dynamics in bio-molecules using accelerated molecular dynamics. Phys Chem Chem Phys 13(45):20053–20065
20. De Vivo M, Masetti M, Bottegoni G, Cavalli A (2016) Role of molecular dynamics and related methods in drug discovery. J Med Chem 59 (9):4035–4061
21. Di Paola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48
22. Di Paola L, De Ruvo M, Paci P, Santoni D, Giuliani A (2013) Protein contact networks: an emerging paradigm in chemistry. Chem Rev 113(3):1598–1613
23. Rivalta I, Sultan MM, Lee NS, Manley GA, Loria JP, Batista VS (2012) Allosteric pathways in imidazole glycerol phosphate synthase. Proc Natl Acad Sci U S A 109(22):E1428–E1436

24. Ricci CG, Silveira RL, Rivalta I, Batista VS, Skaf MS (2016) Allosteric pathways in the pparγ-rxrα nuclear receptor complex. Sci Rep 6:19940
25. Palermo G, Ricci CG, Fernando A, Basak R, Jinek M, Rivalta I, Batista VS, McCammon JA (2017) Protospacer adjacent motif-induced allostery activates crispr-cas9. J Am Chem Soc 139(45):16028–16031
26. Negre CFA, Morzan UN, Hendrickson HP, Pal R, Lisi GP, Loria JP, Rivalta I, Ho J, Batista VS (2018) Eigenvector centrality for characterization of protein allosteric pathways. Proc Natl Acad Sci U S A 115(52):E12201–E12208
27. Sethi A, Eargle J, Black AA, Luthey-Schulten Z (2009) Dynamical networks in tRNA: protein complexes. Proc Natl Acad Sci U S A 106 (16):6620–6625
28. Gasper PM, Fuglestad B, Komives EA, Markwick PR, McCammon JA (2012) Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. Proc Natl Acad Sci U S A 109(52):21216–21222
29. Blacklock K, Verkhivker GM (2014) Computational modeling of allosteric regulation in the hsp90 chaperones: a statistical ensemble analysis of protein structure networks and allosteric communications. PLoS Comput Biol 10(6): e1003679
30. Stolzenberg S, Michino M, LeVine MV, Weinstein H, Shi L (2016) Computational approaches to detect allosteric pathways in transmembrane molecular machines. Biochim Biophys Acta 1858(7 Pt B):1652–1662
31. Wagner JR, Lee CT, Durrant JD, Malmstrom RD, Feher VA, Amaro RE (2016) Emerging computational methods for the rational discovery of allosteric drugs. Chem Rev 116 (11):6370–6390
32. Rivalta I, Lisi GP, Snoeberger NS, Manley G, Loria JP, Batista VS (2016) Allosteric communication disrupted by a small molecule binding to the imidazole glycerol phosphate synthase protein-protein interface. Biochemistry 55 (47):6484–6494
33. Chaudhuri BN, Lange SC, Myers RS, Chittur SV, Davisson VJ, Smith JL (2001) Crystal structure of imidazole glycerol phosphate synthase: a tunnel through a (beta/alpha) (8) barrel joins two active sites. Structure 9 (10):987–997
34. Breitbach K, Köhler J, Steinmetz I (2008) Induction of protective immunity against burkholderia pseudomallei using attenuated mutants with defects in the intracellular life cycle. Trans R Soc Trop Med Hyg 102: S89–S94
35. Gomez MJ, Neyfakh AA (2006) Genes involved in intrinsic antibiotic resistance of acinetobacter baylyi. Antimicrob Agents Chemother 50(11):3562–3567
36. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242. http://www.rcsb.org/pdb/
37. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The amber biomolecular simulation programs. J Comput Chem 26(16):1668–1688
38. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4 (3):435–447
39. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K (2005) Scalable molecular dynamics with namd. J Comput Chem 26 (16):1781–1802
40. See webpage: https://ambermd.org/tutorials/
41. See webpage: http://www.Ks.Uiuc.Edu/training/tutorials/namd-index.Html
42. See webpage: http://www.Gromacs.Org/documentation/tutorials
43. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25 (9):1157–1174
44. Lange OF, Grubmuller H (2006) Generalized correlation for biomolecular dynamics. Proteins 62(4):1053–1061
45. Kraskov A, Stogbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E Stat Nonlinear Soft Matter Phys 69(6 Pt 2):066138
46. See webpage: https://parmed.github.io/parmed/html/parmed.html
47. Floyd RW (1962) Algorithm-97—shortest path. Commun ACM 5(6):345–345
48. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci U S A 99(12):7821–7826
49. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113
50. Lipchock J, Loria JP (2009) Millisecond dynamics in the allosteric enzyme imidazole

glycerol phosphate synthase (igps) from thermotoga maritima. J Biomol NMR 45 (1–2):73–84

51. Lipchock JM, Loria JP (2010) Nanometer propagation of millisecond motions in v-type allostery. Structure 18(12):1596–1607
52. Myers RS, Jensen JR, Deras IL, Smith JL, Davisson VJ (2003) Substrate-induced changes in the ammonia channel for imidazole glycerol phosphate synthase. Biochemistry 42 (23):7013–7022
53. Lisi GP, East KW, Batista VS, Loria JP (2017) Altering the allosteric pathway in igps suppresses millisecond motions and catalytic activity. Proc Natl Acad Sci U S A 114(17): E3414–E3423
54. Gheeraert A, Pacini L, Batista VS, Vuillon L, Lesieur C, Rivalta I (2019) Exploring allosteric pathways of a v-type enzyme with dynamical perturbation networks. J Phys Chem B 123 (16):3452–3461
55. Humphrey W, Dalke A, Schulten K (1996) Vmd: visual molecular dynamics. J Mol Graph Model 14(1):33–38

Chapter 10

The PyInteraph Workflow for the Study of Interaction Networks From Protein Structural Ensembles

Matteo Lambrughi, Valentina Sora, and Matteo Tiberti

Abstract

PyInteraph is a software package designed for the analysis of structural communication from conformational ensembles, such as those derived from in silico simulations, under the formalism of protein structure networks. We demonstrate its usage for the calculation and analysis of intramolecular interaction networks derived from three different types of interactions, as well as with a more general protocol based on distances between centers of mass. We use the xPyder PyMOL plug-in to visualize such networks on the three-dimensional structure of the protein. We showcase our protocol on a molecular dynamics trajectory of the Cyclophilin A wild-type enzyme, a well-studied protein in which different allosteric mechanisms have been investigated.

Key words Protein structure networks, PSN, Structural communication, Allostery, Salt bridge, Hydrogen bond, Atomic contact, Non-covalent interaction, Graph

1 Introduction

1.1 Long-Range Structural Communication in Proteins

Localized changes in the structure of a protein can influence distant regions, resulting in sometimes dramatic changes far from the site of the perturbation in terms of both structure and dynamics. Such perturbations can be, for instance, the binding of a small molecule as inhibitor or effector, but also a mutation or a post-translational modification [1–3]. The localized perturbation at a site imposes stress on the protein structure and changes in dynamics, and such changes propagate throughout the structure to other parts of the protein. The transmission of structural information happens through pathways of interconnected residues that are pre-encoded in the structural ensemble and can be one or more likely several. Mutations in these communication routes may affect the propagation of the signal [4]. In this sense, dynamical motions and intra-residue contacts are the underlying substrates through which structural communication happens.

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_10,

In this context, in silico methods, such as Molecular Dynamics (MD) simulations, have been used to identify the most important residues and pathways involved in structural communication, due to their ability to inform about how the protein behaves in the ensemble at least on a limited timescale [1, 2, 5–7]. To extract information about the most important residues involved in communication and possible structural communication pathways, as well as to identify key structurally important residues and important interactions between residues, protein structure networks (PSNs) have been proved to be a particularly useful representation of the protein structure and dynamics. In the following paragraph, we will summarize how PSNs are derived and analyzed, and detail our approach for their investigation – implemented in the PyInteraph software [8].

1.2 Definition of Protein Structure Networks

Networks or graphs have been used extensively to map and analyze relations between elements of a system in disparate fields. This powerful model allows to keep track of and study the behavior of complex systems in which the interaction of single elements gives rise to emergent behaviors. A graph is a mathematical model composed of unique elements (nodes) that are connected to each other through arcs or edges according to the existence of a specific relation between each pair. Such networks can be represented as square matrices called adjacency matrices, in which each row and column corresponds to an individual node, and a value different from zero is present in the respective position in the matrix if an edge is present between those two nodes. A number or weight can be associated to an edge in order to quantify the extent of their relation. Once a system is represented as a network, analysis techniques can be applied agnostically with respect to the type of data they encode, making many different analyses and methods available. Not surprisingly, the network paradigm has been applied in the field of structural biology as well, among others, in the form of protein structure networks (PSNs), also called residue interaction networks (RINs) or amino acid networks (AANs). Extensive reviews on what PSNs are and how they are used in the context of structural biology are available elsewhere [9–13] and here, we will summarize the main concepts behind them. A PSN is a network representation of the relations between residues in a protein, calculated either from a single experimental structure or an ensemble of conformations, depending on how the network is defined. The definition of what nodes and edges correspond to in the protein structure is what defines the network, the most common choice being to consider the monomeric unit of a protein (i.e., the residue) as a node. Edges are undirected, meaning they have no specified direction, and can be defined in a number of ways depending on which property or feature is used to measure the relation between residues, and if this relation is detected on a single protein structure or ensemble of

structures. Edge weight is often used to quantify the relation so that only edges that are associated with significant weights are retained; however, the decision of connecting two residues with an edge does not necessarily depend on the weight itself, and the two can be independent.

The most popular edge or weight definitions that have been employed so far are based, among others, on atomic side-chain contacts or otherwise defined interatomic distances, on interaction energy based on a force-field or knowledge-based potentials and others. Other network representations of protein dynamics are based, for instance, on local changes in the protein structure [14–16].

1.3 Analysis of PSNs

Once a PSN has been calculated from a protein structure or ensemble, different graph analysis techniques and network parameters can be calculated on the graph to extract useful information. One of the simplest and most important to calculate is the degree of each node, equal to the number of edges that are connected to that node. This allows identifying hubs, i.e. residues that are particularly well connected in the protein structure and are usually important for protein stability and as communication hubs [9].

Paths are successions of residues or edges which allow reaching a target residue from a source one. In this way, a chain of contacting residues through which structural communication may happen can be identified. Shortest paths are usually considered in PSNs as the most straightforward routes of structural communication [9].

Clusters or connected components are related with the global structure of the network and are defined as subgraphs in which paths exist between each pair of nodes but not with the rest of the network, representing more interconnected regions of the protein [9].

1.4 PyInteraph

Most definitions of PSNs rely on distances between atoms, the most popular ones being based on simple atomic contacts. While atomic contacts are indeed descriptive of the intraprotein interactions in an ensemble, they do not account for the different physicochemical properties of the protein amino acids and do not take in consideration specific types of non-covalent interactions that are known to be important for stability and dynamics. Such interactions are sometimes residue-specific (for instance in the case of salt bridges), meaning that their analysis can help in understanding how mutations affect the interaction network in the protein. Under these premises, a PSN based on different classes of the most relevant non-covalent interactions found in proteins may help us get a clearer picture of the interactions occurring among residues and complement more standard PSN analyses based on atomic contacts. This was the drive behind the development of PyInteraph, a set of tools designed to facilitate the generation and analysis of

network types based on specific residue-residue interaction networks. Since long-range communication can happen through less specific contacts, however, we also included a more general way to generate and analyze the PSN which is based on distances between centers of mass (cmPSN).

The primary function of PyInteraph is to compute, on conformational ensembles, three types of intramolecular interactions which are thought to be the most important non-covalent interactions in the protein structure: hydrophobic contacts, hydrogen bonds, and salt bridges. In PyInteraph, a hydrophobic contact is identified when the center of mass of the side chain of two hydrophobic residues is found within a given distance cut-off, which is 5 Å by default.

For salt bridges, the program is able to derive which charged groups belonging to side chains and main chain are present, depending on the topology. Charged groups are defined as those groups of charged residues (as Asp, Glu, Lys, Arg, and His, plus main-chain N- and C-termini) with a certain protonation state. Considering one pair of charged groups at the time, all the distances between atom pairs belonging to them are calculated, and they are selected as taking part in a salt bridge if at least one pair of atoms is found at a distance shorter than 4.5 Å by default.

A hydrogen bond is identified when a hydrogen and an acceptor atom are within 3.5 Å and the donor-hydrogen-acceptor angle is greater than 120°.

Each type of interaction is calculated independently and considered as a separate network at first. For each ensemble conformation, an interaction graph is calculated and two nodes are connected by an edge if at least one interaction of the specified type is found between the residues. The final network is constructed by counting, for each residue pair, the number of ensemble conformations in which two residues were connected by an edge over the total number of frames, times 100 to obtain a persistence value. In this way, one network per interaction type is collected, in which edges are weighted by the persistence of the given interaction in the ensemble. Two residues are connected by an edge in the final network if this value is above 0.0 or above a chosen significance cut-off.

Merging these networks to account for the interactions of each type is just a matter of keeping those edges that are present in at least one network, thus generating the so-called intramolecular interaction network (IIN).

PyInteraph also implements a knowledge-based potential [17] for the calculation of energy networks. Briefly, it is based on a four-distance description of interactions between specific atoms of side chains, which were chosen as those having the largest number of contacts in a dataset of high-resolution protein structures. The potential is calculated as follows:

$$\Delta E = -k_B T \ln (P_{real}/P_{rand})$$

where

$$P = P(\{\text{dist}\}|AA)P(AA)$$

$P(\{\text{dist}\}|AA)$ is the probability of identifying a specific combination of the four distances for a residue pair and $P(AA)$ is the probability to observe a contact between side chains for a residue pair in the structure. P_{real} was derived using experimental structures while P_{rand} was determined using a random model.

PyInteraph also includes scripts for the analysis of the networks, for the determination of persistence significance thresholds, and for the calculation of network properties such as hubs, connected components, and paths.

Finally, the suite includes a PyMOL plug-in for the visualization of the identified single interactions on the protein structure. Visualization of the networks is best done using the xPyder plug-in, detailed in Subheading 1.6.

PyInteraph has been used for the study of several biological systems, giving insight on local and long-range effects and allowing to investigate residues important for structural stability and intramolecular interactions [18–29].

1.5 Customization

PyInteraph has been built to be as customizable as possible and none of the parameters described throughout the text are hard-coded. The user can modify the cut-offs and definitions of the type of interactions described above. Charged groups are defined in a configuration file which can be modified with custom charged groups of any residue, natural or not, or even of other molecules. Similarly, another configuration file is available for hydrogen bonds, and the user can define which atoms are possible hydrogen bond donors or acceptors to consider non-standard residues. More details are available in the Notes.

1.6 Related Tools: xPyder

xPyder [30] is a PyMOL [31] plug-in initially designed to visualize and analyze networks of correlated motions on the protein structure, such as dynamic cross-correlation matrices. The program is, however, agnostic regard to the type of network it is able to plot and analyze, the only requirement being an adjacency matrix file encoding a weighted network which has precisely one node per protein residue in the input structure. Since the PSNs calculated by PyInteraph are by default stored in such a format, xPyder is the ideal tool to plot and visualize them. xPyder represents networks in the protein structures as cylinders connecting residues in the three-dimensional 3D structure, whose thickness depends on the weight of the associated edge. The plug-in includes options to filter the network according to a number of criteria, including edge weight, sequence proximity, distance between nodes, selection of specific

nodes and edges and more. It allows performing network analysis by calculating and visualizing hubs, connected components, and paths between selected residues on the structure. It also allows to calculate the difference matrix between two loaded matrices and visualize it as well. Finally, it supports customizing colors, thickness, and scale of the plotted interactions to produce publication-ready figures. xPyder has been used in a number of studies for the analysis and visualization of different types of networks, including those generated by PyInteraph [21, 32–38].

1.7 Our Test System: Cyclophilin A

We have used Cyclophilin A as a test case for our protocol. CypA is a peptidyl prolyl *cis-trans* isomerase involved in different biological functions and an important therapeutic target due to its involvement in pathological processes such as viral infection, cancer, cardiovascular diseases, neurodegeneration, aging, and others [39]. Different studies investigated CypA's dynamics and allosteric mechanisms and how these are coupled to catalysis and substrate recognition to shed light on its functions and exploit them for inhibitor drugs development and protein engineering [40–44]. It has been shown that dynamics of this enzyme happen on the same timescale as the catalytic turnover and that they are coupled to protein function, including long-range structural communication effects. Mutations of residue Ser99, 14 Å away from the active site, influence the dynamics of the catalytic residue Arg55 and affect reaction rates through a dynamic network of residues [40, 45]. Further experimental and computational works allowed to identify different dynamic pathways and distal regions in CypA that are involved in allosteric communication with the active site. For example, distal regions around the Val29 and Val6 residues have been recently identified to be involved in allosterically coupled dynamic networks in CypA and mutations of these key hotspot residues alter dynamics at the active site, modulating enzymatic function through allosteric networks [42, 44]. These reasons make CypA particularly suitable to be a good test case for our protocol.

2 Materials

The PyInteraph software was downloaded from https://github.com/ELELAB/PyInteraph and the xPyder PyMOL plug-in from https://github.com/ELELAB/xPyder. These packages were installed according to the installation instructions included with each. PyInteraph is written for Python 2.7 and includes C and Cython extensions, which need to be compiled during installation (*see* **Note 1** for more details). xPyder is a PyMOL plug-in and thus requires the PyMOL software to run.

We have used a 1 μs molecular dynamics simulation trajectory of wild-type CypA to demonstrate our protocol, whose set up is detailed elsewhere [25, 40]. Notably, this simulation was

performed using the GROMACS software and the CHARMM22* force field [46] and its trajectory was available in the XTC file format (*see* **Note 2** for details on compatibility with trajectory formats). The steps described below were carried out on a server running Ubuntu Server 14.04 from the command line, or on a common MacBook Pro running macOS when a graphical user interface was required, as when using xPyder.

3 Methods

3.1 Preparation of Topology and Trajectory for the Analysis

As most trajectory analysis tools, PyInteraph requires a topology and a trajectory file to work (*see* **Notes 2** and **3**). As the software relies on atom names to recognize different analysis groups, the names in the topology file should be consistent with atom and group definitions that the software uses, so that PyInteraph can correctly recognize the groups between which interactions are calculated. This means that the user can either modify the configuration file so that it matches the definitions in the topology file or, conversely, modify the topology file so that atom names match the definition in the configuration files. It should be noted that the latter uses standard PDB residue and atom names, meaning that they should fit most cases (*see* **Note 4** for more details on analysis customization). Nonetheless, depending on the setup of each simulation and used software, some adjustments may be required, as MD software frequently assign non-standard names to residues or atoms depending on the force-field definition. Running *pyinteraph* with the verbose option (*−v*) outputs the groups from the topology that the program will use for calculation, meaning that the user can always verify whether the program recognizes groups correctly. In particular:

1. For hydrophobic interactions, no adjustment is usually necessary. Nonetheless, when asking the software to use force-field masses during calculation (option *--ff-masses* of the *pyinteraph* executable), masses will be assigned according to residue and atom names. Standard atom and residue names are required to assign masses correctly, and the software will try to guess the element corresponding to the atom type and to assign a mass value accordingly if they are not already available. This is especially important for force fields that use non-standard masses for heavy atoms (as united-atoms or coarse-grained force fields; see details in **Note 5**).
2. For hydrogen bonds, no adjustment is necessary in case the topology file uses standard atom names.
3. For charged groups, no adjustment is necessary in case the topology file uses standard atom names and standard residue names.

4. The protein or proteins in the trajectory need to be made whole, meaning that no broken molecules should be present because of periodic boundary conditions. It is also good practice to remove any molecule or atom that does not need to be included in the analysis from topology and trajectory.

In the case of CypA, we modified the topology file by changing the non-standard names assigned by GROMACS to some hydrogen atoms to the standard ones, so that charged groups of arginine would be correctly recognized and that atomic masses would not need to be guessed. We also added the missing chain name in the topology pdb file ("A") so that the output files would contain it instead of the generic SYSTEM label, which is used when no chain definition is supplied. In the case of CypA, we filtered the trajectory by keeping only the protein atoms and we made sure that the protein was whole throughout the entire trajectory using the *gmx trjconv* program available in the GROMACS package.

3.2 Interaction Networks

After preparing the system, PyInteraph can be used to calculate the network of each intramolecular interaction type, which are hydrophobic contacts, salt bridges, and hydrogen bonds. The analysis writes as output two main files for each interaction type. One is the interactions file, a text file that contains the list of the non-covalent interactions identified in the ensemble. This file can be used in the interaction plotter PyMOL plug-in which is included in the distribution of PyInteraph (more in **Note 6**). The second output is a graph adjacency matrix file containing the adjacency matrix of the interactions with persistence values as weights, and which can be processed by using *filter_graph* and *graph_analysis* tools and visualized on the structure using the xPyder plug-in in PyMOL (*see* **Note 7** for more details on the file format). The interaction networks are produced as follows:

1. Calculate non-covalent interactions between residues from the MD trajectory. We used the *pyinteraph* program with options *-b* to analyze salt bridges, *-f* to analyze hydrophobic contacts, and *-y* to analyze hydrogen bonds. For each interaction type, we collected the corresponding interaction files and the adjacency matrix files. The files containing the topology of the system and the trajectory were specified with *-s* and *-t* options, respectively. We used the *--ff-masses* option to perform the calculation using the definition of masses from the CHARMM27 force field (*see* **Note 5** for details). The output files are defined by the *--sb-graph*, *--hb-graph*, *--hc-graph* options for the graph adjacency matrix files of salt bridges, hydrogen bonds, and hydrophobic clusters, respectively. We calculated the hydrogen bonds considering only the donor-acceptor atoms located on the side chains of the residues using the option *--hb-class sc-sc*, in order to exclude hydrogen bonds involved in the formation of the

secondary structure. We performed the calculation of salt bridges considering the distances between positively and negatively charged groups, as by default (*see* **Note 4** for more details). In this case, we decided to keep all the edges with a persistence value higher than zero (*see* **Note 8**).

2. Identify the persistence significance threshold (P_{crit}). In order to remove the most transient interactions, we used the *filter_graph* tool to perform an estimation of the threshold of significance for the interaction persistence in the graph (P_{crit}). P_{crit} is calculated by filtering the original graph several times for increasing persistence threshold values, which is accomplished by removing edges having a weight lower than a given cut-off. For each filtered graph, the size of the biggest connected component is then calculated. The plot resulting from the procedure, as the one from the graph of hydrophobic contacts shown in Fig. 1a, generally exhibits an abrupt decrease, often with a sigmoid-like shape for globular proteins, with a central point that represents a threshold with a good balance between having a too interconnected and a too sparse network. Based on this analysis, we set a persistence threshold of 20, which was also compatible and used for the salt bridges and hydrogen bond graphs (not shown). This threshold indicates the minimum percentage of frames in which an interaction should be present over the whole ensemble to be considered.
3. Filter each graph according to the identified persistence threshold. We used the *filter_graph* program to perform graph filtering according to the selected P_{crit} threshold of 20, removing all the edges with weight lower than this value. We specified the graph to be filtered with the option *-d* and the persistence threshold using the *-t* option. Figure 1b–d shows the filtered salt bridges, hydrogen bonds, and hydrophobic cluster networks, respectively.
4. Visualization and analysis of the calculated networks. We plotted and visualized each interaction type on the reference structure, using the xPyder plug-in for PyMOL. To do so, we loaded the topology PDB file in PyMOL, opened the xPyder plug-in, and loaded an adjacency matrix file. We then generated the corresponding graph in the Graph analysis tab and visualized hubs, defined as residues involved in three or more distinct interactions.

It should be noted that network properties, such as hubs and connected components, can be analyzed using either xPyder or the *graph_analysis* tool in PyInteraph. See section 3.5 for an example of the latter.

Salt bridges can have local and long-range effects in proteins, especially in solvent-exposed and disordered regions that have a

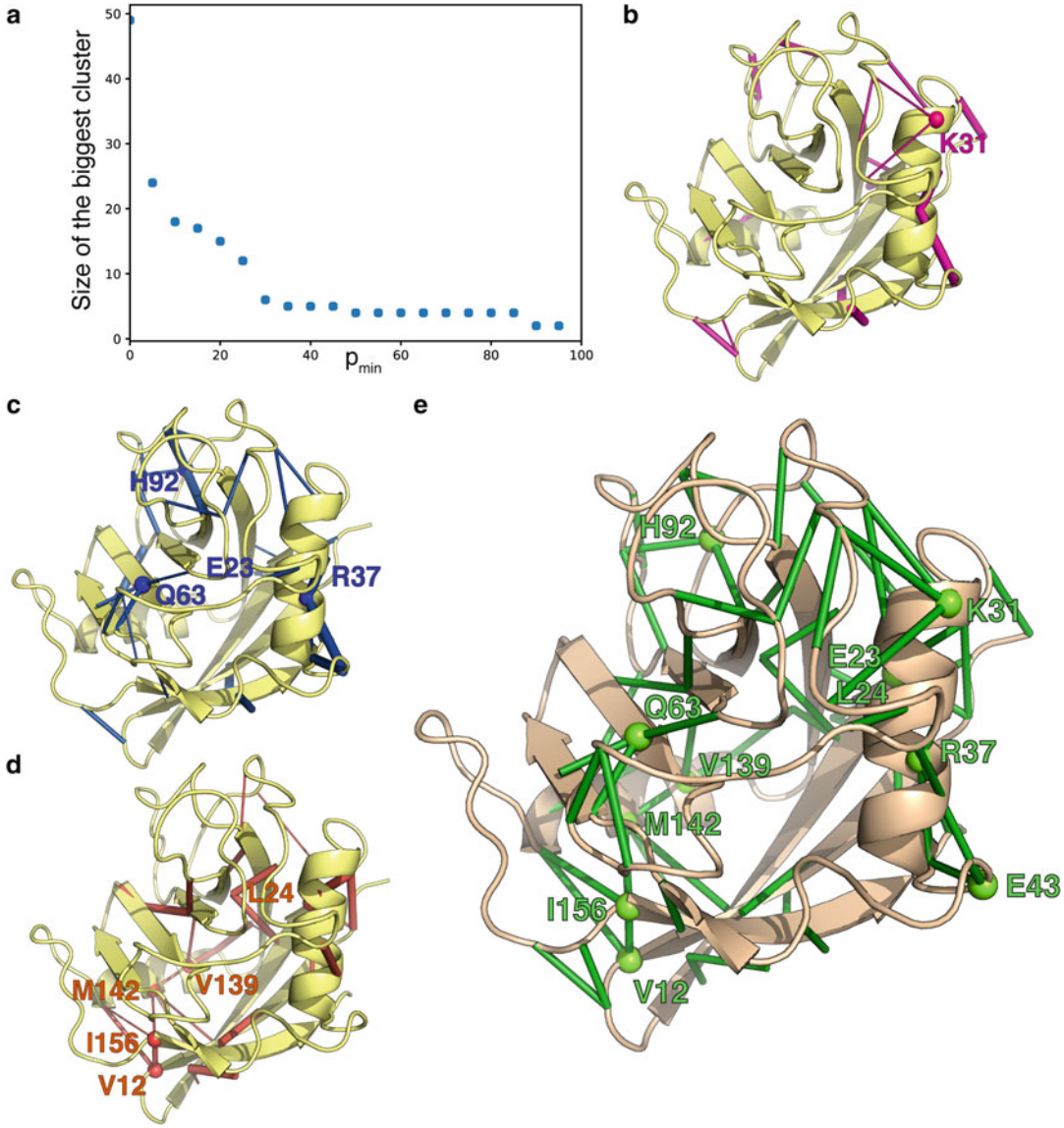

Fig. 1 (**a**): Size of the biggest cluster calculated for the hydrophobic contact network at varying cut-off of persistence (P_{min}). (**b**–**d**) Networks of salt bridges, hydrogen bonds between side chains and hydrophobic clusters, respectively. In these networks, an edge is represented by a cylinder connecting C_α atoms of residues, and the different thickness of the cylinders is proportional with the persistence value. Residues having their C_α shown as sphere and labelled are network hubs, i.e. residues connected by at least three edges. The IIN is plotted according to the same representation, except for cylinder thickness as the IIN is unweighted, in panel (**e**)

high content of charged residues, playing roles in protein stability and function [47, 48]. The identified salt bridges (Fig. 1b) are mostly localized on the protein surface and the only hub identified is Lys31 that makes interactions with Glu81, Glu84, and Glu86. Lys31 is located far away from the active site (~17 Å) and is not

directly involved in catalysis or substrate binding. This residue has been proposed by a recent computational study to be coupled to residues Val6 and Val29 [44], which have been identified as key mutation hotspot residues that communicate by allosterically coupled dynamic networks in CypA, affecting enzyme reaction rates [42]. In the network of hydrogen bonds, we identified as hubs the Arg37, His92, Glu23, and Gln63 (Fig. 1c). Hydrophobic contacts are usually crucial to maintain protein structure and stability, composing the major interactions between the residues in the protein core, that are tightly packed and shielded to the solvent. We identified two clusters in the network of hydrophobic interactions: one localized around the N-terminal of CypA, comprising Leu 24, Pro4, Val6, Ala26, and Ala33 and one at the other side of the antiparallel β-sheet, comprising Val12, Leu17, Val139, Met142, Phe145, Ile156 (Fig. 1d).

3.3 Construction and Analysis of the IIN

PyInteraph can be used to obtain a general view of the interactions in a protein ensemble, without considering the type of each intramolecular interaction and building a common map for their visualization and analysis. In order to do so, the different interaction graphs calculated in the previous section are combined in a meta-Intramolecular Interaction Network (IIN). In this network, an edge between two residues is present if at least one of the interaction graphs has an edge between them. The network is by default unweighted (i.e., all weights are set to 1.0), but weights can be added using a knowledge-based potential implemented in PyInteraph as detailed in Sect. 3.4. The following steps allow to derive the IIN from the calculated interaction graphs:

1. Combine interaction graphs to obtain the IIN. We used the *filter_graph* tool to combine the filtered interaction graphs described previously, by supplying the three filtered interaction graphs using the *-d* option multiple times.
2. Visualization and analysis of the network. We plotted and visualized the IIN on the reference structure, using the xPyder plug-in for PyMOL as reported in Fig. 1e. To obtain the results shown, we loaded the topology PDB structure in PyMOL and used the xPyder plug-in to visualize the IIN, as detailed in the previous paragraph. We visualized hubs, highly connected nodes in the graph involved in three or more different interactions using the Graph analysis tools in xPyder, as explained previously.

The IIN permits to obtain an overall description of all the most persistent non-covalent interactions in the ensemble and their location on the protein structure. Together with the calculation of hubs, it gives an idea of the most relevant residues in the network, possibly important for stability and structural communication. In the IIN, we identified several residues with high connectivity in the

network and possible communication hubs: Val12, Glu23, Leu24, Lys31, Arg37, Glu43, Gln63, His92, Val139, Met142, and Ile156. The central role in the network of contacts of the Gln63 is particularly interesting since it interacts directly with the substrate through hydrogen bonds, and both NMR experiments and computational investigations identified this residue as involved in the major conformational processes of CypA [40, 49]. It should be noted that the IIN contains all the analyzed interactions in the network, without taking into account nonspecific contacts that can still take part in communication pathways in the network. For this reason, we also considered a more generic analysis, cmPSN, which takes into account any possible contact between residues (*see* Sect. 3.5).

3.4 Energy Interaction Network

As outlined in the introduction, PyInteraph can be used to calculate average interaction knowledge-based potential pseudo-energies based on sets of four distances between pairs of residues. This is especially useful to have an idea of how favorable the interactions identified in the IIN or in the cmPSN are, especially for the IIN as it has no associated weights by default. Given the definition of the knowledge-based potential, negative values account for favored interactions while positive values account for unfavored interactions. The weighted IIN has been obtained as follows:

1. Calculate the interaction energy potential for all possible pairs of residues. This is performed similarly as the other analyses, using the *pyinteraph* program with the *-p* option. Lists of interaction energies per pair of residues are written in the file specified by *--kbp-dat* while the same information in adjacency matrix form is saved with *--kbp-graph*. The calculated energy network can be used as-is or it can be used to weight any of the networks we calculate.
2. Assign weights to the IIN. This is done using the *filter_graph* executable, with option *-d* for the IIN adjacency matrix file and option *-w* for the knowledge-based potential graph.
3. Visualize the weighted IIN. We plotted the obtained weighted graph using xPyder, as detailed above for the other interaction graphs. *See* **Note 9** for more details on network analysis on this type of network.

The obtained energy network can be used to understand which interactions are most favored among those identified and is shown in Fig. 2. All except one of the identified interactions are found to be somehow favorable when scored by the knowledge-based potential, however, with different magnitudes. As expected, there is a degree of correlation between the most stable interactions and their strengths, although some highly persistent interactions are not considered to be strong when scored with the potential. This is

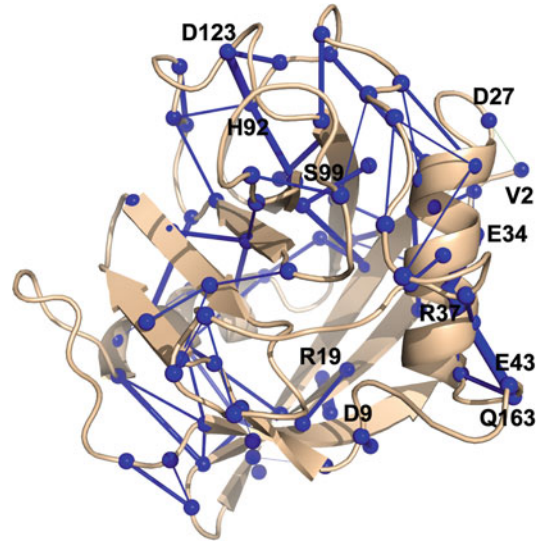

Fig. 2 IIN re-weighted using the knowledge-based potential implemented in PyInteraph. The same representation is used as in Fig. 1, except for spheres that here represent all residues. Blue edges represent negative values, while green edges represent positive values. The thickness of the cylinders is proportional with absolute value of energy

especially true for hydrophobic contacts. For instance, Ala26 contacts Ala33 and Pro4, but the found interactions are scored with low absolute values of potential. Salt bridges are found to be among the most favored interactions, with the associated magnitude roughly correlated with the persistence in the ensemble. The only positive value among the analyzed interactions is detected on a salt bridge, between the charged N-terminal main-chain group and Asp27. It should be noted, however, that the potential considers distances between side-chain atoms exclusively, therefore it doees not consider the interactions between terminal groups and residues. Among the interactions that are more energetically favored, we list three salt bridges (Arg19-Asp9, Arg37-Glu34, Arg37-Glu43), and three very persistent hydrogen bonds (Arg37-Gln163, Asp123-His92, Ser99-His92). We note once again the role of Arg37, which acts as an interaction hub and is involved in two highly favored salt bridges and two persistent hydrogen bonds of which one is particularly energetically favorable. Similarly, also residue His92 was found to be part of a highly persistent network of energetically favorable hydrogen bonds.

3.5 Center of Mass PSN

While the IIN takes into account the most prevalent but specific type of interactions found in proteins, the transmission of structural information can also happen through nonspecific contacts. For this reason, we devised a more general analysis that takes into

consideration all residues with a side chain of any type. This is called the center of mass PSN (cmPSN) and relies on the identification of contacts between side chains when the distance between their center of mass is below a certain threshold. The calculation of the cmPSN works in the same way as for hydrophobic contacts, except that the calculation is extended to any residue except glycine. This is possible as PyInteraph allows to specify which residues the hydrophobic contact interactions should be calculated among, potentially including non-hydrophobic residues as well. Especially when running this type of analysis, the distance cut-off between centers of mass for the definition of a contact has a major influence on the resulting cmPSN topology. We used a 5 Å cut-off, as suggested by Salamanca Viloria et al. [25], which has been determined to be a good value in the context of MD simulation with atomistic force fields; this might be different depending on your setup and system (*see* **Note 10**). In order to generate and analyze the cmPSN, we carried out the following steps:

1. Calculate the cmPSN. We ran the *pyinteraph* command with distance cut-off equal to 5 Å specifying all the residue types except glycine for hydrophobic contacts (option *--hc-residues*). We specified the files containing the topology of the system and the trajectory with the *-s* and *-t* options, respectively, and the output file with the *--hc-graph* option. The latter contains the adjacency matrix representing the cmPSN. We set the software to use masses from the CHARMM27 force field for the reasons described above (*see* **Note 5**) using the *-ff-masses* option. We chose not to set any persistence cut-off for the interactions in this phase, as per **Note 8**.
2. Filter the cmPSN to remove transient interactions. Similarly as done before with the interaction network, we filtered the adjacency matrix file to remove the most transient interactions, which means edges with low persistence values, using the *filter_graph* program. This requires as input the unfiltered cmPSN generated with the previous command (*−d* option), and generates an output file with the same format, whose name is specified with the *-o* option. We specified the persistence threshold using the *-t* option. This threshold represents the minimum percentage of frames in which the interaction must be present over the whole trajectory. We set this threshold to 20, as previously done in previous benchmarking works [25].
3. Graph analysis. We then carried out basic graph analysis on the resulting graphs, using the *graph_analysis* program, as detailed above. We used the filtered cmPSN outputted by *filter_graph* as input and considered any residue connected with 3 or more other residues to be a hub node by setting the *-k* option to *3*, which sets the minimum hubs degree. Similarly,

graph_analysis allows calculating connected components on the network by using opion *-c*. We specified the topology with the *-r* option to be able to map the graph nodes and edges onto the structure. Using the *-ub* option, we wrote a PDB file containing the reference structure along with the degree for each hub residue in the b-factor column. Likewise, option *-cb* allows to write a PDB file containg the connected component number for each residue in the b-factor column.

4. Calculation of shortest communication paths. We used the *graph_analysis* program to identify pathways of structural communication between a specific pair of residues. By default, *graph_analysis* calculates all simple paths between a given pair of residues up to a certain length. We first ran *graph_analysis* with the *-r* and *-a* options alone so that it would print a list of node names, which are used to specify which residues we will calculate paths between. We then ran *graph_analysis* again with the *-p* option so that it would try to calculate paths between the residues specified by options *-r1* and *-r2* (see below for details), using the node names as detailed above. By default, the maximum path length is 3. No paths were present with this length and the program reported a minimum path length of 5, which corresponds to the length of the shortest paths. Finally, we ran the same command line this time specifying a length of 5 with option *-l*. The found paths were printed to the standard output. We used option *-d* to save each path as an independent subgraph (i.e., adjacency matrix).
5. Visualization of connected components on the protein structure. We used the xPyder plug-in to visualize the connected components on the reference structure, as can be seen in Fig. 3b. This was performed by loading the topology PDB structure in PyMOL, loading the adjacency matrix in the xPyder plug-in, then generating the graph and calculating the connected components in the Graph analysis tab, and finally plotting them one by one changing the plotting color in the main tab every time.
6. Visualization of hubs on the protein structure. The hubs were subsequently visualized by loading the PDB file obtained with the *-ub* option into the PyMOL software. We used the *putty b-factor* in PyMOL representation that changes the thickness and color of residues according to the values in the b-factor column. Similarly, the color scale ranges from yellow to red as the value increases.

By performing the graph analysis on the center of mass PSN in CypA, we found 15 connected components, the biggest of which were composed by 58, 19, and 11 residues, with the first connected component comprising most of the secondary structure elements in the protein and representing a large part of the protein core

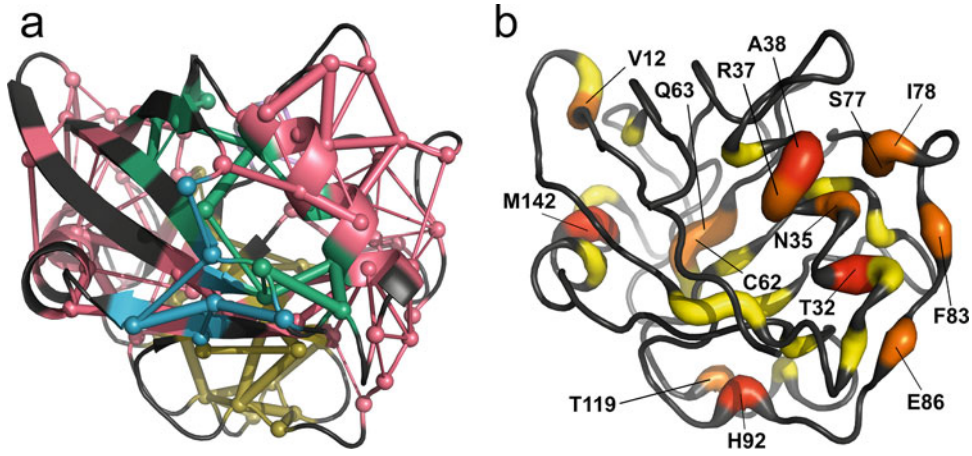

Fig. 3 (**a**) The five most populated connected components mapped on the reference structure. The first connected component is by far the largest and is displayed in pink. The other four connected components are showed in yellow, green, light blue, and purple, respectively. (**b**) Hub nodes. Hubs are color-coded according to their degree: red corresponds to 5, orange to 4, and yellow to 3. Labels are shown only for hubs with degree 4 or 5

(Fig. 3a). Figure 3b shows the identified hub residues colored in yellow, orange, and red. As can be seen, the red hubs Tyr32, Ala38, His92, and Met142 are the ones with the highest degree (degree = 5); hence, the residues most prone to behave as communication hubs. Finally, we calculated shortest paths between the catalytic residue Arg55 and Ser99. It is known that mutations in position 99, which is remote from the active site, can influence reaction rates. Previous works identified a dynamic network connecting these two residues, underlying the existence of a structural communication path between them [40]. PyInteraph also identifies a similar communication network in the form of two five-residue paths that connect the two endpoints. It should be noted that such a network cannot be represented in the IIN as it involves unspecific interactions between residues; nonetheless, the IIN and its composing networks can still be used to recover significant interactions between residues once pathways have been identified in the cmPSN.

4 Notes

1. PyInteraph is not compatible with the most current versions of the MDAnalysis package [50, 51]. We suggest to install MDAnalysis and PyInteraph in a separate environment, such as a Python *virtualenv*, so that different versions of the same library can coexist in the operating system if need be. The required version is specified in the installation instructions. An up to

date version compatible with Python 3 is currently in the making and will be released soon, and will be compatible with the most recent MDAnalysis currently released (1.0.0).

2. PyInteraph is compatible with all the trajectory formats supported by MDAnalysis, the most popular being XTC, NETCDF, and DCD. It is also compatible with many topology formats, with PDB being the most widespread and commonly used.
3. PyInteraph also supports as input a reference structure file. This is useful for cases in which we want chain definition and residue numbering in the output files to be different from the ones in the topology. This can be useful for instance for cases in which the MD software has changed the residue numbering in the topology with respect to the experimental structure or no chain definition is present in the topology file (as happens in GRO files). A reference file needs to have the same number of residues in the same order as topology file, but the number of atoms can differ – so that an experimentally solved structure can be used. When no reference file is provided, the topology is used as the reference instead – as we did in this case for CypA.
4. The default configuration files for PyInteraph are located in the PyInteraph installation directory and allow to set which atom groups constitute charged groups (*charged_groups.ini*) as well as to set which atom types can act as acceptor or donor for hydrogen bonds (*hydrogen_bonds.ini*). The files are read by default from the main installation directory, but the user can specify their specific versions with the command-line options *--sb-cg-file* and *--hb-ad-file*. The file formats are straightforward: for charged groups, the user defines specific charged groups under the [CHARGED_GROUPS] section. Charged groups can have any name but must end either with "p" or "n," respectively, for positively and negatively charged. Each group is defined as a list of atom names; if an atom needs not to present for the group to be considered as charged, its name is prefixed by an exclamation mark. The default_charged_groups entry is a list of charged groups that any residue might have. Charged groups are finally assigned to residues, defined as residue names, in the [RESIDUES] section. The *hydrogen_bonds.ini* file just contains lists of acceptor and donor atoms. By following these conventions, it is possible to include in the analysis even non-natural amino acids or small molecules. Further customization is possible through command-line options. It is possible to calculate electrostatic interactions between residues of the same charge using the --sb-mode option (*--sb-mode*

same_charge) or all of them (*--sb-mode all*). *pyinteraph* also allows to analyze main chain-main chain (*--hb-class mc-mc*), main chain-side chains (*--hb-class mc-sc*), all (*--hb-class all*), or hydrogen bonds between custom groups (*--hb-class custom*). Custom groups of atoms are defined using MDAnalysis selection format through options *--hb-custom-group-1* and *--hb-custom-group-2*. As far as hydrophobic contacts are considered, the user can decide which residue needs to be included in the analysis using the *--hc-residues* option.

5. *pyinteraph* supports assigning different masses to atoms depending on the force field, which is especially important for the calculation of the correct centers of mass. This is relevant when considering united-atom force fields, such as GROMOS, or even coarse-grained systems. PyInteraph supports the GROMOS, AMBER, CHARMM, ENCAD, and OPLS force-field families, but more can be added in the form of JSON-formatted files. In the case of CypA, whose simulations use the CHARMM22* force field, we used atomic mass from CHARMM27 after checking that they were identical in the force-field definition files of GROMACS. Masses files are found in the *ff_masses* directory in the PyInteraph installation directory.
6. *interaction_plotter* is a second PyMOL plug-in designed to plot on the 3D structures single interactions between residues groups encoded in single interaction files (*see* **Note 7**). These include side-chain and main-chain atoms, and each residue can interact through different groups, depending on the interaction definitions. As the xPyder plug-in supports only one node per residue, while intramolecular interactions can involve more than one group per residue, the two formats are not intercompatible.
7. Interaction networks are written by PyInteraph into two different text formats. One details every single interaction found in the ensemble and lists the groups of the two residues that are interacting together with the associated persistence value. The other is a graph adjacency matrix, which is simply an ASCII square symmetric matrix in which every line and every column represents a residue, ordered as the residues in the protein under study. Each position of the matrix represents a single edge weight, which is 0.0 if no interactions were found in the ensemble (which means the edge does not exist in the context of PyInteraph), or a value in the (0.0,100.0] range if they have been found. The simplicity of this matrix format allows it to be easily read in most programming languages and it is compatible with the xPyder PyMOL plug-in.

8. The *pyinteraph* program supports the option of writing filtered graphs directly by specifying a persistence threshold with options *--hc-perco*, *--hb-perco*, and *--sb-perco*. However, we usually prefer not to perform any filtering at this stage so that the original graph can be filtered more than once using *filter_graph*, in case different values of P_{crit} need to be tested.
9. xPyder graph analysis uses positive values only from the loaded matrix. This works when using most of the networks identified in this protocol, but can be problematic when using the knowledge-based potential network in which the favorable interactions are expressed as negative values. In that case we suggest changing the sign of the energy values in the adjacency matrix file before loading it into xPyder in order to consider favored interactions when doing the graph analysis.
10. While this cut-off has been rationalized and validated on some among the most popular force fields, it might not be the best for other cases especially in which the definition of the topology changes significantly (as for coarse-grained systems).

Acknowledgments

The authors would like to thank Elena Papaleo and Emmanuelle Bignon for fruitful comments and suggestions. This work was supported by Carlsberg Foundation Distinguished Fellowship (CF18-0314), The Danish Council for Independent Research, Natural Science, Project 1 (102517), Danmarks Grundforskningsfond (DNRF125) to our group.

References

1. Ribeiro AAST, Ortiz V (2016) A chemical perspective on allostery. Chem Rev 116:6488–6502. https://doi.org/10.1021/acs.chemrev.5b00543
2. Papaleo E, Saladino G, Lambrughi M, Lindorff-Larsen K, Gervasio FL, Nussinov R (2016) The role of protein loops and linkers in conformational dynamics and allostery. Chem Rev 116:6391–6423. https://doi.org/10.1021/acs.chemrev.5b00623
3. Guo J, Zhou H-X (2016) Protein allostery and conformational dynamics. Chemical Reviews 116:6503–6515. https://doi.org/10.1021/acs.chemrev.5b00590
4. del Sol A, Tsai C-J, Ma B, Nussinov R (2009) The origin of allosteric functional modulation: multiple pre-existing pathways. Structure 17:1042–1050. https://doi.org/10.1016/j.str.2009.06.008
5. Feher VA, Durrant JD, Van Wart AT, Amaro RE (2014) Computational approaches to mapping allosteric pathways. Curr Opin Struct Biol 25:98–103. https://doi.org/10.1016/j.sbi.2014.02.004
6. Hertig S, Latorraca NR, Dror RO (2016) Revealing atomic-level mechanisms of protein allostery with molecular dynamics simulations. PLoS Comput Biol 12:e1004746. https://doi.org/10.1371/journal.pcbi.1004746
7. Allain A, Chauvot de Beauchêne I, Langenfeld F, Guarracino Y, Laine E, Tchertanov L (2014) Allosteric pathway identification through network analysis: from molecular dynamics simulations to interactive 2D and 3D graphs. Faraday Discuss 169:303–321. https://doi.org/10.1039/c4fd00024b
8. Tiberti M, Invernizzi G, Lambrughi M, Inbar Y, Schreiber G, Papaleo E (2014)

PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. J Chem Inf Model 54:1537–1551. https://doi.org/10.1021/ci400639r

9. Bhattacharyya M, Ghosh S, Vishveshwara S (2016) Protein structure and function: looking through the network of side-chain interactions. Curr Protein Pept Sci 17:4–25
10. Di Paola L, De Ruvo M, Paci P, Santoni D, Giuliani A (2013) Protein contact networks: an emerging paradigm in chemistry. Chem Rev 113:1598–1613. https://doi.org/10.1021/cr3002356
11. Yan W, Zhou J, Sun M, Chen J, Hu G, Shen B (2014) The construction of an amino acid network for understanding protein structure and function. Amino Acids 46:1419–1439. https://doi.org/10.1007/s00726-014-1710-6
12. Di Paola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48. https://doi.org/10.1016/j.sbi.2015.03.001
13. Hu G, Zhou J, Yan W, Chen J, Shen B (2013) The topology and dynamics of protein complexes: insights from intra- molecular network theory. Curr Protein Pept Sci 14:121–132
14. Sethi A, Eargle J, Black AA, Luthey-Schulten Z (2009) Dynamical networks in tRNA:protein complexes. Proc Natl Acad Sci U S A 106:6620–6625. https://doi.org/10.1073/pnas.0810961106
15. Pandini A, Fornili A, Fraternali F, Kleinjung J (2012) Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. FASEB J 26:868–881. https://doi.org/10.1096/fj.11-190868
16. Pandini A, Fornili A, Fraternali F, Kleinjung J (2013) GSATools: analysis of allosteric communication and functional local motions using a structural alphabet. Bioinformatics 29:2053–2055. https://doi.org/10.1093/bioinformatics/btt326
17. Potapov V, Cohen M, Inbar Y, Schreiber G (2010) Protein structure modelling and evaluation based on a 4-distance description of side-chain interactions. BMC Bioinformatics 11:374. https://doi.org/10.1186/1471-2105-11-374
18. Jónsdóttir LB, Ellertsson BÖ, Invernizzi G, Magnúsdóttir M, Thorbjarnardóttir SH, Papaleo E, Kristjánsson MM (2014) The role of salt bridges on the temperature adaptation of aqualysin I, a thermostable subtilisin-like proteinase. Biochim Biophys Acta 1844:2174–2181. https://doi.org/10.1016/j.bbapap.2014.08.011
19. Lambrughi M, De Gioia L, Gervasio FL, Lindorff-Larsen K, Nussinov R, Urani C, Bruschi M, Papaleo E (2016) DNA-binding protects p53 from interactions with cofactors involved in transcription-independent functions. Nucleic Acids Res 44:9096–9109. https://doi.org/10.1093/nar/gkw770
20. Marino V, Scholten A, Koch K-W, Dell'Orco D (2015) Two retinal dystrophy-associated missense mutations in GUCA1A with distinct molecular properties result in a similar aberrant regulation of the retinal guanylate cyclase. Hum Mol Genet 24:6653–6666. https://doi.org/10.1093/hmg/ddv370
21. Papaleo E, Parravicini F, Grandori R, De Gioia L, Brocca S (2014) Structural investigation of the cold-adapted acylaminoacyl peptidase from Sporosarcina psychrophila by atomistic simulations and biophysical methods. Biochim Biophys Acta 1844:2203–2213. https://doi.org/10.1016/j.bbapap.2014.09.018
22. Óskarsson KR, Nygaard M, Ellertsson BÖ, Thorbjarnardóttir SH, Papaleo E, Kristjánsson MM (2016) A single mutation Gln142Lys doubles the catalytic activity of VPR, a cold adapted subtilisin-like serine proteinase. Biochim Biophys Acta 1864:1436–1443. https://doi.org/10.1016/j.bbapap.2016.07.003
23. Nygaard M, Terkelsen T, Vidas Olsen A, Sora V, Salamanca Viloria J, Rizza F, Bergstrand-Poulsen S, Di Marco M, Vistesen M, Tiberti M, Lambrughi M, Jäättelä M, Kallunki T, Papaleo E (2016) The mutational landscape of the oncogenic MZF1 SCAN domain in cancer. Front Mol Biosci 3:78. https://doi.org/10.3389/fmolb.2016.00078
24. Michetti D, Brandsdal BO, Bon D, Isaksen GV, Tiberti M, Papaleo E (2017) A comparative study of cold- and warm-adapted endonucleases A using sequence analyses and molecular dynamics simulations. PLoS One 12:e0169586. https://doi.org/10.1371/journal.pone.0169586
25. Salamanca Viloria J, Allega MF, Lambrughi M, Papaleo E (2017) An optimal distance cutoff for contact-based protein structure networks using side-chain centers of mass. Sci Rep 7:2838. https://doi.org/10.1038/s41598-017-01498-6
26. Marino V, Dell'Orco D (2016) Allosteric communication pathways routed by Ca2+/Mg2+ exchange in GCAP1 selectively switch target regulation modes. Sci Rep 6:517. https://doi.org/10.1038/srep34277

27. Lambrughi M, Lucchini M, Pignataro M, Sola M, Bortolotti CA (2016) The dynamics of the β-propeller domain in Kelch protein KLHL40 changes upon nemaline myopathy-associated mutation. RSC Adv 6:34043–34054. https://doi.org/10.1039/C6RA06312H
28. Singh B, Bulusu G, Mitra A (2016) Effects of point mutations on the thermostability of B. subtilis lipase: investigating nonadditivity. J Comput Aided Mol Des 30:899–916. https://doi.org/10.1007/s10822-016-9978-0
29. Otaki H, Taguchi Y, Nishida N (2018) Molecular dynamics simulation reveals that switchable combinations of β-sheets underlie the prion-like properties of α-synuclein amyloids. https://doi.org/10.1101/326462
30. Pasi M, Tiberti M, Arrigoni A, Papaleo E (2012) xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. J Chem Inf Model 52:1865–1874. https://doi.org/10.1021/ci300213c
31. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
32. Vijayan RSK, Arnold E, Das K (2014) Molecular dynamics study of HIV-1 RT-DNA-nevirapine complexes explains NNRTI inhibition and resistance by connection mutations. Proteins 82:815–829. https://doi.org/10.1002/prot.24460
33. Hashem S, Tiberti M, Fornili A (2017) Allosteric modulation of cardiac myosin dynamics by omecamtiv mecarbil. PLoS Comput Biol 13:e1005826. https://doi.org/10.1371/journal.pcbi.1005826
34. Guizado TRC (2014) Analysis of the structure and dynamics of human serum albumin. J Mol Model 20:43. https://doi.org/10.1007/s00894-014-2450-y
35. Papaleo E, Renzetti G, Invernizzi G, Ásgeirsson B (2013) Dynamics fingerprint and inherent asymmetric flexibility of a cold-adapted homodimeric enzyme. A case study of the vibrio alkaline phosphatase. Biochim Biophys Acta Gen Subj 1830:2970–2980. https://doi.org/10.1016/j.bbagen.2012.12.011
36. Lambrughi M, Papaleo E, Testa L, Brocca S, De Gioia L, Grandori R (2012) Intramolecular interactions stabilizing compact conformations of the intrinsically disordered kinase-inhibitor domain of Sic1: a molecular dynamics investigation. Front Physiol 3:435. https://doi.org/10.3389/fphys.2012.00435
37. Invernizzi G, Tiberti M, Lambrughi M, Lindorff-Larsen K, Papaleo E (2014) Communication routes in ARID domains between distal residues in helix 5 and the DNA-binding loops. PLoS Comput Biol 10:e1003744. https://doi.org/10.1371/journal.pcbi.1003744
38. Invernizzi G, Lambrughi M, Regonesi ME, Tortora P, Papaleo E (2013) The conformational ensemble of the disordered and aggregation-protective 182-291 region of ataxin-3. Biochim Biophys Acta 1830:5236–5247. https://doi.org/10.1016/j.bbagen.2013.07.007
39. Nigro P, Pompilio G, Capogrossi MC (2013) Cyclophilin A: a key player for human disease. Cell Death Dis 4:e888–e888. https://doi.org/10.1038/cddis.2013.410
40. Papaleo E, Sutto L, Gervasio FL, Lindorff-Larsen K (2014) Conformational changes and free energies in a proline isomerase. J Chem Theory Comput 10:4169–4174. https://doi.org/10.1021/ct500536r
41. Schlegel J, Armstrong GS, Redzic JS, Zhang F, Eisenmesser EZ (2009) Characterizing and controlling the inherent dynamics of cyclophilin-A. Protein Sci 18:811–824. https://doi.org/10.1002/pro.89
42. Holliday MJ, Camilloni C, Armstrong GS, Vendruscolo M, Eisenmesser EZ (2017) Networks of dynamic allostery regulate enzyme function. Structure 25:276–286. https://doi.org/10.1016/j.str.2016.12.003
43. Doshi U, Holliday MJ, Eisenmesser EZ, Hamelberg D (2016) Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. Proc Natl Acad Sci U S A 113:4735–4740. https://doi.org/10.1073/pnas.1523573113
44. Rodriguez-Bussey I, Yao X-Q, Shouaib AD, Lopez J, Hamelberg D (2018) Decoding allosteric communication pathways in cyclophilin a with a comparative analysis of perturbed conformational ensembles. J Phys Chem B 122:6528–6535. https://doi.org/10.1021/acs.jpcb.8b03824
45. Fraser JS, Clarkson MW, Degnan SC, Erion R, Kern D, Alber T (2009) Hidden alternative structures of proline isomerase essential for catalysis. Nature 462:669–673. https://doi.org/10.1038/nature08615
46. Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? Biophys J 100:L47–L49. https://doi.org/10.1016/j.bpj.2011.03.051
47. Jelesarov I, Karshikoff A (2009) Defining the role of salt bridges in protein stability. Methods Mol Biol 490:227–260. https://doi.org/10.1007/978-1-59745-367-7_10

48. Kumar S, Nussinov R (2002) Close-range electrostatic interactions in proteins. Chembiochem 3:604–617. https://doi.org/10.1002/1439-7633(20020703)3:7<604::AID-CBIC604>3.0.CO;2-X

49. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, Kern D (2005) Intrinsic dynamics of an enzyme underlies catalysis. Nature 438:117–121. https://doi.org/10.1038/nature04105

50. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. J Comput Chem 32:2319–2327. https://doi.org/10.1002/jcc.21787

51. Gowers RJ, Linke M, Barnoud J, Reddy TJE, Melo MN, Seyler SL, Dotson DL, Domanski J, Buchoux S, Kenney IM, Beckstein O (2016) MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In: Rostrup S, Benthall S (eds) th Python in science conference. Austin, Texas, pp 102–109

Chapter 11

The Allosteric Effect in Antibody-Antigen Recognition

Jun Zhao, Ruth Nussinov, and Buyong Ma

Abstract

We studied the molecular details of the recognition of antigens by the variable domain of their cognate antibodies in as well as those elicited by the constant domains, which do not directly interact with antigens. Such effects are difficult to study experimentally; however, molecular dynamics simulations and subsequent residue interaction network analysis provide insight into the allosteric communication between the antigen-binding CDR region and the constant domain. We performed MD simulations of the complex of Fab and prion-associated peptide in the apo and bound forms and follow the conformational changes in the antibody and cross-talk between its subunits and with antigens. These protocols could be generally applied for studies of other antigens-antibody recognition systems.

Key words Antibody-antigen interaction, Motion correlation, Dynamic network, Community analysis, Disulfide bond, Allosteric effect

1 Introduction

Protein conformational dynamics and fluctuations in water are intrinsic thermodynamic phenomena, with the distributions of the states on the energy landscape determined by statistical thermodynamics. Protein dynamics and conformational changes have been optimized by evolution to perform biological functions [1]. Proteins are intrinsically allosteric, and their residue interaction networks are controlled by the protein energy landscape [2–6].

The antibody variable regions are necessarily flexible to enable recognition of diversified targets. Recognition is associated with structural transitions [7–9]. The variable domains, especially CDRs, mainly control the specificity and affinity [10], while the constant domains modulate the isotype/effector [11] and independently the variable and constant domains functions. Recent studies indicate that besides the variable domains, the constant domain also plays an essential role in antigen binding [12–16]. There is direct communication between the variable domains of the light and heavy chains [17] and distant communication between the variable

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_11,

and constant domains [18]. Redistribution of flexibility in stabilizing fragments of mutant antibodies was observed in anti-lymphotoxin-beta receptor antibody [9]. Antibody-antigen recognition appears to also involve allosteric effects [19]. Pritsch et al. suggested that antibodies with identical variable domains, but different isotypes have significantly different affinities when binding to tubulin [13]. Oda et al. showed that the binding of antigen causes conformational changes in protein G and protein A binding sites on the heavy-chain constant domains [20]. A recent study surveyed over 100 crystal structures of antibodies in either the apo or bound form and found that distant loops from CH1-1 undergo significant fluctuation upon antigen binding and this fluctuation is common among these structures [21].

Though several studies have shown that antibody constant domains respond to antigen binding, molecular details and dynamics are still unknown. We have studied conformational changes in the antibody and cross-talk between its subunits and with antigens, using MD simulations of the complex of Fab and prion-associated peptide in the apo and bound forms. This allowed us to show that the inter-chain disulfide bond between the CH-1 and CL domains restrains the conformational changes of Fab, especially the loops in the CH1 domain, resulting in inhibition of the cross-talk between Fab subdomains which thereby may prevent prion peptide binding. Structural cross-talks between the constant domains and the antigen were shown by several negative and positive correlations of motions between the peptide and Fab constant domains. The cross-talk was influenced by the inter-chain disulfide bond which reduced the number of paths between them. Importantly, network analysis of the complex and its bound water molecules observed that those water molecules form an integral part of the Fab/peptide network of potential allosteric pathways. This chapter aims to provide robust and general strategies to study allosteric effects in the antibody-antigen recognition, which may help to develop strategies to incorporate these network communications—including the associated water molecules—in antibody design.

2 Simulation Protocols

2.1 Materials

2.1.1 Structures of Isolated Antibody and Prion Peptide

1. The structures of the apo forms of the Fab are based on crystal structures PDB IDs 1cr9 [22]. The unbound form was obtained by manually removing the peptide in the bound (1cu4) structure (*see* **Note 1**, Fig. 1).
2. The isolated prion peptide was simulated independently starting from the conformation in 1cu4.

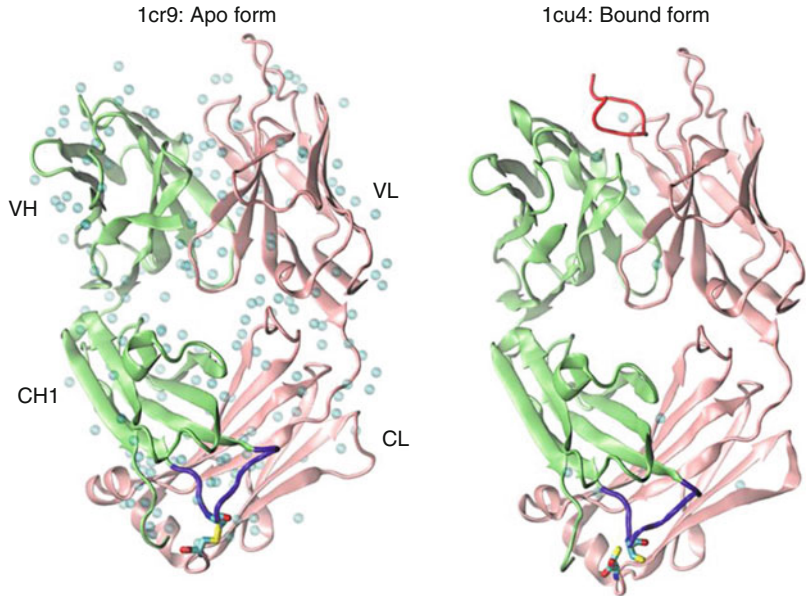

Fig. 1 Crystal structure of Fab 3F4 in Apo form (PDB:1cr9) and the complex (PDB:1cu4) with its cognate peptide (SHaPrP104-113). Light chain, heavy chain, the cognate peptide, CH1-1 loop, the inter-chain disulfide bond and the water molecules in the crystal are colored in pink, lime, red, purple, yellow, and cyan respectively

3. A 500-step steepest descent (SD) minimization with backbone atoms fixed is used to relax the initial structures of antibody and prion peptide.

2.1.2 Structures of Antibody-Antigen Complexes

1. The structures of the bound forms of the Fab/peptide complex are based on crystal structures PDB IDs 1cu4 (*see* **Note 1**, Fig. 1).
2. The two unresolved N-terminal residues of the heavy chain in the bound form were modeled using MODELLER.
3. The N- and C-termini are capped by NH_3+ and COO- groups. The tautomeric state of HIS residues is assigned based on local environment.
4. A 500-step SD minimization with backbone atoms and key hydrogen bonds/salt bridges fixed is performed to refine the overall structure (*see* **Note 2**, Fig. 1).

2.1.3 Setup of Disulfide Bond, Water Molecules, and Numbering System

1. The inter-chain disulfide bond was either kept or removed for the bound, unbound, and apo structures to consider the effects of the inter-domain disulfide bond.
2. Crystallized water molecules in the crystal structures were kept (*see* **Note 3**, Fig. 1).

3. As the non-sequential Kabat numbering scheme is used in the crystal structures, we renumber the residues for convenience in the simulation.

2.2 Simulation Methods

2.2.1 All-Atom MD Simulations

1. The systems were then solvated by TIP3P water molecules in the cubic water box with minimal margin of 15 Å from any protein atom to any edge of water box, and sodium and chlorides were added to neutralize the system to a total concentration of ~150 mM.
2. The resulting solvated systems were energy minimized for 5000 conjugate gradient steps, with the protein fixed and water molecules and counterions allowed to move, followed by additional 5000 conjugate gradient steps, where all atoms could move.
3. In the equilibration stage, each system was gradually relaxed by performing a series of dynamic cycles, in which the harmonic restraints on proteins were gradually removed to optimize the protein-water interactions.
4. In the production stage, all simulations were performed using the NPT ensemble at 310 K.
5. All MD simulations were performed using the NAMD software [23] with CHARMM36 force field [24]. MD trajectories were saved by every 2 ps for analysis.
6. To reduce the statistical noise, the systems of 1cu4, 1cu4 with the inter-domain disulfide bond, 1cr9, and 1cr9 without the inter-domain disulfide were repeated independently. The initial structure of each individual repeat system was minimized by using different energy minimization protocols. The systems were then re-solvated randomly by water molecules and ions. The initial velocity distribution of each repeat system was also set differently. Thus, the repeat simulations were started from alternate conformations.

2.2.2 Binding Energy Evaluation

1. To evaluate the binding energy between Fab and the prion peptide, the trajectory for each bound and apo system was extracted from the last 20 ns of explicit solvent MD without water molecules and ions.
2. The solvation energies of all systems were calculated using the generalized Born method with molecular volume (GBMV) after 500 steps of energy minimization to relax the local geometries caused by the thermal fluctuations which occurred in the MD simulations.
3. In the GBMV calculation, the dielectric constant of water is set to 80 and no distance cutoff is used. The binding energy between two Fab and the prion peptide was calculated by $\langle E_{\text{bind}} \rangle = \langle E_{\text{complex}} \rangle - \langle E_{\text{Fab}} \rangle - \langle E_{\text{peptide}} \rangle$.

2.2.3 Evolutionary Analysis

1. The sequence of Fab (PDB ID:1cu4) was used to generate relative conservation scores for each amino acid position using the Bayesian method implemented in the ConSurf server [25].
2. The sequence homologues of Fab were identified after three iterations of CS-BLAST algorithm against UNIREF90 database.
3. The selected sequence homologues with an *E*-value <0.0001 were then aligned using MAFFT [26].
4. The resulting Multiple Sequence Alignment (500 sequences with $\geq$35% similarity) was used to calculate the conservation scores, which ranged from 1 (variable) to 9 (highly conserved). The Fab residues were color-coded based on conservation scores obtained.

2.2.4 Correlation Analysis

1. Correlations between all the residues in the six systems were analyzed for the entire 100-ns MD trajectory (25,000 frames) using the normalized covariance to characterize the correlation in motion of protein residues [27], ranging from -1 to 1.
2. If two residues move in the same (opposite) direction in most the frames, the motion is considered as (anti-)correlated, and the correlation value is close to -1 or 1. If the correlation value between two residues is close to zero, they are generally uncorrelated. The correlation evaluation was performed by using CARMA [28] (*see* **Notes 4–6**, Fig. 2).

2.2.5 Weighted Network, Community Analysis, Optimal/Suboptimal Paths in Fab/Peptide Systems

1. A network is defined as a set of nodes with connecting edges. The nodes in this work represent the amino acid residues and the essential bound water molecules. An edge between two nodes was defined if any heavy atoms from the two residues/water molecules are within 4.5 Å of each other in over 75% of the analyzed frames (*see* **Note 7**, Fig. 3).
2. Neighboring residues in sequence are not considered to be in contact because they will form numerous trivial suboptimal paths in the weighted network.
3. The dynamical networks were constructed based on the 100-ns trajectories.
4. To study the water effect on the network, the crystallized water molecules on the antibody-antigen interface were kept from the system.
5. The communities within a network were defined as the sub-structures of the network in which the nodes are more heavily interconnected to each other than to other nodes. The community was identified by the Girvan-Newman algorithm [29]. For simplicity and clarity, communities with

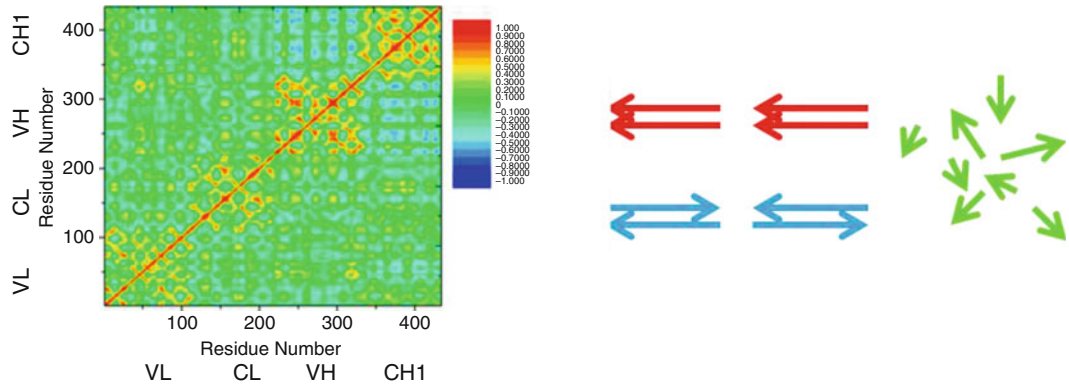

Fig. 2 Correlation analysis (*Cij*) of the motion during a 100-ns MD simulation of the Fab/peptide complex. Residues with highly correlated motion to loosely correlated motion, and to highly anti-correlated motion are colored from red, green/yellow, to blue

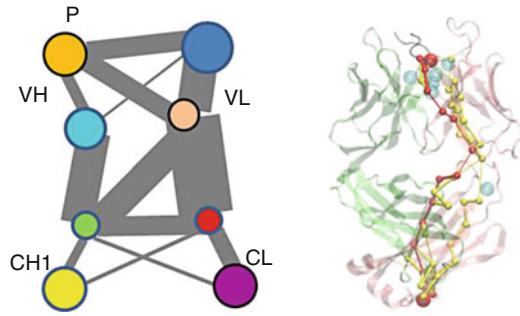

Fig. 3 Community and network analysis of the allosteric effects. The communities (left panel) are represented in circles in different colors and the size of the communities is proportional to the diameter. The grey lines represented the communications between different communities and the communication strength (betweeness) is proportional to width of the lines. *P* the prion peptide, *VL* light chain variable domain, *CL* light chain constant domain, *VH* heavy chain variable domain, *CH1* heavy chain constant domain-1. Light chain, heavy chain, and the peptide are colored by pink, lime and black respectively. Optimal and suboptimal pathways (right panel) are colored by red and yellow respectively. Key antibody/antigen residues, and the associated water molecules are represented by beads

member number <10 and edges with betweenness <1000 were not considered.

6. The length of a path between two distant nodes is the sum of the edges weights between consecutive nodes along the path. The shortest path is obtained by optimization of this length.
7. Suboptimal paths are determined in addition to the shortest path to measure the path degeneracy of the network.
8. The network, community, and optimal/sub-optimal paths analysis of Fab/peptide systems were performed by using NetworkView [30] module in VMD.

2.3 Simulation Notes

1. Missing residues, especially the residues on the CH1-1 loop, are important for allosteric effects. Thus, these missing residues should be reconstructed.
2. The initially constructed structures may contain some unreasonable steric overlaps between sidechains and backbones. The SD minimization with CHARMM36 force field is required to eliminate any bad atom contact. To minimize the structural disruption of the complex, especially the residues on the antibody-antigen interface, the hydrogen bonds and salt bridges are harmoniously constrained during the SD minimization.
3. The crystallized interfacial water molecules are crucial for the antibody-antigen recognition. When constructing the systems, these crystallized water molecules should not be excluded.
4. For cluster analysis, RMSD and Rg data for each structure in the equilibrium trajectories are calculated. The structures with the RMSD and Rg difference less than 4 Å are assigned as the same cluster.
5. Correlation analysis (*Cij*) of the motion required both simulation conformations and initial crystal conformations, thus the more conformations considered, the more accurate of the motion correlation analysis.
6. During the network analysis, the oxygen of the water molecules should be renamed to CA in order to be recognized by "network view" program.
7. The sub-optimal path (default is 20) evaluation should be set up carefully. In some cases when the complex is large, the sub-optimal path should set up <20, or there will be unexpected long time to evaluate all the possible sub-optimal path.

Acknowledgements

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract number HHSN261200800001E. This research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. JZ was supported in part by the Intramural Research Program of the NIH, NIDCD.

References

1. Wei G, Xi W, Nussinov R, Ma B (2016) Protein ensembles: how does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell. Chem Rev 116(11):6516–6551. https://doi.org/10.1021/acs.chemrev.5b00562
2. Gunasekaran K, Ma B, Nussinov R (2004) Is allostery an intrinsic property of all dynamic proteins? Proteins 57(3):433–443
3. Hu Z, Bowen D, Southerland WM, del Sol A, Pan Y, Nussinov R, Ma B (2007) Ligand binding and circular permutation modify residue

interaction network in DHFR. PLoS Comput Biol 3(6):e117
4. Ma B, Nussinov R (2014) Druggable orthosteric and allosteric hot spots to target protein-protein interactions. Curr Pharm Des 20 (8):1293–1301
5. Zhan C, Qi R, Wei G, Guven-Maiorov E, Nussinov R, Ma B (2016) Conformational dynamics of cancer-associated MyD88-TIR domain mutant L252P (L265P) allosterically tilts the landscape toward homo-dimerization. Protein Eng Des Sel 29(9):347–354. https://doi.org/10.1093/protein/gzw033
6. Ma B, Nussinov R (2016) Protein dynamics: conformational footprints. Nat Chem Biol 12 (11):890–891. https://doi.org/10.1038/nchembio.2212
7. Keskin O (2007) Binding induced conformational changes of proteins correlate with their intrinsic fluctuations: a case study of antibodies. BMC Struct Biol 7:31. https://doi.org/10.1186/1472-6807-7-31
8. Thielges MC, Zimmermann Jr YW, Oda M, Romesberg FE (2008) Exploring the energy landscape of antibody– antigen complexes: protein dynamics, flexibility, and molecular recognition. Biochemistry 47(27):7237–7247
9. Li T, Tracka MB, Uddin S, Casas-Finet J, Jacobs DJ, Livesay DR (2014) Redistribution of flexibility in stabilizing antibody fragment mutants follows Le Chatelier's principle. PLoS One 9(3):e92870. https://doi.org/10.1371/journal.pone.0092870
10. Mian IS, Bradwell AR, Olson AJ (1991) Structure, function and properties of antibody binding sites. J Mol Biol 217(1):133–151
11. Torres M, Casadevall A (2008) The immunoglobulin constant region contributes to affinity and specificity. Trends Immunol 29(2):91–97
12. Adachi M, Kurihara Y, Nojima H, Takeda-Shitaka M, Kamiya K, Umeyama H (2003) Interaction between the antigen and antibody is controlled by the constant domains: Normal mode dynamics of the HEL–HyHEL-10 complex. Protein Sci 12(10):2125–2131
13. Pritsch O, Hudry-Clergeon G, Buckle M, Pétillot Y, Bouvet JP, Gagnon J, Dighiero G (1996) Can immunoglobulin C (H) 1 constant region domain modulate antigen binding affinity of antibodies? J Clin Investig 98(10):2235
14. Dam TK, Torres M, Brewer CF, Casadevall A (2008) Isothermal titration calorimetry reveals differential binding thermodynamics of variable region-identical antibodies differing in constant region for a univalent ligand. J Biol Chem 283(46):31366–31370
15. Tudor D, Yu H, Maupetit J, Drillet A-S, Bouceba T, Schwartz-Cornil I, Lopalco L, Tuffery P, Bomsel M (2012) Isotype modulates epitope specificity, affinity, and antiviral activities of anti–HIV-1 human broadly neutralizing 2F5 antibody. Proc Natl Acad Sci 109 (31):12680–12685
16. Li T, Tracka MB, Uddin S, Casas-Finet J, Jacobs DJ, Livesay DR (2015) Rigidity emerges during antibody evolution in three distinct antibody systems: evidence from QSFR analysis of fab fragments. PLoS Comput Biol 11(7):e1004327. https://doi.org/10.1371/journal.pcbi.1004327
17. Pellequer JL, Chen SW, Roberts VA, Tainer JA, Getzoff ED (1999) Unraveling the effect of changes in conformation and compactness at the antibody VL-VH interface upon antigen binding. J Mol Recognit 12(4):267–275
18. Janda A, Casadevall A (2010) Circular Dichroism reveals evidence of coupling between immunoglobulin constant and variable region secondary structure. Mol Immunol 47 (7):1421–1425
19. Sela-Culang I, Kunik V, Ofran Y (2015) The structural basis of antibody-antigen recognition. Immune system modeling and analysis. Front Immunol. https://doi.org/10.3389/fimmu.2013.00302
20. Oda M, Kozono H, Morii H, Azuma T (2003) Evidence of allosteric conformational changes in the antibody constant region upon antigen binding. Int Immunol 15(3):417–426
21. Sela-Culang I, Alon S, Ofran Y (2012) A systematic comparison of free and bound antibodies reveals binding-related conformational changes. J Immunol 189(10):4890–4899
22. Kanyo ZF, Pan K-M, Williamson RA, Burton DR, Prusiner SB, Fletterick RJ, Cohen FE (1999) Antibody binding defines a structure for an epitope that participates in the PrP C→ PrP Sc conformational change. J Mol Biol 293 (4):855–863
23. Kale L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K (1999) NAMD2: greater scalability for parallel molecular dynamics. J Comput Phys 151(1):283–312
24. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and

dynamics studies of proteins. J Phys Chem B 102(18):3586–3616

25. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33(suppl 2):W299–W302
26. Katoh K, Misawa K, Ki K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30(14):3059–3066
27. Ichiye T, Karplus M (1991) Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. Proteins 11 (3):205–217
28. Glykos NM (2006) Software news and updates carma: a molecular dynamics analysis program. J Comput Chem 27(14):1765–1768
29. Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826
30. Eargle J, Luthey-Schulten Z (2012) NetworkView: 3D display and analysis of protein· RNA interaction networks. Bioinformatics 28 (22):3000–3001

Chapter 12

Distal Regions Regulate Dihydrofolate Reductase-Ligand Interactions

Melanie Goldstein and Nina M. Goodey

Abstract

Protein motions play a fundamental role in enzyme catalysis and ligand binding. The relationship between protein motion and function has been extensively investigated in the model enzyme dihydrofolate reductase (DHFR). DHFR is an essential enzyme that catalyzes the reduction of dihydrofolate to tetrahydrofolate. Numerous experimental and computational methods have been used to probe the motions of DHFR through the catalytic cycle and to investigate the effect of distal mutations on DHFR motions and ligand binding. These experimental investigations have pushed forward the study of protein motions and their role in protein-ligand interactions. The introduction of mutations distal to the active site has been shown to have profound effects on ligand binding, hydride transfer rates and catalytic efficacy and these changes are captured by enzyme kinetics measurements. Distal mutations have been shown to exert their effects through a network of correlated amino acids and these effects have been investigated by NMR, protein dynamics, and analysis of coupled amino acids. The experimental methods and the findings that are reviewed here have broad implications for our understanding of enzyme mechanisms, ligand binding and for the future design and discovery of enzyme inhibitors.

Key words Allostery, Protein motions, Dihydrofolate reductase, Point mutation

1 Introduction

The interaction between a ligand and its biomolecular target forms the basis for many physiological processes [1, 2]. Understanding the molecular mechanisms by which ligands recognize and bind their targets remains a fundamental question in biochemistry and biophysics. Over the years, several models have been proposed. Early on, the "lock-and-key" model, where a protein and ligand are a perfect match of rigid complementary structures, dominated (Fig. 1) [3, 4]. In this model, the protein's ligand binding site is assumed to have a single, rigid shape and the ligand distinguishes between different proteins in the cell based on the different shapes of their ligand binding sites [4]. As technology has advanced, so has our understanding of protein structure and dynamics, and it is now

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_12,

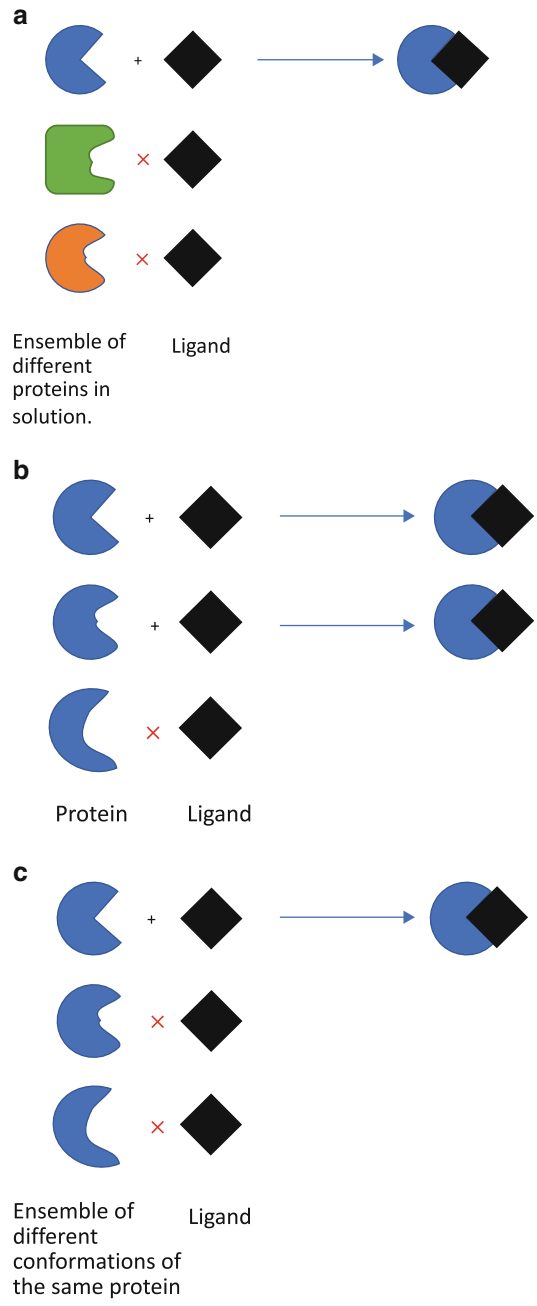

Fig. 1 Schematic representation of (**a**) "lock-and–key," (**b**) "induced fit," and (**c**) "conformational selection" models. In the "lock-and-key" model (**a**), binding occurs when there is an exact geometric fit between a protein and the ligand, among an ensemble of many different proteins. The cross sign indicates that the shapes of the protein and ligand are not a match. In the induced fit model (**b**), there is no exact fit between the protein and ligand before binding occurs. The ligand binds a protein molecule, inducing the protein shape to fit the ligand. In the conformational selection model (**c**), the ligand selects a conformer from the ensemble of conformers of the same protein, whose shape is complementary to the ligand

universally accepted that protein–ligand interactions are rarely rigid [5]. Proteins are inherently flexible molecules that adopt an ensemble of conformations in solution, which exist in a state of dynamic equilibrium. This realization led to the formation of the "induced fit" and, subsequently, the "conformational selection" models of protein–ligand and enzyme–substrate interactions [1–17]. These models account for conformational changes in ligand binding [9].

The induced fit and conformational selection models account for the dynamic state of a protein, but differ in when in the binding process the conformational change occurs [9]. In the induced fit model, the ligand binds to the predominant, free conformation in solution followed by a conformational change to the preferred ligand-bound conformation (Fig. 1). The conformational selection model proposes that a given protein exists in a state of dynamic equilibrium between several conformations, termed the conformational ensemble and the "ligand-bound" conformation already exists as part of the conformational ensemble in solution in a low population state. The population state of a conformation refers to the amount of protein in a particular conformation. Therefore, the ligand recognizes and selectively binds to the conformer in the favored state, shifting the conformational equilibrium to make it the predominant conformation in the ensemble [10]. Conformational selection may appear similar to the "lock-and-key" model because selection occurs via a match in the "shape" in both models. However, in conformational selection, the selection is of a conformer out of many different conformers of a single protein rather than selection of a protein out of many different proteins, as in the lock-and-key model (Fig. 1). In the conformational selection model, ligand binding induces a change in the equilibrium of the states, which forces the system to re-equilibrate, shifting the population of the conformational ensemble toward the preferred conformer. The population shift described in the conformational selection model cannot be present in the lock-and-key model because the ensemble is composed of different proteins rather than different conformers [4]. Thus, the key difference between these two models is the presence of a dynamic equilibrium which allows for a population shift to occur upon ligand binding [16].

Conformational motions in enzymes are inherently linked to their function and have a direct impact on binding of substrate or cofactor, product release, and allosteric regulation [12, 15, 18, 19]. Enzymes are common therapeutic targets but the movement between different conformational states is often ignored during drug discovery and design [12]. When the flexibility of a protein is acknowledged during drug design, it is often assumed to follow the induced fit model. The possibility that small molecule or drug binding may occur through the conformational selection mechanism is usually ignored, which may hinder drug discovery and development efforts [20]. Understanding the mechanism by

which a ligand recognizes its target is a crucial prerequisite for rationally designing novel and effective drugs and therapeutics [1]. By designing small molecules that target a specific conformation, it may be possible to design more potent or selective drugs that bind to the preferred conformer of the target enzyme. This would in turn shift the equilibrium toward this state, redistributing the conformational ensemble so that the favored state predominates in solution [4]. Therefore, the innate flexibility and conformational dynamics of proteins and enzymes may be exploited to improve the efficacy of rational drug design.

In this paper, we use the enzyme dihydrofolate reductase (DHFR) to illustrate how the conformational dynamics influence substrate, cofactor, and inhibitor binding. DHFR has become an important model system for investigating the link between protein dynamics and catalytic function for several reasons. As we discuss the many studies on the link between DHFR structure, motions, and catalysis, we provide an overview of the methods used in these studies. DHFR is a known drug target for inhibiting DNA synthesis in rapidly proliferating cancer cells and microbial infections [21, 22] and a wealth of enzymological studies that have focused on DHFR make it a unique system for investigating the role of dynamics in catalysis [18, 21].

2 Dihydrofolate Reductase

2.1 Kinetic Mechanism and Structure

DHFR kinetics have been extensively reviewed [18, 21, 23]. A notable feature of the catalytic cycle is that hydride transfer, the chemical step of the reaction, is not immediately followed by the release of product. Instead, following hydride transfer, the oxidized $NADP^+$ cofactor is released and NADPH rebinds before product release. Thus, free enzyme is not generated under physiological conditions and the enzyme remains "primed" for the next round of catalysis [18, 23]. The coordination of substrate binding and product release is maintained via a synergistic interaction between the substrate and cofactor in the binding site, in which the off-rates of $NADP^+$ increase in the presence of bound product and the off-rates of product (THF) increase in the presence of the reduced cofactor, NADPH [18, 24].

DHFR displays a high degree of structural homology in different species, despite low sequence homology [18, 25]. The first X-ray crystal structure of a DHFR was of the *Escherichia coli* enzyme (ecDHFR) and was published almost 40 years ago [26]. The ecDHFR structure has been extensively discussed elsewhere [26, 27]. The major subdomain is dominated by a set of three flexible loops that are located on the ligand binding face that surround the active site. The loops are designated Met20 (residues 9-24), F-G (residues 116-132), and G-H (residues 142-150).

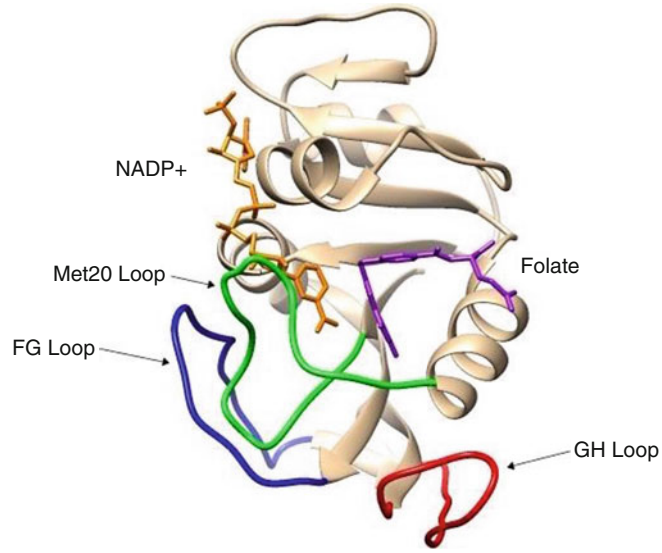

Fig. 2 Overview of 3D structure of *E. coli* DHFR (PDB entry 1RX2). The ternary complex with $NADP^+$ cofactor and folate substrate in their respective binding sites. The major subdomain loops are labeled. Met20 loop is illustrated in green. F-G loop is illustrated in blue. G-H loop is illustrated in red. $NADP^+$ and folate are displayed in orange and purple, respectively. (This figure was prepared using the program Chimera [28])

These loop regions serve as a gate, closing over the bound ligands in the ternary E:DHF:NADPH complex, but are mobile in the apoenzyme and holoenzyme [27]. The substrate and cofactor bind in a hydrophobic cleft at the juncture of the two subdomains and hinge bending motions about residues Lys38 and Val88 allow the adenosine binding domain to move relative to the major domain upon ligand binding, closing off the active site cleft to the surrounding environment (Fig. 2) [18, 27].

2.2 Active Site Structure and Loop Conformations

In the ternary complex, the pterin ring of the substrate and the nicotinamide ring of the cofactor bind in close proximity in the active site, with the hydride donor atom (C4 of NADPH) and the hydride acceptor (C6 of the pterin ring) in van der Waals contact [29]. The DHFR active site contains an invariant carboxylic acid residue, Asp27 in bacteria and a glutamic acid residue in vertebrates (Glu30 in humans) [30]. Mutational studies have supported that Asp27 has a crucial role in the hydride transfer step [30, 31].

The three flexible loops play a critical role in ecDHFR catalysis. The Met20 loop lies directly over the active site and shields the reactants from solvent. The F-G and G-H loops impart stability via hydrogen bonding interactions with the Met20 loop. From X-ray crystallographic data, it has been demonstrated that the Met20 loop assumes four characteristic conformations in the crystalline

state, designated the occluded, closed, open, and disordered states [21]. The occluded and closed conformations have also been observed in NMR experiments [32]. The open conformation has only been observed in certain crystal forms. The disordered states described cases in which motion renders the loop unclear or invisible in crystallographic experiments. The four loop conformations are best characterized by their secondary structure, interactions with nicotinamide-ribose moiety, and hydrogen bonding with the F-G and G-H loops. The open conformation displays characteristics between those of the closed and occluded conformations. The disordered conformation displays characteristics of a time-averaged exchange of closed and occluded conformations [21].

The conformation of the active site loops depends on the ligands bound in the substrate and cofactor binding sites. If only the substrate site is occupied, the enzyme adopts the occluded loop conformation. Binding of the nicotinamide-ribose moiety of NADPH within its binding site produces the closed conformation, in which the Met20 loop is packed against the nicotinamide ring of NADPH, closing the active site off to the surrounding solvent. Only the closed conformation allows for the proper positioning of the NADPH and substrate reactive centers, such that they are in close enough proximity to facilitate the reaction. Thus, it is apparent that movement of the Met20 loop is directly coordinated with the stages of the catalytic cycle (Fig. 3) [18, 33].

The occluded and closed conformations differ in structure in the central portion of the Met20 loop and in the pattern of hydrogen bonds formed between the Met20 loop and the F-G and G-H loops. In the occluded state, the central region of the Met20 loop forms a 3_{10}-helix, with residues Met16 and Glu17 projecting into the active site, where they "occlude" the binding site for the nicotinamide ring moiety of NADPH. The occluded conformation is stabilized by hydrogen bonding interactions between Asn23 (backbone CO and NH) in the Met20 loop and Ser148 (NH and Oγ) in the G-H loop. In the closed conformation, residues 16-19 form a β-hairpin structure. Met16 and Glu17 are flipped out of the active site, thereby allowing nicotinamide binding, while the side chains of Asn18 and Met20 pack down over the bound substrate and cofactor. The Asn23/Ser148 hydrogen bonds are disrupted, and new hydrogen bonds are formed between the backbone NH and Oδ of Asp122 in the F-G loop and the backbone CO and NH of Gly15 and Glu17, respectively [21].

Sawaya et al. proposed a detailed structural model for the conformational changes that occur during the ecDHFR catalytic cycle [21]. They based their model on the analysis of isomorphous crystal structures ($P2_12_12_1$) of different ecDHFR complexes to ensure that the packing interactions are constant between the different structures in the series and that conformational differences are due to differences in ligand binding. To investigate the

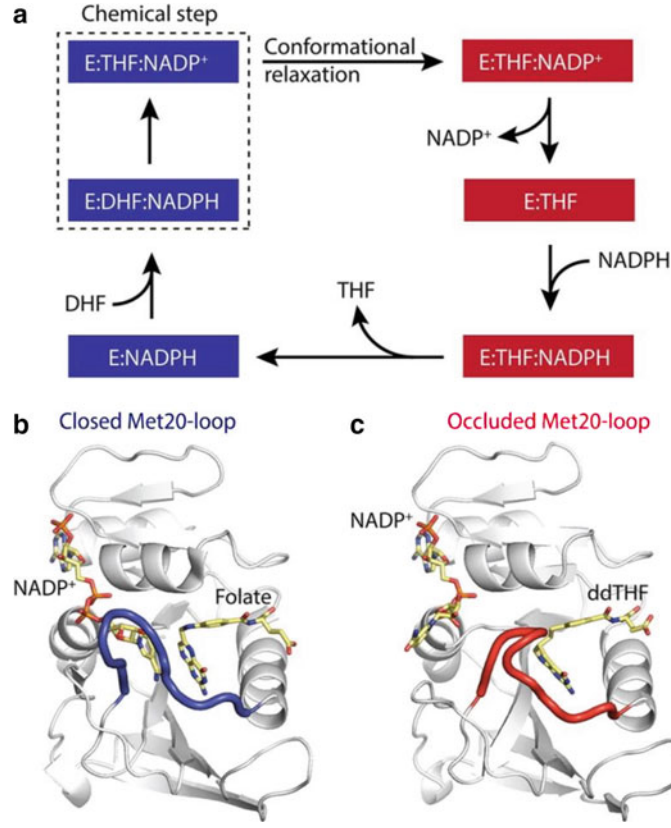

Fig. 3 Conformational changes during the catalytic cycle of ecDHFR. Blue indicates the conformation of the Met20 loop in the closed conformation. Red indicates that the Met20 loop is in the occluded conformation. The crystal structures illustrate the conformational change in the Met20 loop upon the hydride transfer reaction. (Figure reproduced from ref. 33)

conformational changes in different crystal packing environments, the authors also solved some of the ligand-bound structures using crystals with different space groups. The complexes were analogous to the five kinetic intermediates and to the transition state for the hydride transfer reaction. The authors used analogues because actual intermediates are transient and it takes several hours to collect data for an X-ray structure. The choice of analogue for such studies is significant. For example, to mimic the transition state, the authors used methotrexate. This compound has a unique binding geometry, which has been found to induce the transition state protein conformation even though methotrexate does not resemble a transition state structure. The structures suggest that the Met20 loop is in the closed conformation in the holoenzyme, the Michaelis complex, and the transition state and in the occluded conformation in the three product complexes. In the Michaelis complex and transition state, the nicotinamide-ribose moiety is

predicted to occupy its binding pocket within the active site, in close proximity to the pterin ring of the substrate. However, in the occluded D:THF:NADP+ and D:THF:NADPH product complexes, this moiety projects into the solvent. Movement of the adenosine binding domain relative to the major domain during the catalytic cycle modulates the width of the p-aminobenzoylglutamate (pABG) binding cleft. From the Michaelis to the transition state analogue complex, they observed that rotation between the two domains closes the pABG binding cleft by approximately 0.5 Å. The resulting enhancement of contacts with the pABG moiety may stabilize puckering of C6 of the pterin ring in the transition state. The domain rotation is further adjusted by cofactor induced movements of the α_B and α_C helices, producing a larger pABG cleft in the product complexes. The domain rotations are suggested to play a role in transition state stabilization and NADPH-assisted product release (Fig. 3) [18, 21]. In this work function was intimately linked to both macromolecular structure and dynamics by solving crystal structures at different stages of the catalytic cycle by using substrate analogues. The findings showed that regions distal from the active site play a role and have inspired experiments to investigate these regions further. Using the isomorphous structures and GIFMerge, the authors created a movie that illustrates DHFR's range of inferred subdomain and loop movements.

2.3 Species-Specific Structural Features of DHFR

2.3.1 Human DHFR and Comparison to Bacterial DHFR

The structural differences between vertebrate DHFRs and bacterial DHFRs are important to the specificity of DHFR inhibitors. Human DHFR (hDHFR) is a monomeric, 186-amino-acid protein with a molecular weight of approximately 21.5 kDa. Like ecDHFR, hDHFR has an eight-stranded β-sheet consisting of seven parallel strands and a carboxy-terminal antiparallel strand. Five α-helices are packed against the beta-sheet core, denoted αB, αC, αE, αE′, and αF. The αE′ helix is perpendicular to αE and may have emerged via a five-residue insertion mutation in human DHFR relative to bacterial DHFR [34]. In addition, hDHFR has one left-handed, type II polyproline-like helix, which is not present in ecDHFR. Another variation from ecDHFR is the presence of a cis-peptide linkage between residues Arg65 and Pro66. The other cis-peptide linkage is a conserved structural feature of all DHFRs and is between residues Gly116 and Gly117, which are located near the nicotinamide binding site. Just as in ecDHFR, the active site cleft in hDHFR is formed at the junction of the two domains. However, in the case of hDHFR the larger subdomain is the adenosine binding domain and the second is the smaller loop domain. The acidic residue, Glu30, is analogous to Asp27 in ecDHFR [34].

The conformation of vertebrate and human DHFRs is more rigid than ecDHFR [34]. However, the amino acids required for catalysis and the general secondary structural features, as well as the

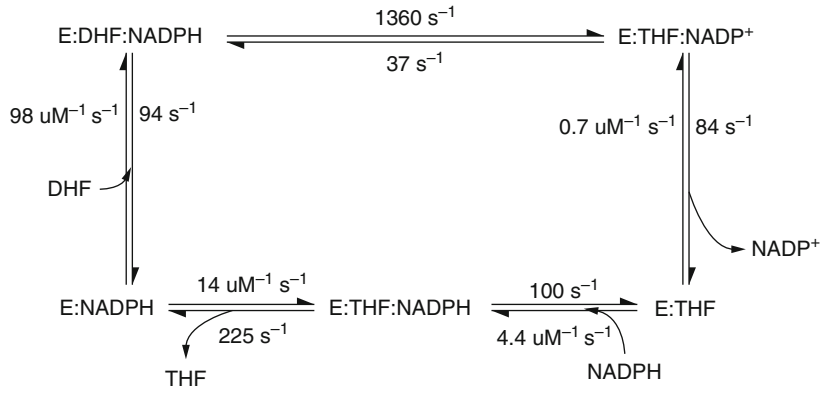

Fig. 4 Kinetic scheme of the catalytic cycle of human DHFR. The rate constants for each step are included for each of the five steps. The most notable difference in the kinetic scheme of hDHFR and ecDHFR is the rate of the hydride transfer step, about sixfold larger for hDHFR. (This figure was generated using the program ChemDraw [35])

kinetic pathways, are highly conserved (Figs. 4 and 5) [36, 37]. Comparison of hDHFR and ecDHFR reveals that hDHFR lacks the Met20 loop subdomain motion that is observed in ecDHFR. In the human DHFR enzyme, the Met20 loop remains in the closed conformation throughout the catalytic cycle. In contrast to ecDHFR, ligand binding in hDHFR occurs through a hinge opening motion of the adenosine subdomain relative to the loop subdomain (Fig. 5). Hinge 1 is defined as residues Thr39-Leu49 and hinge 2 as His127-Leu131. The E:NADPH complex exists as the "hinge-open" conformation. The substrate complexes adopt the "hinge-closed" conformation, in which the active site is tightly packed, thus favoring hydride transfer. The product complexes also exist in the hinge-closed conformation [36].

In addition to the lack of Met20 loop motions, another significant difference between the vertebrate and bacterial DHFRs is that vertebrate DHFRs are more rigid and conformationally restrained. There are three key structural differences between vertebrate and ecDHFR that give rise to the conformational rigidity of hDHFR: (a) insertion of the left-handed polyproline-type helix in vertebrate DHFRs loop1; (b) Gly20 in the loop 1 of vertebrate DHFR instead of an Asn, which creates a stable β-hairpin; and (c) the vertebrate's G-H loop is shorter relative to the ecDHFR G-H loop, preventing the formation of hydrogen bonds with the Met20 loop. These structural and conformational differences are important to designing therapeutics that target DHFR from a specific disease or pathogen specifically and to minimize the off-target effects of inhibitors [34].

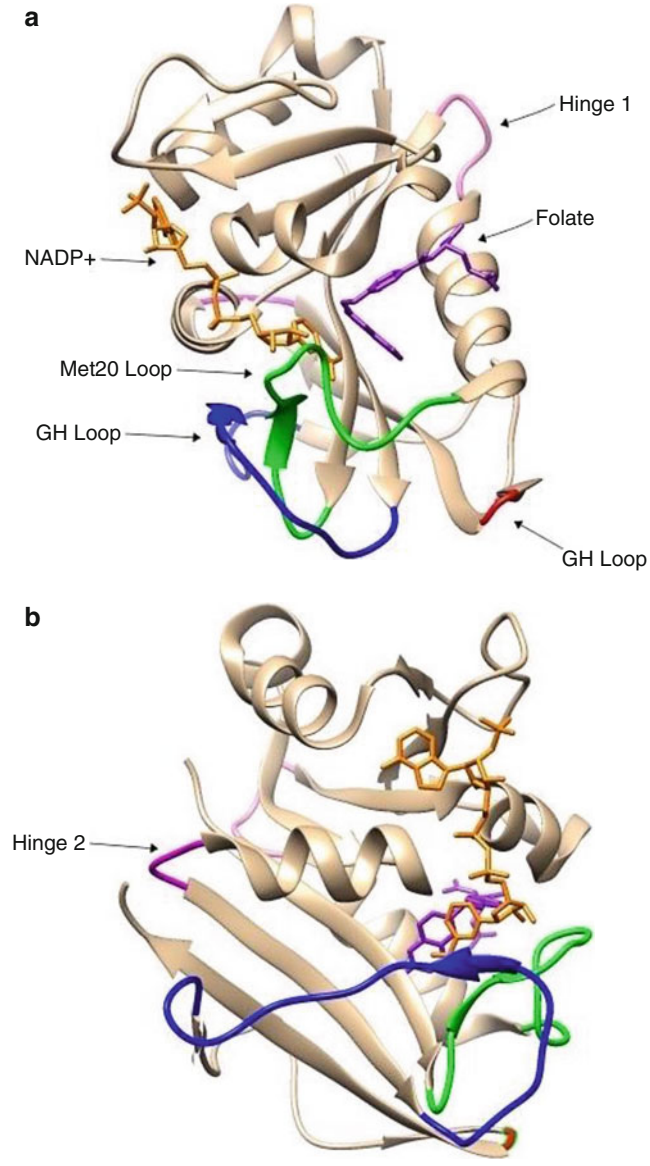

Fig. 5 (**a**) Structure of hDHFR complexed with $NADP^+$ and folate (PDB entry 4M6K). The Met20 loop (residues 12-27, displayed in green) remains closed throughout the entire hDHFR catalytic cycle. The FG loop (residues 139-159) is colored blue and the GH loop (residues 172-175) is colored red. $NADP^+$ and folate are displayed in orange and purple, respectively. Hinge 1 is illustrated in light pink. (**b**) View rotated 90° relative to panel A structure so that hinge 2 can be observed (magenta). In the hDHFR:NADPH complex (hinge-open complex), helix αF slides toward the active site, relative to the hDHFR:DHF:NADPH, hDHFR:THF:$NADP^+$, and hDHFR:NADPH complexes (hinge-closed complexes). (This figure was prepared using the program Chimera [28])

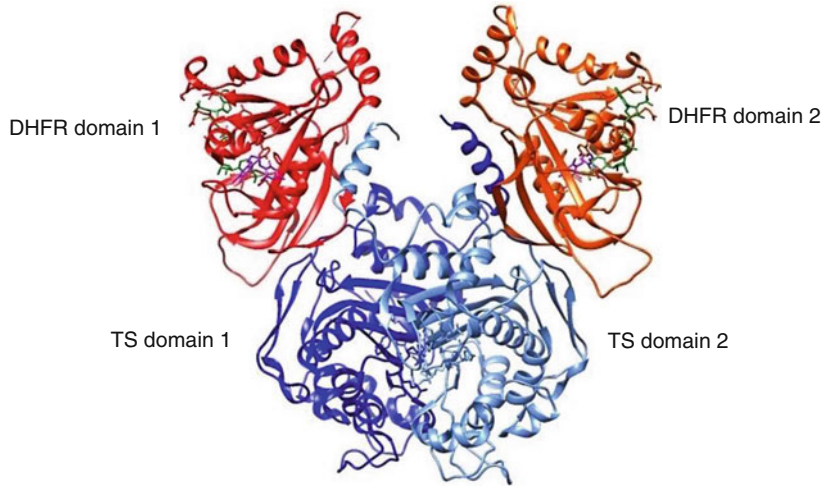

Fig. 6 Three-dimensional structure of DHFR-TS from the malaria parasite *Plasmodium falciparum* (PDB entry 1J3I). The DHFR domains are shown in red and orange. The TS domains are shown in light and dark blue. The DHFR domains are bound to NADPH (green) and the inhibitor molecule WR99210 (purple). (This figure was prepared using the program Chimera [28])

2.3.2 Plasmodial DHFR-TS

A major difference between protozoal DHFR, such as *Plasmodium falciparum* (pfDHFR) and *Plasmodium vivax* (pvDHFR), and the DHFRs from other species is that the *Plasmodium* enzyme exists as a bifunctional enzyme called dihydrofolate reductase-thymidylate synthase (DHFR-TS) in which DHFR and TS are two domains of a single homodimeric protein (Fig. 6). The two subunits are associated via extensive contact between the two TS domains [38, 39]. In humans and bacteria, DHFR and TS exist as two separate monofunctional proteins [38]. Each polypeptide of the pfDHFR-TS homodimer is comprised of 608 amino acids, the first 231 residues of which comprise the DHFR domain of the polypeptide [40]. The next portion of the sequence is the 89-residue "junction region" connecting the DHFR domain to the TS domain. The remaining 288 residues constitute the TS domain of the polypeptide [40]. The key residues in the active site of the pfDHFR domain are Ile14, Ala16, Trp48, Asp54, Phe58, Ser108, Ile164, and Thr185, which interact with DHF, NADPH, and/or inhibitors [41]. The DHFR domain has some similarities to other DHFRs in that it is comprised of eight central β-strands (β_A-β_H) and four α-helices (α_B, α_C, α_E, and α_F). In addition to these structural features, there are three short α-helices in pfDHFR, which are designated α_A, α_D, and $\alpha_{D'}$. Each monomer of pfDHFR-TS contains two inserts in the DHFR domain: Insert 1 which contains a short 3_{10}-helix α_{i1} (residues 33-36), and Insert 2 which contains a long helix α_{i2} (residues 67-95) [40]. While Insert 1 extends away from the domain surface and does not interfere with the core DHFR subunit structure, part of the moiety

interacts with the TS domain and aids in stabilizing the interdomain attachment. The junction region contains the a_{j1} helix which links the DHFR and TS domains and interacts with the DHFR domain of the other polypeptide chain [42]. These features and the major differences in amino acid sequence, structure, and function between *Plasmodial* DHFR-TS and bacterial and vertebrate DHFRs have allowed for the development of species-specific inhibitors.

3 Ligand-Dependent Conformational Dynamics during DHFR Catalysis

The first indication that conformational dynamics may play a role in ecDHFR catalysis came from kinetic measurements, which showed that the apoenzyme exists as two isoforms, E1 and E2 [43–45]. In these experiments, the authors made stopped-flow measurements of association and dissociation rates of ligands to and from the enzyme. They mixed together solutions of the enzyme and ligand and recorded the resulting tryptophan fluorescence over time using the excitation wavelength of 290 nm and a 341 nm interference filter. Alternatively, the authors excited the tryptophans and recorded the enhancement of coenzyme (NADPH) fluorescence by energy transfer using an excitation wavelength of 290 nm and a 449 nm interference filter. The data was analyzed to obtain the rates for formation and dissociation of binary complexes of both forms of the enzyme. The second isoform was detected as a slow ligand-independent phase that followed an initial ligand-dependent burst phase when mixed with NADPH and substrate or inhibitor in a stopped-flow fluorescence experiment by Dunn et al. [43–45] They found that NADPH appears to bind rapidly and exclusively to the E1 isoform and that binding of a ligand or inhibitor to either the substrate or cofactor sites resolves the ambiguity between conformational states and allows a single conformation that readily binds further ligands to form ternary complexes to be observed [18]. This observation highlights the connection between protein conformation and ligand binding.

In a more recent study, Reddish et al. used tryptophan fluorescence probed temperature-jump spectroscopy to observe the kinetics of ligand binding and ligand-induced conformational changes of three DHFR complexes to attempt to establish the relationship between conformational changes and catalytic steps along the DHFR pathway [46]. Temperature-jump spectroscopy can be used to measure rapid reaction rates in the microsecond timescale, which is a significant timescale for catalysis and allostery. In this method, temperature is rapidly increased perturbing the system. The fluorescence of the system is then observed as the system reaches equilibrium with a new equilibrium constant. The three

complexes examined were DHFR:folate, DHFR:$NADP^+$, and DHFR:$NADP^+$:folate, which are models for the binary product complex, the holoenzyme, and the Michaelis complex, respectively. They observed two kinetic events in the temperature-jump transients of each of the complexes: a fast relaxation event and a slow relaxation event. The slow off-pathway conformational rearrangement observed can be interpreted as evidence for conformational selection as a mechanism for ligand binding. The millisecond conformational rearrangements observed using Trp fluorescence are not coupled to a binding event as would be expected for an induced fit model of ligand binding. The presence of the slow relaxation event regardless of the ligand state suggests that it corresponds to fluctuations of the protein that would be necessary for the conformational selection process. In addition, the dependence of the rate of the slow event on the ligand identity is consistent with a ligand-dependent population shift to a favored conformation, as is expected in the conformational selection model [46].

4 Drug Binding

The role of enzyme dynamics in the binding of three DHFR inhibitors will be discussed here. Methotrexate (MTX), trimethoprim (TMP), and pyrimethamine (PYR) were the first DHFR inhibitors in clinical use (Fig. 7) [47]. While MTX is a potent inhibitor of essentially all DHFRs (most likely due to its similarity to the natural substrate), TMP and PYR show strong selectivity for bacterial and protozoal enzymes, respectively [48]. For example, TMP shows 14- and 6000-fold selectivity for ecDHFR over *P. berghdi* and rat liver DHFR, respectively. In contrast, PYR displays 1400- and 5000-fold selectivity for *P. berghdi* over rat liver DHFR and ecDHFR, respectively [48]. Studying the differences in binding of these small molecules to DHFR from different species and the structural features that drive selectivity can provide insight into the role of conformational dynamics in inhibitor binding. Furthermore, understanding the role of conformational selection in the binding of MTX, TMP, and PYR will guide development of novel DHFR inhibitors with improved potency and selectivity.

4.1 Methotrexate

Methotrexate (MTX) was first introduced as an effective treatment for acute leukemia in 1950 and for other solid tumors during the 1950s and 1960s [49]. It is a folic acid analogue and an extremely potent competitive inhibitor of all DHFRs, including the human enzyme. Rapidly proliferating cancer cells need a continuous supply of THF for nucleic acid synthesis and replication. Inhibition of the folate pathway eventually results in cell death. In addition to cancer chemotherapy, MTX has been successfully used to treat rheumatoid arthritis, juvenile idiopathic arthritis, uveitis, graft vs. host disease

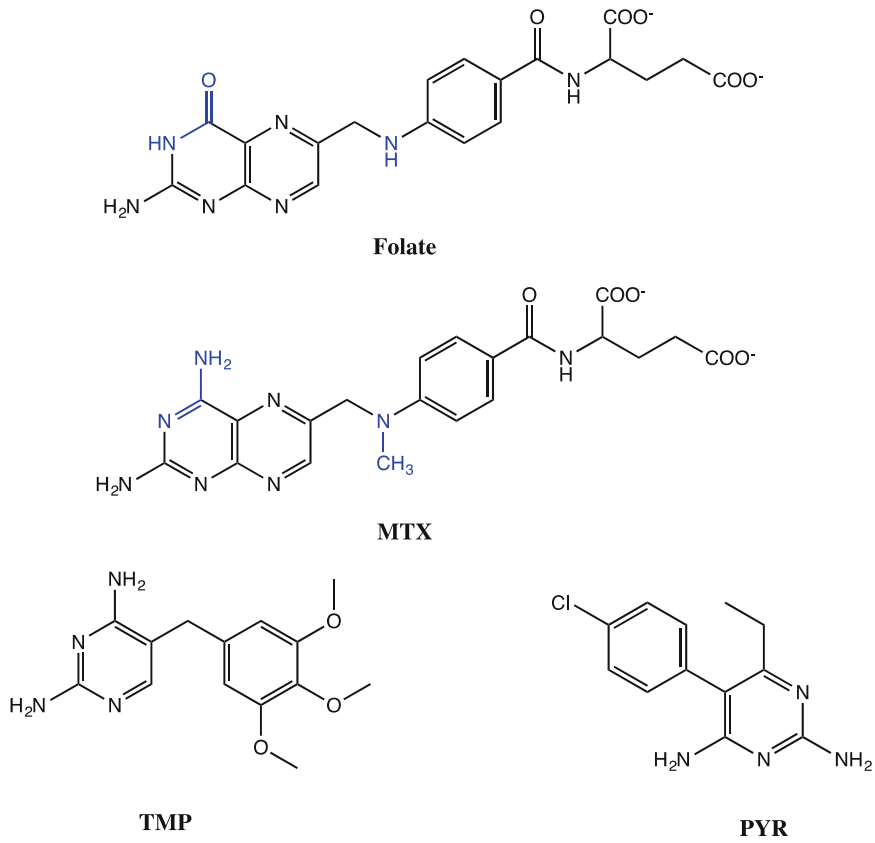

Fig. 7 Structures of folate and the antifolate drugs methotrexate, trimethoprim, and pyrimethamine. Methotrexate is a structural analogue of the natural substrate for DHFR, folate. The differences in the structures are highlighted in blue. MTX is a potent inhibitor of human DHFR and used to treat cancer and some other conditions, including rheumatoid arthritis. TMP is a potent inhibitor of bacterial DHFRs and is therefore used as an antibacterial agent. PYR is an inhibitor of certain protozoal DHFRs, including *Plasmodium falciparum* DHFR-TS and is used as an anti-malarial drug. (This figure was generated using the program ChemDraw [35])

(GVHD), psoriasis, Crohn's disease, and other inflammatory conditions [48–51].

The inhibition of DHFR by MTX has been extensively studied. It has been shown via two-dimensional ^{1}H NMR methods that the conformation of bound MTX to ecDHFR has the pteridine ring rotated about the C6-C9 bond by about 180° relative to that of bound DHF [52]. In this orientation, hydrogen bonds are formed between the protein and the 2,4-diamino groups on the pyrimidine ring portion of the molecule. Crystal structures also indicate that a hydrogen bond between a protonated N1 and Asp27 is present. This interaction is not present in the DHF complex. The enhanced affinity of MTX over DHF has been attributed to this charged interaction [52–54]. MTX also binds hDHFR in the same "non-productive" orientation (Fig. 8) [56, 57].

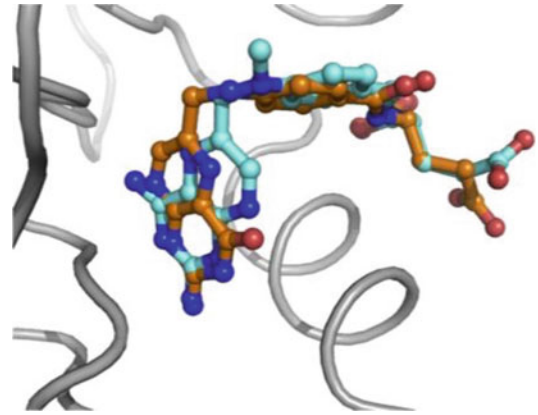

Fig. 8 Comparison of the binding orientation of MTX and DHF. MTX is shown in cyan and DHF in orange. MTX and DHF do not bind DHFR in the same orientation despite their highly similar structures. (Figure reproduced from ref. 55)

In addition, the conformation of MTX-bound DHFR has been studied in numerous species. Using two-dimensional ^{1}H NMR methods, Falzone et al. found that the binary ecDHFR:MTX complex exists as two slowly interconverting conformers and that this interconversion does not take place in the DHF binary complex [52]. Specifically, the authors observed that many of the resonances in the DHFR-methotrexate complex were broadened or doubled. The two distinct set of resonances were attributed to the presence of two protein isomers. The authors collected NOESY spectra at 303 and 323 K and saw no interconversion between the isomers at the lower temperature but did see exchange cross-peaks in a 700 ms NOESY spectrum at 323 K, indicating interconversion between the two isomers at the NMR timescale. These observations were supported by other crystallographic and NMR experiments [58]. It was also observed that two conformers of DHFR exist in solution and interconvert slowly prior to methotrexate binding. The existence of a conformational equilibrium prior to ligand binding may be evidence to support the conformational selection model. In addition, the population of each of the conformers of the MTX complex appears to be pH dependent, with conformation 2 being preferred at high pH [58]. Interestingly, the opposite situation appears to exist for the *Lactobacillus casei* enzyme (lcDHFR). In contrast to ecDHFR, lcDHFR exists in at least two conformations in its binary complex with DHF and three conformations in its ternary complex with DHF and $NADP^{+}$, but only exists as a single conformation in the MTX complex [59–61]. The different conformational equilibriums among bacterial species provide insight into structural differences between the complexes and such information may eventually assist in the design of more species-specific inhibitors [59].

4.2 Trimethoprim

Trimethoprim is a widely used antibacterial drug that is often used in combination with the antibiotic sulfamethoxazole to treat numerous bacterial infections including *E. coli*, *Staphylococcus aureus*, *Shigella* species, *Streptococcus pneumoniae*, and many more [62]. TMP is a potent inhibitor of bacterial DHFRs but a much weaker inhibitor of vertebrate DHFRs. In fact, TMP binds to ecDHFR about 10,000 times stronger than it does hDHFR (IC_{50} values against *E. coli* and human enzymes are approximately 5×10^{-9} M and 3×10^{-4} M, respectively) [63]. The strong selectivity for the bacterial enzyme is what allows TMP to be an effective antibiotic, because inhibition leads to cell death in the pathogenic bacterial cells, but not in the human host cells [64]. In contrast, MTX is a potent inhibitor of both bacterial and human DHFR, therefore it is too toxic to be used as an antibiotic [65]. The main contribution to the selectivity of TMP binding is the large positive cooperative effect between TMP and the NADPH cofactor during the formation of the ternary complex with the bacterial DHFR. TMP binds to bacterial DHFR 135 times more tightly in the presence of NADPH [66]. There is a large positive cooperative effect because the binding of one ligand (NADPH) greatly increases the affinity of DHFR for the second ligand (TMP). This large cooperative binding effect is not observed in the complex of TMP with human DHFR [67, 68]. Several explanations have been proposed for the cooperative binding effect including: direct interaction of the ligands with each other, allosteric effects due to conformation change of the protein upon ligand binding, resonance effects that strengthen networks of hydrogen bonds and electrostatic interactions in the ternary complex relative to the binary complexes, and correlated movements of the ligands [64]. There is also a direct hydrophobic interaction between TMP and NADPH in the ternary complex [69]. The analogous interaction in hDHFR is much weaker because there is greater separation between the ligands [67]. However, it is unlikely that the direct contact between ligands alone is enough to explain the large cooperative effect. The additional free energy change may be due to a conformational change in the enzyme resulting in a more favorable interaction of DHFR with both ligands [64].

Another interesting aspect of TMP binding to bacterial DHFR is the observation of two coexisting conformational states in the ternary complex of the *Lactobacillus casei* enzyme [70, 71]. The complex of lcDHFR with TMP and $NADP^+$ exists in solution as a mixture of approximately equal amounts of two slowly interconverting conformational states [72, 73]. NMR experiments were used to characterize the conformational equilibrium of the *L. casei* enzyme complexes, which showed that the active site is clearly involved. There are significant differences in the environment of the bound ligands between the two conformations but conformational effects are not restricted to the active site

[71]. Involvement of six of the seven His residues was observed, yet only two of these are likely to be in close contact with the bound ligands. This implies that residues distal to the active site are also involved in establishing an equilibrium between the two interconverting conformers. In addition, the differences in protein structure between the two conformations appear to determine the nature of the difference in ligand environment between the two conformations. Complexes formed with various structural analogues of TMP or $NADP^+$ clearly exist in the same two conformations, but the equilibrium constant between the two varies from less than 0.1 to 2.3 in different complexes [71–73]. Another interesting observation regarding the conformational dynamics in TMP-bound DHFR is that, while two conformers are observed for the *L. casei* enzyme, only one conformation is observed for the *E. coli* DHFR complex [70]. Thus, conformational equilibrium does not only vary based on the ligands bound, but also varies among species.

4.3 Pyrimethamine

Pyrimethamine acts as an anti-malaria agent by selectively inhibiting the DHFR domain of pfDHFR-TS, and other *Plasmodium* species (Fig. 9). Important amino acid residues involved in the binding of PYR to DHFR include Ile14, Cys15, Asp54, Phe58, Pro113, and Ile164 [40]. PYR is a potent and selective inhibitor of pfDHFR, with an inhibitory constant of 0.2 ± 0.02 nM. It also inhibits pvDHFR with an inhibitory constant of 0.16 ± 0.03 nM

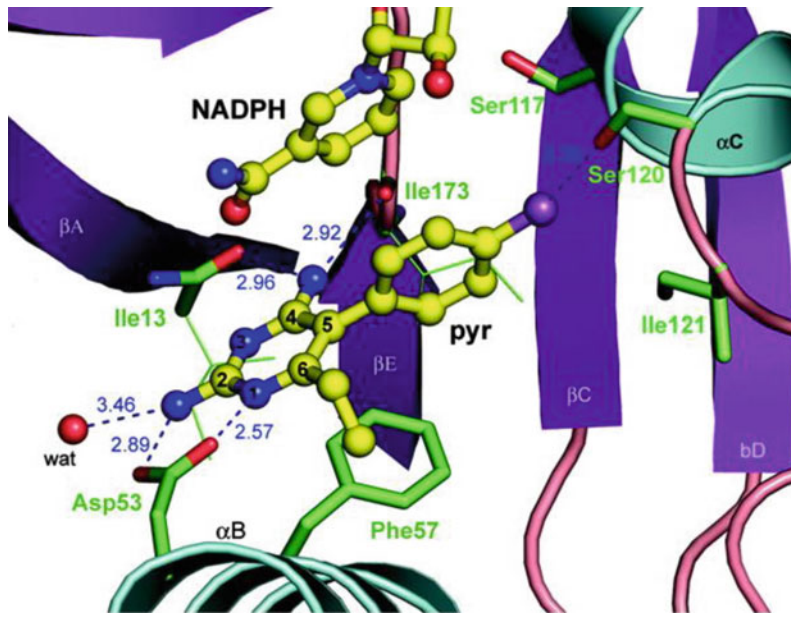

Fig. 9 PYR bound to the active site of the pvDHFR domain of pvDHFR-TS. Interactions between PYR and the enzyme include electrostatic interactions, shown as dotted lines. The numbers next to the lines are bond distances in Å. (Figure reproduced from ref. 74)

Table 1
Comparison of inhibitory constants for MTX, TMP, and PYR in bacterial, vertebrate, and plasmodial DHFRs

Species	$K_{I(MTX)}$ (nM)	$K_{I(TMP)}$ (nM)	$K_{I(PYR)}$ (nM)	References
E. coli	0.0010	0.080	ND[a]	[75]
H. sapiens	0.0034	200	120	[75, 76]
P. falciparum	0.24	11	0.2–1.5	[40, 74, 75, 77]
P. vivax	5.2	98	0.16	[74, 75]

The difference in K_I values highlights the species-specific nature of different DHFR inhibitors
[a]Not determined

[74]. In contrast, human and bacterial DHFRs have a much lower affinity for PYR. PYR displays a more than 1000-fold decrease in binding affinity for vertebrate DHFR and a 5000-fold decrease for ecDHFR [48]. In addition, MTX and TMP show a notable decrease in inhibitory activity for pfDHFR and pvDHFR compared to bacterial and human DHFR [75]. While MTX is still a potent inhibitor of pfDHFR and pvDHFR, the K_I values for MTX binding to these enzymes show a several hundred to several thousand-fold decrease in inhibitory activity compared to the bacterial and human enzyme. TMP does not appear to effectively inhibit either pfDHFR or pvDHFR. This data further illustrates the significance of enzymatic differences among species and the opportunity to develop novel species-specific inhibitors (Table 1).

Despite being a potent inhibitor of PfDHFR, pyrimethamine resistance is extremely common and problematic. Pyrimethamine-resistant strains of PfDHFR have been reported as early as the 1950s [78]. Since the emergence of pyrimethamine-resistant malaria has emerged, many studies have studied these mutants to determine important residues that confer resistance. A S108N mutation was determined to be responsible for many different PYR-resistant strains of *P. falciparum* [75, 79]. The S108N mutation is seen in single, double, triple, and quadruple PYR-resistant mutants [75]. S108 is an active site residue. Even though it does not interact with PYR in the WT enzyme, it clearly is an important residue since the S108N mutant is implicated in almost all naturally occurring PYR-resistant strains of *P. falciparum*. In addition to PYR, *P. falciparum* has developed resistance to other common inhibitors of the DHFR domain of PfDHFR-TS, including cycloguanil and WR99210. The emergence of antifolate resistance in malaria parasites highlights the importance of developing novel small molecule inhibitors that bind PfDHFR selectively.

5 Effects of Distal Mutations and Evidence for Conformational Selection in DHFR

Over the last 30 years, the effect of mutations distal from the active site on catalytic activity and inhibitor binding has become the focus of many studies. The terms distal and allosteric are used to indicate that an amino acid residue is located away from the active site [80]. To understand the effects of distal mutations, it is necessary to recognize the flexible nature of proteins. Enzymes are intrinsically flexible molecules and undergo conformational changes upon ligand binding and throughout catalysis. A "network" of amino acids located in and away from the active site may allosterically regulate the protein and form "global protein dynamics" that are crucial to enzyme catalysis [81, 82]. The term "global protein dynamics" refers to the overall motions exhibited by all atoms in the protein. Therefore, the effect of a distal mutation on catalysis and ligand binding is transmitted indirectly, through the network of amino acids comprising the protein [80]. The wealth of information available about DHFR makes it a great enzyme for studying global protein dynamics. In this work, research and methods on the role of distal residues and effect of distal mutations in DHFR are outlined. We describe the application of these methods to address several questions: (a) Do allosteric mutations effect the conformational motions associated with inhibitor binding? (b) Are the conformational motions inhibitor specific? (c) Do allosteric mutations alter the conformational equilibrium of DHFR prior to inhibitor binding? and (d) What is the significance of these effects on inhibitor binding and specificity?

Numerous DHFR mutagenesis studies have been published which focus on the effect of distal residues of DHFR catalysis and inhibitor binding. Several mechanisms have been proposed to explain how distal mutations change enzyme function. One proposes that the effect is due to changes in the conformational equilibria between the different conformers. The equilibrium is disturbed because some of the intermolecular contacts that stabilize particular conformers have been altered by the mutation, which emphasizes the delicate balance of the conformational ensemble [80]. A second mechanism proposes that changes caused by a distal mutation result in a different pattern of interactions with the rest of the protein and that these changes are propagated throughout the protein [80]. A third mechanism involves changes in protein conformational motion. In this case, the functional effect is due to changes in flexibility and mobility, rather than structural changes in the protein [80]. In this section, we will review several single and multiple site mutation studies on the catalytic activity and inhibitor binding of DHFR.

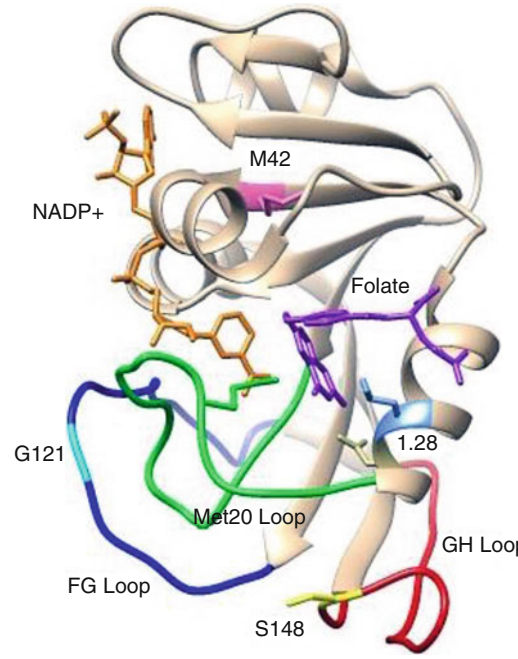

Fig. 10 The structure of DHFR in complex with $NADP^+$ (orange) and DHF (purple). The Met20 loop (green) is in the closed conformation. The FG loop and GH loops are illustrated in dark blue and red, respectively. The distal residues L28 (light blue), M42 (pink), G121 (cyan), and S148 (yellow) are displayed and labeled. (This figure was generated using the program Chimera [28])

5.1 Mutations in the Adenosine Binding Subdomain

M42 in ecDHFR is a highly conserved residue located in the hydrophobic core of the adenosine binding domain, approximately 10 Å from the site of catalysis in bacterial DHFRs (Fig. 10) [83]. Although M42 is distal to the active site, mutations at this site have been shown to have a profound effect on catalysis and ligand affinity. One well-studied mutation is the substitution of M42 with Trp (M42W) [55, 80, 83–87]. Hydride transfer rates of this mutant were measured using stopped-flow kinetics using both single and multiple turnover conditions. The concentrations in a single turnover experiments are set up such that the enzyme concentration exceeds the substrate concentration and thus only one conversion of substrate to product takes place. Single turnover conditions allow the researchers to isolate particular events in the catalytic cycle and avoid repeated cycling. The authors recorded the changes in absorbance and emission of the cofactor NADPH over time after mixing the enzyme-cofactor complex with the substrate DHFR. M42W showed a 41-fold reduction in the rate of hydride transfer, and increases the rate of product dissociation, making hydride transfer the rate-limiting step of catalysis [80]. In addition, kinetic studies similar to what is described above revealed that this mutation introduces a structural rearrangement step into the reaction cycle, which has a significant impact on catalysis [83].

To gain insight into how the mutations influence activity, Rod and coworkers employed several long (10 ns) molecular dynamics simulations of M42W-DHFR and wild type. They simulated the Michaelis complexes using explicit solvent and the 1RX2 crystal structure. They examined whether the mutations impact correlated motions and/or the distribution of conformers sampled by the enzyme and the resulting effects on hydrogen bonds. The authors quantitated the coupling between residues by calculating the covariance between the fluctuations of the two residues. The results indicate that the dynamics of the closed conformation in the M42W mutant are altered because the mutation disrupts a network of coordinated motion that promotes hydride transfer [84, 88].

In a study by Mauldin et al., NMR relaxation data was used to examine the dynamics of M42W-DHFR in the ternary complex with THF. The authors used a strategy where they measured the conformational fluctuations of backbone amide and side-change methyl groups on multiple timescales. They attributed changes in picosecond to nanosecond dynamics to mutational effects propagating throughout a network of interacting residues and micro- to millisecond timescale changes to the mutation resulting in an increased rate of switching in the catalytic core. They observed two distinct groups of residues that experience R_2 dispersion in M42W-DHFR: 15 residues within the catalytic core of the protein and a cluster of 5 residues lining the pABG binding cleft. They found that conformational switching within the pABG binding cleft may act to eject THF from M42W-DHFR [83]. In addition, they proposed that M42 acts as a "dynamic hub" in DHFR by coordinating motion on multiple timescales and that disrupting these dynamic interactions may be an effective method of allosterically modulating DHFR function [83].

5.2 Mutations in the Met20 Loop

The direct role of the Met20 loop in DHFR catalysis was outlined in Subheadings 2.2 and 2.3. Although the Met20 loop is directly implicated in the catalytic cycle, some residues are considered distal because the side chains are oriented toward the solution and do not contact cofactor or substrate [80]. An investigation of the contribution of the Met20 loop to DHFR catalysis was performed by constructing a DHFR deletion mutant of four residues in the Met20 loop (residues 16-19) [89]. Three of these four residues are considered distal. However, N18 is not considered distal because it is within contact distance to the cofactor. The deletion mutant (DL1) resulted in a 400-fold decrease in the rate of hydride transfer (950 s^{-1} for WT enzyme and 1.7 s^{-1} for DL1). The K_M and K_d values for DHF and NADPH increased, but not drastically. These observations support Met20 loop acting as an active site gate that affects the organization of bound substrate and cofactor to form an active complex [80].

5.3 Mutations in the FG Loop

5.3.1 G121V

G121 in ecDHFR is one of the most thoroughly studied distal residues and has been the subject of numerous kinetic and mutagenesis studies. It is universally conserved among all prokaryotic DHFRs, which suggests that it plays a crucial role in the function of DHFR [55]. G121 is located on the FG loop, 19 Å from the active site of the enzyme (Fig. 10) [80]. The substitution of Gly with Val (G121V) has been the subject of numerous studies. Gekko et al. demonstrated that the G121V mutant decreased the rate of steady-state catalysis 20-fold, concluding that amino acid substitutions at position 121 significantly influence its enzymatic function [55, 90]. The fact that enzyme function is affected by mutations in residues far from the active site suggests global dynamics of the protein play an important role in catalysis [90]. In a later study, Cameron et al. analyzed the full kinetic scheme of G121V DHFR and determined that the rate of hydride transfer decreased 170-fold. In addition, G121V was found to introduce a catalytically significant conformational exchange preceding the hydride transfer step, at a rate of 3.5 s^{-1} [91]. In contrast to WT DHFR, the closed conformation is energetically disfavored for G121V DHFR [92]. As mentioned earlier, the Michaelis complex for ecDHFR is in the closed conformation. The mutant remains in the occluded conformation, which interferes with coupled loop movements and impairs catalysis by destabilizing the closed Michaelis complex and introducing an extra conformational exchange step into the kinetic pathway [92].

In a study by Boehr et al. nuclear magnetic resonance relaxation experiments were used to determine the mechanism by which the G121V mutation effects DHFR kinetics and dynamics on the picosecond to nanosecond and microsecond to millisecond timescales. The authors recorded the 15 N relaxation data for both wild type and G121V DHFR with folate bound. They found that the mutant ternary complex (G121V:NADPH:THF) adopts an occluded conformation which is very similar to that of the wild-type ternary complex. However, this mutation causes substantial changes in dynamics, restricting the amplitude and altering the timescale of motions of residues in the FG loop and the Met20 loop. The effects of the G121V mutation are transmitted to distal sites through subtle changes in the accessible conformational space by molecular fluctuations on the picosecond to nanosecond timescales. Therefore, they conclude that their results are consistent with theoretical experiments that suggest that long-range allostery in DHFR arises from a redistribution of the conformational ensemble, with significant contributions from perturbations of the protein dynamics [93].

In another study, Mauldin et al. used NMR to investigate whether the decrease in catalytic efficacy caused by the G121V mutant was the result of changes in structure, flexibility, or both [94]. Since the G121V mutants favor the occluded conformation,

they used MTX and the reduced cofactor NADPH to create a model of the transition state. The high affinity of MTX for DHFR effectively "locks" the enzyme in the closed conformation in WT ecDHFR; thus, they reasoned that it may do the same for the mutant enzyme. This allowed them to isolate the mutant in the closed conformation, which is the catalytically relevant conformation of DHFR. Using ^{15}H and ^{2}H NMR spin relaxation experiments, they observed that the most dramatic effect of the G121V mutation involves changes in the dynamics of the FG and Met20 loops on the μs–ms timescale. In the WT DHFR:NADPH:MTX complex, loop motion is suppressed so that the complex favors the closed conformation. However, in the mutant complex, the FG and Met20 loops undergo fluctuations from the closed conformation. These dynamic fluctuations serve to decrease the population of conformers having the correct active site conformation for catalysis, providing an explanation for the decrease in catalytic activity observed for the G121V mutant [94].

5.3.2 ΔG121

The effect of deletion and insertion mutations in the FG loop has also been studied to probe the role of distal residues in DHFR catalysis. Deletion of the G121 residue (ΔG121) results in decreased binding to NADPH by 20-fold, as well as a 550-fold decrease in the hydride transfer rate [80, 95]. Additionally, ΔG121 requires conformational changes dependent on the initial binary complex to attain the Michaelis complex. Insertion mutants also displayed a significant decrease in substrate and cofactor binding. However, the insertion of a glycine residue into a modified FG loop eliminated the conformational changes required to attain the Michaelis complex seen in the ΔG121 mutant [95]. These observations suggest that the FG loop plays a role in the formation of liganded complexes and proper orientation of substrate and cofactor during catalysis. Through a transient interaction with the Met20 loop, alterations to the FG loop can coordinate proximal and distal effects on ligand binding and catalysis that implicate a variety of enzyme conformations involved in the catalytic cycle [95].

5.4 Mutations in the GH Loop

5.4.1 S148

S148 is located in the GH loop (residues 142-149) of ecDHFR, 18 Å away from the active site. This residue is involved in hydrogen bonding interactions that modulate the conformation of the Met20 loop [80, 95]. To gauge the importance of these hydrogen bonding interactions, S148 was replaced by Asp, Ala, and Lys. These mutations increased the affinity for the NADPH cofactor, but significantly decreased the affinity of the enzyme for DHF (Table 2). Further analysis revealed that these mutations predominantly effected the ligand release rates. Mutations at residue 148 altered the preferred catalytic pathway by introducing branches at key intermediates [80].

Table 2
Thermodynamic and kinetic data for WT DHFR and single mutants

DHFR	$K_{d(DHF)}$ (μM)	$K_{d(NADPH)}$ (μM)	k_{hyd} (s^{-1})	k_{cat} (s^{-1})	References
WT	0.22 ± 0.06	0.33 ± 0.06	220	12	[23]
M42F	0.35 ± 0.05	0.22 ± 0.04	159 ± 17		[85]
M42W	0.43 ± 0.1	0.27 ± 0.03	5.6 ± 0.4		[85]
ΔG121	0.26 ± 0.03	3.2 ± 0.4	3.7 ± 0.4		[95]
G121S	0.39 ± 0.05	3.2 ± 0.04	3.7 ± 0.4		[85]
G121V	0.36 ± 0.02	14.2 ± 0.8	1.4 ± 0.2		[91]
S148A	1.06 ± 0.09	0.049 ± 0.003	157 ± 3	6.6 ± 0.8	[96]
S148D	0.18 ± 0.02	0.15 ± 0.01	319 ± 3	4.6 ± 0.1	[96]
S148K	0.72 ± 0.14	0.16 ± 0.01	162 ± 2	5.7 ± 0.1	[96]

S148 has also been shown to play an important role in the occluded conformation of ecDHFR through mutational analysis. The occluded conformation is stabilized via two hydrogen bonds between Asn23 in the Met20 loop and Ser148. In a study conducted by Behiry et al. S148 was replaced with proline (S148P). Pro cannot form the hydrogen bond interactions necessary to stabilize the occluded conformation and the S148P mutant is useful for investigating the importance of the occluded conformation in DHFR catalysis. Their results indicated that the occluded conformation assists in the release of the oxidized cofactor $NADP^+$ and progression through the catalytic cycle [97].

5.5 Mutations in the DHF Binding Cleft

The residues in the DHF binding site of ecDHFR are I5, A6, A7, M20, D27, L28, F31, R52, R57, I94, and T113. Among these residues, the backbone of I5 and side chains of D27, R52, R57 directly interact with DHF via hydrogen bonding interactions [98]. The other residues can be considered distal in this context because they are not directly involved in binding of substrate or the catalytic reaction. In a recent study by Abdizadeh et al., replacement of L28 with an Arg residue (L28R) was investigated to structurally and dynamically characterize the WT and L28R ecDHFR in the presence of DHF and TMP. They used the NAMD package to model the molecular dynamics of the protein–water systems to determine the conformational space, loop dynamics, and hydrogen binding interactions for the WT and mutant enzyme. The protein was soaked in a cubic solvent box with at least a 10 Å layer of solvent in each direction from any atom of the protein to the edge of the box and the ionic strength in the simulations was kept at 150 mM. All systems were subjected to 10,000 steps of energy minimization. The resulting structures were

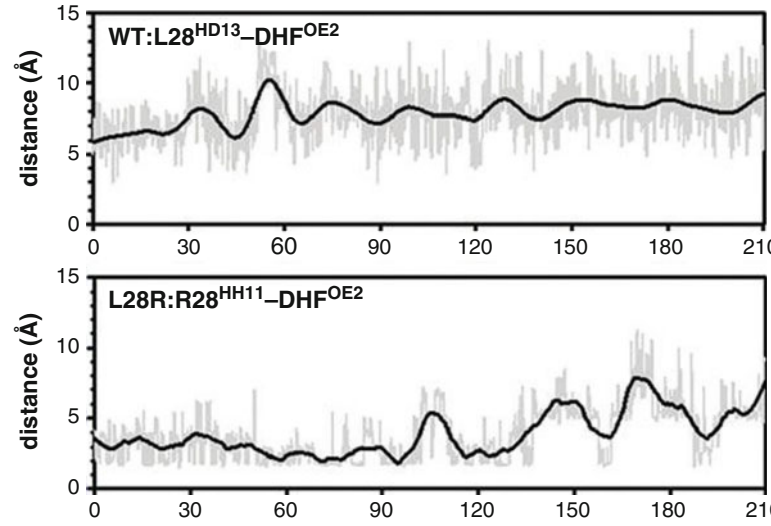

Fig. 11 Comparison of WT DHFR:DHF and L28R:DHF complexes. In the WT enzyme, L28 does not form hydrogen bond interactions with the substrate. In the mutant enzyme, R28 interacts with the p-aminobenzoyl glutamate tail via its side chain further stabilizing the protein–ligand interaction and altering the binding conformation of DHF in the binding site. (This figure was reproduced from Fig. 4 in ref. 98)

analyzed at 1 atm and 310 K until volumetric fluctuations were stabilized and the desired average pressure was maintained. The DHF-bound WT DHFR structure (PDB code 1rx2) was used in the molecular dynamics simulations. In addition, experimental values for k_{cat}, K_m, and K_I values were measured. Isothermal titration calorimetry (ITC) measurements were conducted that distinguish enthalpic and entropic contributions to TMP binding [98]. This thorough study provides insight into the binding kinetics and dynamics of TMP to WT DHFR and the L28R mutant, which confers resistance to TMP [98]. L28R is a common TMP resistance-conferring mutation [98]. In WT ecDHFR, L28 does not exhibit any hydrogen bond interactions with DHF (Fig. 11). However, for the L28R mutant, sample donor–acceptor distances between the methyl group of R28 and DHF indicate the presence of hydrogen bonds between the p-aminobenzoyl glutamate tail of DHF and the α-amino group of R28 (Fig. 11) [98]. This interaction stabilizes the DHF-bound complex and provides a unique mechanism of resistance. Typically, mutations that confer resistance to a competitive inhibitor make enzymes more promiscuous and decrease affinity for both inhibitor and the natural substrate. For the L28R mutant, an increase in DHF affinity and a decrease in TMP affinity are observed (Table 3). While this also results in slower product release and catalytic rate for the L28R mutant, the

Table 3
Kinetic and competitive inhibition measurements for WT and L28R DHFR bound to the substrate, DHF, and the competitive inhibitor, trimethoprim

DHFR	k_{cat} (s^{-1})	K_M (MM)	K_I (nM)
WT	8.16 ± 3.27	1.49 ± 0.14	2.39 ± 1.06
L28R	1.30 ± 0.01	0.62 ± 0.01	24.63 ± 1.34

Table adapted from Table 2 in ref. 98

Table 4
ITC measurement data for WT and L28R DHFR

DHFR	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	ΔS (cal mol^{-1} k^{-1})	K_d (nM)
WT	−11.2 ± 0.5	−13.1 ± 1.4	−6.5 ± 3.0	4.5 ± 0.9
L28R	−10.8 ± 0.2	−6.8 ± 0.5	13.4 ± 2.6	13.1 ± 1.5
Difference	+0.4 ± 0.5	+6.3 ± 1.5	+19.9 ± 4.0	

Table adapted from Table 2 in ref. 98

enzyme is still able to maintain a sufficient rate of production formation to be catalytically viable (Tables 3 and 4) [98].

The overall structure of the TMP-DHFR complex is essentially unaltered upon the introduction of the L28R mutation. However, large differences in thermodynamic and kinetic parameters are observed between the WT and mutant enzymes (Tables 3 and 4) [98]. In the WT:DHF complex, there are strong correlations between different regions throughout the protein. In particular, the Met20 loop is correlated to residues 47-59, 85-89, 119-126, and 142-149 [98]. Abdizadeh et al. found that the L28R mutation suppresses the overall cross correlations between different regions of DHFR in the presence of substrate compared to the WT enzyme. Notably, the concerted movements of the GH loop with the rest of the enzyme essentially disappear [98]. Therefore, the L28R mutation does not only alter DHFR activity via increased hydrogen bonding with substrate. The effects of the mutation can also be observed by the altered global protein dynamics.

5.6 Multiple Mutations

The results discussed in Subheadings 5.1–5.5 show that single mutations, distant from the active site, are capable of affecting enzyme activity. These observations prompted work on the effect of multiple mutations at residues that are distal to the active site and spatially separated from each other [80]. Some of the most thoroughly studied multiple DHFR mutants and their effects on catalysis and ligand binding will be discussed in this section.

Table 5
Kinetic data and nonadditive mutational effects on hydride transfer rates for M42-G121 DHFR double mutants

DHFR	k_{hyd} (s^{-1})	k_{hyd} ratio, WT/Mutant	Additivity factor[a]	Nonadditivity[b]	Ref.
WT	220 ± 8	1.0			[81]
M42F	159 ± 17	1.4			[81, 85]
M42W	5.6 ± 0.4	41			[81]
G121A	38 ± 3	6.0			[81]
G121S	3.7 ± 0.4	62			[81]
G121V	1.4 ± 0.2	163			[81]
M42F-G121A	1.3 ± 0.2	175	8.4 ± 1.4	21 ± 5	[81]
M42F-G121S	0.46 ± 0.08	496	87 ± 16	5.7 ± 1.5	[81]
M42W-G121A	0.27 ± 0.04	844	34	248	[81]
M42W-G121S	0.07 ± 0.01	3257	21	155	[81]
M42W-G121V	0.030 ± 0.005	7600	7.8	974	[81]

[a]Additivity factor is the factor by which k_{hyd} would be lowered based on an additive effect of the individual mutations
[b]The additive effect of the individual mutations does not match the actual reduction for the double mutant

5.6.1 M42-G121 Double Mutants

Several different M42-G121 mutants have been constructed and analyzed to determine whether these residues are coupled through global protein dynamics and to determine why and how distal mutations affect the hydride transfer step of the DHFR catalytic cycle [84]. These mutants include M42-G121A, M42F-G121S, M42W-G121A, M42W-G121S, and M42W-G121V (Table 3) [85]. These double mutants generally showed little changes in substrate and cofactor binding bit synergistic decreases in the rate of hydride transfer rates. For example, the hydride transfer rates of the double mutants were decreased by 3200- and 7600-fold in M42W-G121S and M42W-G121V compared to the WT enzyme, respectively (Table 5) [80]. In addition, kinetic measurements indicate that double mutants involving these residues are nonadditive, meaning that the effect of a double mutant is much greater than the sum of the effects of the single mutations. For example, if the effect of the M42F-G121A double mutant was additive, k_{hyd} would decrease 8.4-fold ($1.4 \times 6.0 = 8.4$). However, this is not the case and therefore the effect is nonadditive. The nonadditivity observed is the factor by which k_{hyd} is lowered more than would be predicted by a simple additive effect. For example, for M42F-G121A k_{hyd} is lowered 21-fold more than predicted by a simple additive effect ($175/8.4 = 21$) [81]. The observed nonadditive effects suggest a coupling of the FG loop to distant regions of the enzyme [99].

Table 6
Comparison of hydride transfer rates for the M42F-G121S-S148A triple mutant to associated mutants and WT DHFR

DHFR	k_{hyd} (s^{-1})	k_{hyd} ratio, WT/Mutant
WT	220	1
M42F	159	1.4
G121S	3.9	56
S148A	157	1.4
G121S-S148A	18	12
M42F-S148A	92	2.4
M42F-G121S	2.9	76
M42F-G121S-S148A	12	18

Table was adapted from ref. 101

5.6.2 M42F-G121S-S148A Triple Mutant and Associated Mutants

In a study by Wong et al. a comprehensive analysis of coupled motions correlated to hydride transfer rates was applied to the triple M42F-G121S-S148A and the associated single and double mutants. Because these three residues are all located in different regions of the enzyme, analysis of the triple mutant may provide further insight into the coupled interactions and motions between different regions of DHFR (Fig. 10). The k_{hyd} rates of the single mutants M42F, G121S, and S148A were compared with the double mutants G121S-S148A, M42F-S148A, M42F-G121S and the triple mutant M42F-G121S-S148A. Hydride transfer rates for all the associated mutants decreased significantly and the triple mutant displayed an 18-fold decrease in k_{hyd} (Table 4). The results illustrate that each mutant samples a unique set of motions and nonadditivity was observed for the hydride transfer rates, which may be explained by nonadditive modulations of the network of coupled motions involved in the hydride transfer step. In addition, their calculations indicated that distal mutations can introduce subtle structural perturbations that impact the hydride transfer rate by altering the conformational ensemble of DHFR. Since distal mutations are coupled to each other through long-range electrostatic and hydrogen bonding networks, the introduction of site-specific mutations alters the motions of the entire enzyme [100, 101] (Table 6).

6 Future Prospects: Protein Dynamics and Conformational Selection in Drug Design

Classically, drug design efforts have ignored protein motion and flexibility for several reasons, including time constraints and methodology. Thus, protein motions are normally regarded as small

perturbations that can be disregarded and that static models are sufficient for drug discovery efforts [20]. However, based on what we now know about the role of protein dynamics in catalysis and inhibitor binding, assuming that protein motions can be ignored may be a mistake. Mauldin et al. used ^{15}N and ^{2}H NMR spin relaxation experiments to investigate how the functional motions of DHFR respond to MTX and TMP. They profiled the motions of ecDHFR bound to NADPH in the presence and absence of the two inhibitors and found that drug binding at the substrate binding site breaks up the usual μs–ms loop motions of the holoenzyme into smaller, unproductive clusters of local motion, which they refer to as "dynamic dysfunction" [102]. Interestingly, they found that both MTX and TMP cause the same "dysfunction," which suggests that these dynamic changes may be important to inhibitory activity. These results demonstrate that MTX and TMP do not just block substrate binding. They also cause a breakdown of communication within the network of collective motions required for DHFR catalysis [20, 102]. The results of this study are supported by the results of a more recent investigation of TMP binding by Abdizadeh et al. They also found that TMP effectively "locks" DHFR in the closed conformation where the closed conformation dominates in the ensemble. They propose that binding of TMP sends a "signal" for conformational change on the μs–ms timescale for the drugged complex to remain locked in the closed conformation. If this is true for DHFR inhibitors, then it is reasonable to assume that other drugs may act in a similar fashion.

The study conducted by Mauldin et al. highlights several points relevant for drug discovery. First, flexibility-function studies can indicate new modes of drug action that would not be observed using traditional, static model drug discovery and design strategies. Second, drug action is most likely broader than we originally thought. Competitive inhibitors may also inhibit functional protein dynamics in addition to preventing substrate from binding to the active site. Third, protein functional motions can be distributed among networks of amino acid residues throughout the protein. Thus, motions at one site can be inhibited by binding at a distant site. Although this makes the drug discovery and design process much more complex, it also widens the range of potential inhibitor binding sites, facilitating the design of more specific inhibitors. Finally, protein dynamics studies should be complemented by studying the dynamics of the ligand. Since drug design entails modifying the ligand, not the protein target, the conformation of the ligand when bound to the protein can provide added insight [20].

In addition, as described earlier, inhibitor binding modulates the conformational ensemble of a protein by shifting the conformational equilibrium. By studying the conformational motions of a protein or enzyme, we may be able to identify the ligands preferred

conformer and design inhibitors accordingly. If inhibitors can be designed based on the structure of the preferred conformer, these inhibitors would bind the preferred conformer more selectively. Binding of inhibitor would then shift the conformational equilibrium in favor of that conformer, increasing drug binding to the target [12]. In summary, incorporation of protein motion and conformation studies represents a great opportunity for drug discovery design. Despite such research being ill-suited to high-throughput methods, it may be necessary because we may be overlooking a realm of novel drug design possibilities.

7 Conclusion

In this work, DHFR was used as a case study for the methods and approaches to study the role of conformational selection and motions in ligand binding, the effects of distal mutations on the conformational motions of DHFR on inhibitor binding, and the potential implications these findings may have on drug discovery and design. The studies as a whole show that conformational motions play a crucial role in ligand binding during the catalytic cycle and binding of inhibitor molecules such as MTX, TMP, and PYR. Distal mutations affect ligand binding and the conformational motions associated with inhibitor binding. Novel methods are needed to investigate the hypothesis that proteins exist as a conformational ensemble in a state of dynamic equilibrium and a network of amino acids located both near and away from the active site are required for protein function. Distal mutations most likely exert their effects by modulating conformational motions indirectly through this network of amino acids that make up the global protein dynamics and these effects can only in part be captured through enzyme kinetics and NMR. Finally, these observations may have useful implications in drug design. Conformational changes associated with inhibitor binding were shown to be inhibitor specific, which implies that drug action may be broader than we originally thought. Inhibitors may exert their effects through disrupting functional protein dynamics instead of simply blocking substrate binding to the active site. Therefore, methods to study the conformational motions that may be potential "drug targets" can reveal opportunities for the design of better and more selective inhibitors.

References

1. Vogt AD, Di Cera E (2012) Conformational selection or induced fit? A critical appraisal of the kinetic mechanism. Biochemistry 51 (30):5894–5902
2. Vogt AD, Pozzi N, Chen Z, Di Cera E (2014) Essential role of conformational selection in ligand binding. Biophys Chem 186:13–21
3. Csermely P, Palotai R, Nussinov R (2010) Induced fit, conformational selection and

independent dynamic segments: an extended view of binding events. Trends Biochem Sci 35(10):539–546

4. Nussinov R, Ma B, Tsai CJ (2014) Multiple conformational selection and induced fit events take place in allosteric propagation. Biophys Chem 186:22–30
5. Michel D (2016) Conformational selection or induced fit? New insights from old principles. Biochimie 128-129:48–54
6. Changeux JP, Edelstein S (2011) Conformational selection or induced fit? 50 years of debate resolved. F1000 Biol Rep 3(19):1–15
7. Frederick KK, Marlow MS, Valentine KG, Wand AJ (2007) Conformational entropy in molecular recognition by proteins. Nature 448(7151):325–329
8. Gianni S, Dogan J, Jemth P (2014) Distinguishing induced fit from conformational selection. Biophys Chem 189:33–39
9. Hammes GG, Chang Y, Oas TG (2009) Conformational selection or induced fit: a flux description of reaction mechanism. Proc Natl Acad Sci U S A 106(33):13434–13741
10. Hatzakis NS (2014) Single molecule insights on conformational selection and induced fit mechanism. Biophys Chem 186:46–54
11. Ma B, Nussinov R (2010) Enzyme dynamics point to stepwise conformational selection in catalysis. Curr Opin Chem Biol 14 (5):652–659
12. Teague SJ (2003) Implications of protein flexibility for drug discovery. Nat Rev Drug Discov 2(7):527–541
13. Vega S, Abian O, Velazquez-Campoy A (2016) On the link between conformational changes, ligand binding and heat capacity. Biochim Biophys Acta 1860(5):868–878
14. Vogt AD, Di Cera E (2013) Conformational selection is a dominant mechanism of ligand binding. Biochemistry 52(34):5723–5729
15. Watt ED, Shimada H, Kovrigin EL, Loria JP (2007) The mechanism of rate-limiting motions in enzyme function. Proc Natl Acad Sci U S A 104(29):11981–11986
16. Weikl TR, Paul F (2014) Conformational selection in protein binding and function. Protein Sci 23(11):1508–1518
17. Zhou HX (2010) From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions. Biophys J 98(6): L15–L17
18. Schnell JR, Dyson HJ, Wright PE (2004) Structure, dynamics, and catalytic function of dihydrofolate reductase. Annu Rev Biophys Biomol Struct 33:119–140
19. Eisenmesser EZ et al (2002) Enzyme dynamics during catalysis. Science 295 (5559):1520–1523
20. Peng JW (2009) Communication breakdown: protein dynamics and drug design. Structure 17(3):319–320
21. Sawaya MR, Kraut J (1997) Loop and subdomain movements in the mechanism of E. coli DHFR-Crystallographic evidence. Biochemistry 36:586–603
22. Huennekens FM (1994) The methotraxate story: a paradigm for development of cancer chemotherapeutic agents. Adv Enzyme Regul 34:379–419
23. Fierke CA, Johnson KA, Benkovic SJ (1987) Construction and evaluation of the Kinetic Scheme associated with DHFR from Escherichia coli. Biochemistry 26:4085–4092
24. Verma CS, Caves LSD, Hubbard RE, Roberts GCK (1997) Domain motions in dihydrofolate reductase: a molecular dynamics study. J Mol Biol 266:776–796
25. Wallace LA, Robert Matthews C (2002) Highly divergent dihydrofolate reductases conserve complex folding mechanisms. J Mol Biol 315(2):193–211
26. Matthews DA, Alden RA, Bolin JT, Freer ST, Hamlin R, Xuong N, Kraut J, Poe M, Williams M, Hoogsteen K (1977) DHFR-X-ray structure of the binary complex with methotrexate. Science 197(4302):452–455
27. Bystroff C, Kraut J (1991) Crystal structures of E. coli DHFR-the NADP+ holoenzyme and the folate-NADP+ complex. Substrate binding and a model for the transition state. Biochemistry 30:2227–2239
28. Petterson EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF chimera--a visualization system for exploratory research and analysis. J Comput Chem 25(13):1605–1612
29. Bystroff C, Oatley SJ, Kraut J (1990) Crystal structures of E. coli DHFR-the NADP+ holoenzyme and the folate-NADP+ complex. Substrate binding and a model for the transition state. Biochemistry 29(13):3263–3277
30. Appleman JR, Howell EE, Kraut J, Blakely RL (1990) Role of Aspartate 27 of DHFR from E coli in interconversion of active and inactive enzyme conformers and binding of NADPH. J Biol Chem 265(10):5579–5584
31. Howell EE, Villafranca JE, Warren MS, Oatley SJ, Kraut J (1986) Functional role of Aspartic acid-27 DHFR revealed by mutagenesis. Science 231(4742):1123–1128

32. Osborne MJ, Schnell J, Benkovic SJ, Dyson HJ, Wright PE (2001) Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism. Biochemistry 40(33):9846–9859
33. Oyen D, Fenwick RB, Stanfield RL, Dyson HJ, Wright PE (2015) Cofactor-mediated conformational dynamics promote product release from Escherichia coli dihydrofolate reductase via an allosteric pathway. J Am Chem Soc 137(29):9459–9468
34. Abali EE, Skacel NE, Celikkaya H, Hsieh YC (2008) Regulation of human dihydrofolate reductase activity and expression. Vitam Horm 79:267–292
35. PerkinElmer Informatics (2016) *ChemDraw Professional*, 15.1.0.144
36. Tuttle LM, Dyson HJ, Wright PE (2014) Side chain conformational averaging in human dihydrofolate reductase. Biochemistry 53 (7):1134–1145
37. Bhabha G, Ekiert DC, Jennewein M, Zmasek CM, Tuttle LM, Kroon G, Dyson HJ, Godzik A, Wilson IA, Wright PE (2013) Divergent evolution of protein conformational dynamics in dihydrofolate reductase. Nat Struct Mol Biol 20(11):1243–1249
38. Anderson AC (2005) Targeting DHFR in parasitic protozoa. Drug Discov Today 10 (2):121–128
39. Chanama M, Chitnumsub P, Yuthavong Y (2005) Subunit complementation of thymidylate synthase in Plasmodium falciparum bifunctional dihydrofolate reductase-thymidylate synthase. Mol Biochem Parasitol 139(1):83–90
40. Yuvaniyama J, Chitnumsub P, Kamchonwongpaisan S, Vanichtanankul J, Sirawaraporn W, Taylor P, Walkinshaw MD, Yuthavong Y (2003) Insights into antifolate resistance from malarial DHFR-TS structures. Nat Struct Biol 10(5):357–365
41. Mokmak W, Chunsrivirot S, Hannongbua S, Yuthavong Y, Tongsima S, Kamchonwongpaisan S (2014) Molecular dynamics of interactions between rigid and flexible antifolates and dihydrofolate reductase from pyrimethamine-sensitive and pyrimethamine-resistant Plasmodium falciparum. Chem Biol Drug Des 84(4):450–461
42. Yuthavong Y, Yuvaniyama J, Chitnumsub P, Vanichtanankul J, Chusacultanachai S, Tarnchompoo B, Vilaivan T, Kamchonwongpaisan S (2005) Malarial (Plasmodium falciparum) DHFR-thymidylate synthase-structural basis for antifolate resistance and development of effective inhibitors. Parasitology 130:249–259
43. Dunn SMJ, Batchelor JG, King RW (1978) Kinetics of ligand binding to DHFR-binary complex formation with NADPH and coenzyme analogues. Biochemistry 17 (12):2356–2363
44. Dunn SMJ, King RW (1980) Kinetics of ternary complex formation between DHFR, coenzyme, and inhibitors. Biochemistry 19 (4):766–773
45. Cayley PJ, Dunn SMJ, King RW (1981) Kinetics of substrate, coenzyme, and inhibitor binding to E. coli DHFR. Biochemistry 20 (4):874–879
46. Reddish MJ, Vaughn MB, Fu R, Dyer RB (2016) Ligand-dependent conformational dynamics of Dihydrofolate Reductase. Biochemistry 55(10):1485–1493
47. Karabulut S, Sizochenko N, Orhan A, Leszczynski J (2016) A DFT-based QSAR study on inhibition of human dihydrofolate reductase. J Mol Graph Model 70:23–29
48. Schweitzer BI, Dicker AP, Bertino JR (1990) Dihydrofolate reductase as a therapeutic target. FASEB J 4:2441–2452
49. Hashkes PJ, Becker ML, Cabral DA, Laxer RM, Paller AS, Rabinovich CE, Turner D, Zulian F (2014) Methotrexate: new uses for an old drug. J Pediatr 164(2):231–236
50. Lv C, Zhao W, Guo R (2012) Methotrexate: pharmacology, clinical application and adverse effects. In: Castillo VS, Moyano LA (eds) Methotrexate. Nova Science Publishers, Inc., New York, pp 217–226
51. Verma U, Verma N (2012) Methotrexate: pharmacology, clinical uses and adverse effects. In: Castillo VS, Moyano LA (eds) Methotrexate. Nova Science Publishers, Inc., New York, pp 107–126
52. Falzone CJ, Wright PE, Benkovic SJ (1991) Evidence for two interconverting protein isomers in the methotrexate complex of DHFR from E. coli. Biochemistry 30:2184–2191
53. Appleman JR, Howell EE, Kraut J, Kühl M, Blakely RL (1988) Role of Aspartate 27 in the binding of methotrexate to DHFR from E coli. J Biol Chem 263(19):9187–9198
54. Cannon WR, Garrison BJ, Benkovic SJ (1997) Consideration of the pH dependent inhibition of DHFR by methotrexate. J Mol Biol 271:656–668
55. Mauldin RV (2009) Multi-timescale dynamic effects of antifolate binding and mutations in DHFR

56. Stockman BJ, Nirmala NR, Wagner G, Delcamp TJ, DeYarman MT, Freishman JH (1991) Methotrexate binds in a non-productive orientation to human dihydrofolate reductase in solution, based on NMR spectroscopy. FEBS Lett 283 (2):267–269
57. Oefner C, D'Arcy A, Winkler FK (1988) Crystal structure of human DHFR complexed with folate. Eur J Biochem 174:377–385
58. Huang FY, Yang QX, Huang T (1991) 15N NMR studies of the conformation of E. coli DHFR in complex with folate or methotrexate. FEBS Lett 289(2):231–234
59. Cheung HT, Birdsall B, Feeny J (1992) 13CNMR studies of complexes of Ecoli DHFR formed with methotrexate and with folic acid. FEBS Lett 312(2):147–151
60. Curtis N, Moore S, Birdsall B, Bloxsidge J, Gibson CL, Jones JR, Feeny J (1994) 3H-NMR studies of multiple conformations and dynamic processes in complexes of folate and methotrexate with L. casei DHFR. Biochem J 303:401–405
61. Bolin JT, Filman DJ, Matthews DA, Hamlin RC, Kraut J (1982) Crystal structures of E coli and Lactobacillus casei DHFR refined at 1.7A resolution: general features and binding of methotrexate. J Biol Chem 257 (22):13650–13662
62. Kielhofner MA (1990) Trimethoprim-sulfamethoxazole-pharmacokinetics, clinical uses, and adverse reactions. Tex Heart Inst J 17 (2):87–93
63. Baker DJ, Beddell CR, Champness JN, Goodford PJ, Norrington FEA, Smith DR, Stammers DK (1981) The binding of trimethoprim to bacterial dihydrofolate reductase. FEBS Lett 126(1):49–52
64. Kovalevskaya NV, Smurnyi ED, Birdsall B, Feeney J, Polshavoc VI (2007) Structural factors determining the binding selectivity of the antibacterial drug trimethoprim to dihydrofolate reductase. Pharm Chem J 41(7):350–353
65. Lemke TL, Roche VF, Williams DA, Zito SW (2013) Foye's principles of medicinal chemistry, 7th edn. Lippincott Williams & Wilkins, a Wolters Kluwer business, Baltimore
66. Feeney J, Birdsall B, Kovalevskaya NV, Smurnyy YD, Navarro Peran EM, Polshakov VI (2011) NMR structures of apo L. casei dihydrofolate reductase and its complexes with trimethoprim and NADPH: contributions to positive cooperative binding from ligand-induced refolding, conformational changes, and interligand hydrophobic interactions. Biochemistry 50(18):3609–3620
67. Kovalevskaya NV, Smurnyy YD, Polshakov VI, Birdsall B, Bradbury AF, Frenkiel T, Feeney J (2005) Solution structure of human dihydrofolate reductase in its complex with trimethoprim and NADPH. J Biomol NMR 33(1):69–72
68. Champness JNS (1986) D.K.; Beddell, C.R., crystallographic investigation of the cooperative interaction between trimethoprim, reduced cofactor and dihydrofolate reductase. FEBS Lett 199(1):61–67
69. Polshakov VI, Smirnov EG, Birdsall B, Kelly G, Feeney J (2002) Letter to the editor: NMR-based solution structure of the complex of lactobacillus casei DHFR with trimethoprim and NADPH. J Biomol NMR 24:67–70
70. Huang F, Yang Q, Huang T, Gelbaum L, Kuyper LF (1991) The conformations of trimethoprim/E. coli dihydrofolate reductase complexes: A 15N and 31P NMR study. FEBS Lett 283(1):44–46
71. Birdsall B, Bevan AW, Pascaul C, Roberts GCK, Feeny J, Gronenborn A, Clore GM (1984) Multinuclear NMR characterization of two coexisting conformational states of the Lactobacillus casei DHFR-trimethoprim-NADP+ complex. Biochemistry 23:4733–4742
72. Gronenborn A, Birdsall B, Hyde E, Roberts GCK, Feeney J, Burgen ASV (1981) Direct observation by NMR of two coexisting conformations of an enzyme ligand complex in solution. Nature 290:273–274
73. Gronenborn A, Birdsall B, Hyde E, Roberts G, Feeney J, Burgen A (1981) 1H and 31P NMR characterization of two conformations of the trimethoprim-NADP+ DHFR complex. Mol Pharmacol 20:145–153
74. Kongsaeree P, Khongsuk P, Leartsakulpanich U, Chitnumsub P, Tarnchompoo B, Walkinshaw MD, Yuthavong Y (2005) Crystal structure of dihydrofolate reductase from plasmodium vivax-pyrimethamine displacement linked with mutation-induced resistance. Proc Natl Acad Sci U S A 102(37):13046–13051
75. Volpato JP, Pelletier JN (2009) Mutational 'hot-spots' in mammalian, bacterial and protozoal dihydrofolate reductases associated with antifolate resistance: sequence and structural comparison. Drug Resist Updat 12 (1–2):28–41

76. Goodey NM, Herbert KG, Hall SM, Bagley KC (2011) Prediction of residues involved in inhibitor specificity in the dihydrofolate reductase family. Biochim Biophys Acta 1814 (12):1870–1879
77. McKie JH, Douglas KT, Chan C, Roser SA, Yates R, Read M, Hyde JE, Dascombe MJ, Yuthavong Y, Sirawaraporn W (1998) Rational drug design approach for overcoming drug resistance-application to pyrimethamine resistance in malaria. J Med Chem 41 (9):1367–1370
78. Wilson T, Edeson JFB (1953) Treatment of acute malaria with pyrimethamine. Br Med J 1:253–255
79. Cowman AF, Morry MJ, Biggs BA, Cross GAM, Foote SJ (1988) Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of plasmodium falciparum. Proc Natl Acad Sci U S A 85(23):9109–9113
80. Lee J, Goodey NM (2011) Catalytic contributions from remote regions of enzyme structure. Chem Rev 111(12):7595–7624
81. Agarwal PK, Billeter SR, Ravi Rajagopalan PT, Benkovic SJ, Hammes-Schiffer S (2002) Network of coupled promoting motions in enzyme catalysis. Proc Natl Acad Sci U S A 99(5):2794–2799
82. Okondo M (2015) Effects of allosteric mutations on DHFR. Thesis-Montclair State University
83. Mauldin RV, Lee AL (2010) Nuclear magnetic resonance study of the role of M42 in the solution dynamics of Escherichia coli dihydrofolate reductase. Biochemistry 49 (8):1606–1615
84. Rod TH, Radkiewicz JL, Brooks CL 3rd (2003) Correlated motion and the effect of distal mutations in dihydrofolate reductase. Proc Natl Acad Sci U S A 100 (12):6980–6985
85. Rajagopalan PT, Lutz S, Benkovic SJ (2002) Coupling interactions of distal residues enhance DHFR catalysis-mutational effects on hydride transfer rates. Biochemistry 41:12618–12628
86. Wang L, Goodey NM, Benkovic SJ, Kohen A (2006) The role of enzyme dynamics and tunnelling in catalysing hydride transfer: studies of distal mutants of dihydrofolate reductase. Philos Trans R Soc Lond Ser B Biol Sci 361(1472):1307–1315
87. Fan Y, Cembran A, Ma S, Gao J (2013) Connecting protein conformational dynamics with catalytic function as illustrated in dihydrofolate reductase. Biochemistry 52 (12):2036–2049
88. Boehr DD, Dyson HJ, Wright PE (2008) Conformational relaxation following hydride transfer plays a limiting role in dihydrofolate reductase catalysis. Biochemistry 47 (35):9227–9233
89. Li L, Falzone CJ, Wright PE, Benkovic SJ (1992) Functional role of a mobile loop of Escherichia coli DHFR in transition-state stabilization. Biochemistry 31(34):7826–7833
90. Gekko K, Kunori Y, Takeuchi H, Ichihara S, Kodama M (1994) Point mutations at glycine-121 of E. coli DHFR-important roles of a flexible loop in the stability and function. *J.* Biochemist 116:34–41
91. Cameron CE, Benkovic SJ (1997) Evidence for a functional role of the dynamics of glycine-121 of E. coli DHFR obtained from kinetic analysis of a site-directed mutant. Biochemistry 36:15792–15800
92. Venkitakrishnan RP, Zaborowski E, McElheny D, Benkovic SJ, Dyson HJ, Wright PE (2004) Conformational changes in he active site loops of dihydrofolate reductase during the catalytic cycle. Biochemistry 43:16046
93. Boehr DD, Schnell JR, McElheny D, Bae SH, Duggan BM, Benkovic SJ, Dyson HJ, Wright PE (2013) A distal mutation perturbs dynamic amino acid networks in dihydrofolate reductase. Biochemistry 52 (27):4605–4619
94. Mauldin RV, Sapienza PJ, Petit CM, Lee AL (2012) Structure and dynamics of the G121V dihydrofolate reductase mutant: lessons from a transition-state inhibitor complex. PLoS One 7(3):e33252
95. Miller GP, Benkovic SJ (1998) Deletion of a highly motional residue affects formation of the Michaelis complex for E. coli DHFR. Biochemistry 37:6327–6335
96. Miller GP, Wahnon DC, Benkovic SJ (2001) Interloop contacts modulate ligand cycling during catalysis by Escherichia coli dihydrofolate reductase. Biochemistry 40(4):867–875
97. Behiry EM, Luk LY, Matthews SM, Loveridge EJ, Allemann RK (2014) Role of the occluded conformation in bacterial dihydrofolate reductases. Biochemistry 53 (29):4761–4768
98. Abdizadeh H, Tamer YT, Acar O, Toprak E, Atilgan AR, Atilgan C (2017) Increased substrate affinity in the Escherichia coli L28R dihydrofolate reductase mutant causes trimethoprim resistance. Phys Chem Phys 19 (18):11416–11428

99. Watney JB, Agarwal PK, Hammes-Schiffer S (2003) Effect of mutation on enzyme motion in dihydrofolate reductase. J Am Chem Soc 125:3745–3750
100. Hammes-Schiffer S, Benkovic SJ (2006) Relating protein motion to catalysis. Annu Rev Biochem 75:519–541
101. Wong KF, Selzer T, Benkovic SJ, Hammes-Schiffer S (2005) Impact of distal mutations on the network of coupled motions correlated to hydride transfer in dihydrofolate reductase. Proc Natl Acad Sci U S A 102 (19):6807–6812
102. Mauldin RV, Carroll MJ, Lee AL (2009) Dynamic dysfunction in dihydrofolate reductase results from antifolate drug binding: modulation of dynamics within a structural state. Structure 17(3):386–394

Chapter 13

Investigating Conformational Dynamics and Allostery in the p53 DNA-Binding Domain Using Molecular Simulations

Elena Papaleo

Abstract

The p53 tumor suppressor is a multifaceted context-dependent protein, which is involved in multiple cellular pathways, with the ability to either keep the cells alive or to kill them through mechanisms such as apoptosis. To complicate this picture, cancer cells that express mutant p53 becomes addicted to the mutant activity, so that the mutant variant features a myriad of gain-of-function activities, opening different venues for therapy. This makes essential to think outside the box and apply new approaches to the study of p53 structure–(mis)function relationship to find new critical components of its pathway or to understand how known parts are interconnected, compete, or cooperate. In this context, I will here illustrate how to integrate different computational methods to the identification of possible allosteric effects transmitted from the DNA binding interface of p53 to regions for cofactor recruitment. The protocol can be extended to any other cases of study. Indeed, it does not necessarily apply only to the study of DNA-induced effects, but more broadly to the investigation of long-range effects induced by a biological partner that binds to a biomolecule of interest.

Key words p53, DNA-binding domain, Transcription factor, Molecular dynamics, Protein structure network, Allostery, Structural communication, Metadynamics

1 Introduction

To appreciate the protocol illustrated here, a general introduction to p53 complexity is needed. Indeed, the pathways regulated by the p53 tumor suppressor are extremely complex. Even if p53 has been under the radar in the last 40 years, the mechanisms in which it is involved are still elusive at the molecular and atom level [1, 2]. P53 was originally discovered as an oncogene that is overexpressed in cancer to realize then that it is one of the most important tumor suppressors and it was named as the guardian of the human genome [3].

In recent years, it was revised as the "guardian of homeostatis" [4]. Indeed, recently, the attention is turned again to the fact that in

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_13,

certain biological contexts, p53 supports cell survival, even if in this case the beneficiaries are the cancer cells [4]. A similar dual role has been emerging in cancer for many other genes and processes, suggesting that we need a more sophisticated understanding of what these genes do at different cancer stages and in different contexts.

Indeed, p53 is known for its role in multiple cellular pathways, such as response to DNA damage or various cellular stresses, cell cycle arrest, senescence, autophagy, and cell death [1]. P53 is also responsible for maintaining homeostasis by repairing or eliminating cells with a damaged genome [4]. P53 has even more multifaceted functions, and it is also crucial for cell survival by promoting autophagy and metabolism in starvation [1]. P53 signaling pathways are overall cell type- and context-specific [2].

P53 is mostly known as a transcriptional activator of several genes through the recognition and binding of specific DNA sequences [2, 5]. In normal cells, p53 is detectable at very low levels, whereas it is post-translationally modified and stabilized in response to stimuli such as DNA damage, ribosomal or metabolic stress, and other alterations [2, 6]. P53 can then activate the transcription of multiple genes that determine the cell fate toward a survival or a death response [1]. P53 not only can initiate autophagy as a prosurvival mechanism but it is also tightly regulated by autophagy itself, which downregulate p53 functions to prevent cell damage [7]. In general, p53 can be controlled at multiple levels. For example, p53 is involved in an elegant feedback loop in which the protein can signal its destruction via the activation of Mdm2 (murine double minute 2) to restore normal conditions [1]. Mdm2 is an E3 ubiquitin ligase, which ubiquitinates p53 and targets it for proteasomal degradation [8].

In light of its manifold functions, P53 has often been referred to as the guardian of the human genome. Indeed, it monitors and orchestrates the activities or slow down certain processes to maintain a properly functioning environment (i.e., the cell) [1]. In this sense, P53 signaling pathways are overall highly cell type- and context-specific. It becomes thus crucial to understand how, where, and when it is activated and regulated in fine details.

P53 is the gene more frequently mutated in human cancers [9]. In contrast to many other tumor suppressors, the most common alterations of p53 in cancer are missense mutations that can result in the loss of transcriptional activity or in gain-of-function (GOF) that triggers aggressive phenotypes. Indeed, missense mutations account for approximately 75% of all p53 alteration in human cancers [10]. The fact that most p53 alterations in tumors are missense mutations suggests that cancer cells expressing mutant p53 have an advantage over the deletion of p53 [11].

The GOF of mutant p53 could be, in principle, achieved in various ways, i.e., promoting the expression of different target

genes or through the binding to different biological partners, reshaping the network of p53 protein–protein interactions, and introducing or silencing post-translational modification sites or a combination of the mechanisms mentioned above. The emerging mechanisms by which mutant p53 exhibits its GOF when it comes to different protein-protein interactions are: (a) the formation of complexes with other proteins that modify their activities (such as p63 or p73); (b) the interaction with other transcription factors (mainly VDR, SREBP, and Est2) that allows regulating promoters that are generally not under wild-type p53 control; and (c) the remodeling of chromatin through interactions with chromatin-remodeling proteins (such as SWI-SNF CRC and Pontin) or inducing chromatin regulatory gene expression (for example MLL1, MLL2, and MOZ). These mechanisms are not mutually exclusive in the onset of different cancers and are also likely to be context-dependent [1].

Mutant p53 proteins have been for a long time expected to be "undruggable" but recent studies suggest the opposite, as recapitulated in a comprehensive review article [12]. These studies provide important proofs of concept that it is possible to rescue the structural mutants of p53. However, several efforts are still required in the direction of cancer therapy and personalized medicine. In this context, the identification of the complex mechanisms that trigger mutant p53 GOF activities becomes essential to provide tailored solutions for new treatments to undermine mutant p53 activities.

It is also essential to understand in detail the spectrum of protein–protein interactions and how the partners of interaction modify the p53 structure and dynamics in wild-type and mutant p53 to properly understand the regulatory mechanism of p53 stability due to the interactions with ubiquitinating enzymes and chaperones. Indeed, the stabilization of mutant p53 is a prerequisite for its oncogenic GOF phenotype [1].

To make the scenario even more complicated, we have to consider that although it is true that a large number of human cancers feature p53 mutations or deletions, there is a large number of other alterations that could indirectly impact on the p53 pathways. An example is provided by the amplification of its negative regulators [13] that can modify p53 function and structure.

As an ideal master regulator of homeostasis, p53 features an "antagonistic and paradoxical bifunctionality" a term coined in studies of biological circuits and recently translated to p53 [4]. Indeed, as stated above, it exerts opposing effects on the cell, including prosurvival activities that might sound contradictory with its canonical pro-apoptotic functions, as well as it can have opposing effects on cell migration, metabolism, and differentiation [4]. This is typical of a cancer gene with a dual role. P53 can indeed regulate the expression of genes exerting diametrically opposite

effects on the same process and can thus be envisioned as characterized by a dynamic "sliding scale" of functions that go from canonical tumor suppressor to common oncogenic properties [4].

The molecular mechanisms underpinning these opposing responses induced by p53 have been linked to the diversity of the DNA response elements, to a different spectrum of protein conformational changes and/or different chromatin configuration [4]. In this context, p53 could be described as a transcriptional "super hub" that dictates cell homeostasis, and ultimately decides the cell fate by governing other secondary hubs in a tightly orchestrated manner [4].

A major challenge is to decipher the mechanisms behind its hub role and how p53 selects which hubs to engage and how its preferences can be modulated. An intriguing hypothesis is that the paradoxical effects exerted by p53 could be related to differences in their protein conformation [4, 14].

1.1 Cytosolic and Mitochondrial Functions of p53

P53 is also known for its transcription-independent functions [15–17] related, for example, to mitochondrial apoptosis [16, 18]. Indeed, p53 mutants with defects in transcription are still capable of inducing apoptosis [19]. p53 can translocate to the mitochondrial outer membrane in response to DNA damage. Here, p53 binds to Bak/Bax to promote Bak/Bax oligomerization [20]. As a consequence, mitochondrial outer membrane permeabilization (MOMP) is promoted, along with cytochrome c release, caspase activation, and consequent apoptosis [21]. P53 can physically interact with pro- and anti-apoptotic members of the Bcl-2 family, i.e., not only Bak but also Bcl-2 and Bcl-xL, for example.

Different mechanisms could retain p53 in the cytosol and prevent its mitochondrial translocation to restrict apoptosis under normal conditions. Bcl-2, Bcl-xL and, Mcl-1 are known to directly sequester the cytosolic p53 [22]. Moreover, K63-linked ubiquitination is associated with protein trafficking, and the cytosolic pool of p53 is ubiquitinated through the K63 linkage, but such modification was not detected for the mitochondrial p53. Screening a panel of E3 ligases, TRAF6 emerged as critical to control p53 mitochondrial translocation. TRAF6 indeed triggers p53 K63-linked ubiquitination in the cytoplasm, and it can reduce the interaction between p53 and Mcl-1/Bak preventing localization at the mitochondria and mediating activation of Bak. This mechanism is prevented in genotoxic stress condition, in which TRAF6 can also move to the nucleus where mediates the ubiquitination of p53 and promotes, in this case, its acetylation and gene expression induction of genes for cell survival under stress conditions [23].

Recently, another mechanism of p53-mediated transcription-independent functions has been proposed [24] that cytoplasmatic p53 can stimulate the accumulation of Ca2+ ions within the endoplasmic reticulum (ER) by physically interacting with the ATPase

Ca2+ transporting fast twitch 1 (ATP2A or SERCA). As a consequence, the efficiency of the transfer of the calcium ions between ER and mitochondria increases, as well as the propensity to apoptosis.

It becomes now essential to link all these emerging mechanisms. Indeed, for example, anti-apoptotic members of the Bcl-2 protein family can also localize at the ER and modulate Ca2+ homeostasis, but their role in the p53-SERCA-mediated process is unknown, along with the structural domain of p53 for p53-SERCA interaction [25].

1.2 p53 Protein Architecture and DNA-Binding Domain (DBD)

P53 forms a homotetramer with a dimer-of-dimers topology, where each monomer accounts for 393 amino acids and include multiple domains. In particular, a p53 monomer is composed of: an intrinsically disordered N-terminal transactivation domain (TAD, residues 1-42), a proline (Pro)-rich region with multiple copies of the PXXP sequence (residues 61-94), a core DNA-binding domain (DBD, residues 101-292), a tetramerization domain (324-355) connected via a flexible linker, and an intrinsically disordered C-terminal regulatory domain (356-393) [12, 26]. P53 modular structure is typical of signaling proteins, and it provides conformational plasticity to adapt to the interaction with a myriad of different partners and be modulated by a diverse range of PTMs [27].

A recent review article focused on the property and function of the tetramer [12], whereas this protocol focuses on the DNA-binding domain in the context of conformational ensemble and long-range communication, as well as transcription-dependent and -independent functions.

The p53 DBD folds into an immunoglobulin-like β-sandwich architecture with an extended DNA-binding surface (Fig. 1), which is formed by a loop-sheet-helix motif (including loop L1, F113 to T123) and two large loops (L2, i.e., K164-C176 and L3, i.e., M237-P250) that are held together by zinc coordination [29]. The L1 loop can adopt an extended conformation and interacts directly with DNA via lysine (Lys) 120 [28]. The L1 loop conformations can also explore recessed conformations without direct DNA contacts [28] (Fig. 1).

Compelling evidence suggests that p53 DBDs evolved to be only marginally stable, and there is a clear correlation between their thermodynamic stability and the corresponding optimum temperature of the organism of origin [30].

Most cancer-associated mutations are located in the DBD [10]. All but seven residues of p53 have been the target of at least one mutation in human cancer [10, 31, 32].

The effects at the structural and functional levels of p53 cancer mutations, as stated above, can be very different. Some of them are likely to remove important DNA interaction sites, and other can

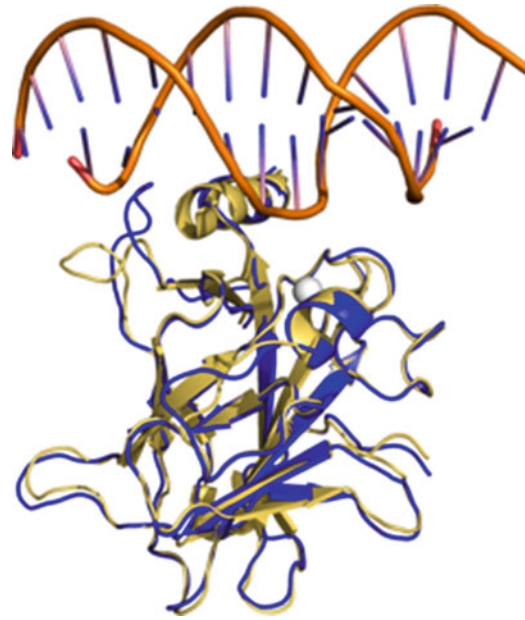

Fig. 1 p53 DBD in complex with DNA. The PDB entry 3Q06 [28] is used to illustrate the L1 in extended (blue, Chain B) and recessed (gold yellow, Chain A) states. The zinc ion is depicted as a grey sphere, the p53 DBD and DNA as cartoons in *Pymol*

perturb the structure of the DBD with consequent effects on its stability [12].

Nevertheless, a detailed investigation in the context of p53 conformational ensemble and its interfaces for cofactor recruitment is still missing for the mutant variants. We cannot rule out, for some of them, that apart from the major well-known local effects elicited by the mutations, more complex and long-range mechanisms are in act. Moreover, some of the "structural mutations" might still induce local changes that impair the p53 functions such as DNA binding or protein-protein interactions in the folded state. Even in these cases, more detailed studies in light of transcription-dependent and -independent functions need to be carried out to reach a complete overview on the effects of the p53 mutations.

1.3 Computational Framework

Several recent computational studies have been carried out to model the full-length p53 and its tetramer form and its interaction with DNA [33–36], which are beyond the scope of this protocol, which focuses on the study of long-range effects and the structural ensemble of the DBD. Moreover, we refer to the recent review by Saha et al. [37] for more details on the computational studies of p53.

Several molecular dynamics (MD) studies have explored the local conformational changes of the p53 DBD in the presence and

absence of the DNA [37–41] and identified loops L1 and L3 as the most critical regions undergoing conformational changes (*see* **Notes 1** and **2**). In other studies, the local consequences of phosphorylation have also been addressed, such as the phosphorylation at S269 and S215, which also feature more distal effects, reducing the affinity and specificity for DNA [42].

It is nowadays well established that despite the static view that X-crystallography or the average ensembles of tens of conformation that are deposited upon NMR structural determination, proteins are highly dynamic entities that can undergo multiple conformational changes. Those changes can take place also on a broad range of time scales [43–46]. They can account for small structural changes in side-chain dihedrals and conformations or involve more pronounced changes in loop conformations or even concerted motions of large regions or domains of a protein structure [46]. The different states might be important for biological functions, and they can also be observed with different populations and kinetics in both unbound/unmodified and bound/modified states of a protein.

Of particular interest to understand the function of proteins so complex as the super-hub p53 are those conformational changes that are promoted long-range and that could unveil allosteric mechanisms [46–49], i.e., those changes that occur at sites distal from the modification or ligand-binding site. Allostery can manifest in the form of both large conformational changes [50] but also subtle localized changes in protein dynamics or structure [51].

In this view, it becomes crucial to understand how a biological partner can exert its effect over long distances in p53, as well as where the distal site interested by the distal communication is and what is its function. Moreover, in the view of pre-existing minor populated states of biological relevance, it is also important to describe with accuracy if any of these conformational changes induced long range are pre-existing in the free protein in solution and how the population shift occurs.

An accurate understanding at the atom level could be achieved by all-atom explicit solvent (*see* **Note 1**) MD simulations [52–54] also coupled to NMR or other biophysical data that accounts for the structural propensity over different time scales [45, 55–57]. Indeed, all-atom MD, especially when coupled with enhanced sampling techniques [58–60] provide information on protein dynamics on timescales that go from the femto to the milliseconds. In several applications of these methods, we are witnessing high accuracy on the estimates of the different conformational states of proteins and how mutations or post-translational modifications can affect them [14, 61–65] and we finally have the tools to achieve a detailed view of the free energy landscape associated with protein conformational transitions (*see* **Note 3**).

Moreover, many methods for analyses of MD conformational ensembles provide information on paths of long-range communications [46, 54, 66–74], such as the ones based on network theory (*see* **Note 4**). Recently, their robustness to different force field descriptions for MD has been shown in proteins of different size and fold [75, 76].

Despite p53 DBD importance, few studies have been devoted to unraveling the long-range communication in p53 DBD in its free, modified or DNA-bound state. However, in recent years, progress has been made in this field, and the results are promising [14, 77–79].

We know from NMR and structural studies that p53 DBD does not appear to undergo significant conformational changes either upon binding with other proteins or DNA, but a slow conformational exchange in the proximity of the disordered N-terminal region has been identified by NMR [80]. This NMR study together with our enhanced sampling and classical MD simulations [14] also pointed out the need of including at least part of the disordered region, which is N-terminal to the p53 DBD in the structural experimental and computational studies of p53 DBD since these residues tightly modulate p53 DBD conformational propensities. In other structures of p53 DBD in complex with protein interactors [22, 81], we are also observing conformational changes that are not in the shape of the domain but often interest specific residues or loop regions. These changes should not be underestimated since it is known, in many protein systems, that rearrangements of short loops or even cascade of rearrangements in side-chain conformation can have a major impact on the protein activity and function and that allostery can occur without changes in shape [46, 82].

Recently, my group and coworkers developed a suitable platform to understand long-range effects at distal sites in transcription factors such as p53 integrating analyses of classical MD and an enhanced sampling approach, based on metadynamics [14]. Indeed, classical MD conformational ensembles can be analyzed with dimensionality reduction [83–87] or higher-order statistics techniques [88–90] to generate working hypotheses on protein regions that are distantly coupled and could be interested in long-range communication or allostery. Metadynamics [58] can be then used to test these hypotheses and to unveil with high accuracy the changes in the free energy landscape of the protein due to binding, mutations or modifications. Protein Structure Network approaches can also complement the overall picture suggesting at the atom level the structural pathways from which one site communicates with the distal one [46, 54, 66–74]. Moreover, if available, NMR-derived parameters that are probes of protein dynamics on different time scales can be used for cross-validation, as we did [14] using backbone chemical shifts of p53 DBD previously published [22]. Indeed, different and accurate chemical

shifts predictors from structural ensembles have been developed [91–94].

It is important to emphasize that unbiased MD simulations of some hundreds of nanoseconds or even microseconds are unlikely to provide enough statistical power to be used to sample conformational changes that are related to long timescale dynamics like the one revealed by the p53 NMR experiments. Indeed, in classical MD even using multiple replicate approaches, only a few transitions could be observed among different states, whereas a proper investigation would require the sampling for multiple times of the same event. When we simulated p53 DBD without the N-terminal tail, we observed higher flexibility of the S6-S7 loop in the ns timescale. This observation raises serious concerns for the usage of a DBD construct lacking the N-terminal disordered residues to study the properties of the DBD regions that can be modulated by intramolecular interaction with the tail [14]. Indeed, such fast motions were not expected. Supporting this notion, classical unbiased MD simulations of the construct including the tail (91–289) did not show any substantial differences between the DNA-bound and -unbound forms of the loop [14], indicating that different techniques for conformational sampling needed to be applied.

In our study [14], we applied the framework above to the study of p53 DBD in its unbound, DNA-bound, and phosphorylated state and on a specific region of the protein (S6-S7 loop). The work can be envisaged as a proof of concept for future applications to unveil the complexity of the p53 signaling function. Indeed, the novelty of our work is not so much about the fact that we identified a coupling between conformational rearrangements at the interface for DNA-binding and changes in a loop (S6-S7 loop, residues 207–213). Other previous works already suggested a long-range coupling in the proximity of the S6-S7 loop using classical MD only and techniques such as principal component analyses [39]. More importantly, we showed that the conformational changes in the L1 loop at the DNA-binding interface are tightly coupled to changes in the S6-S7 loop, which in turn is in proximity to the N-terminal disordered tail, which is also involved in the mechanism (Fig. 2). We showed with high accuracy that DNA modulates the conformational ensemble of the S6-S7 loop conformational ensemble. We also identified key residues that are involved in the paths of structural communication between the two distal sites and that also include the N-terminal disordered region. The proposed mechanism still holds in the context of the p53 tetramer, as shown by a comparison of the structure samples by the p53 DBD simulations with the known experimental structures of the p53 quaternary assembly [14]. Indeed, the different S6-S7 conformations fit into the tetramer without clashes and with most of the residues solved exposed and available for interaction in S6-S7 more “open” states.

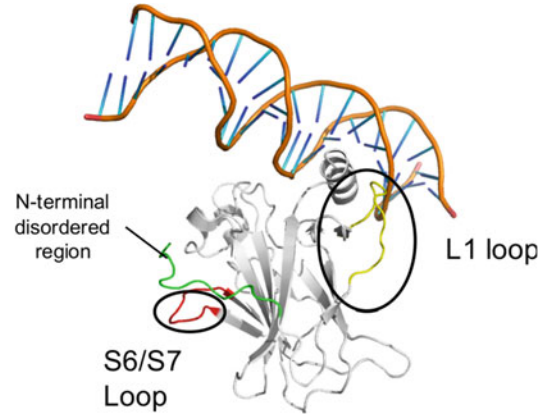

Fig. 2 p53 DBD main components of a new regulatory mechanism to select for transcription-dependent and -independent functions of p53. The PDB entry 2XWR [95] is used and the DNA has been superimposed from the PDB entry 1TSR [29]. The three main players of the mechanism that we predicted using the protocol here described are shown in green (N-terminal residues), red (S6-S7 loop), and yellow (L1 loop), respectively

What is intriguing is that these changes are at the base of different states of DBD p53 with different functional and biological implications and a population shift among the states is observed upon DNA binding. Indeed, our study proposed a new regulatory mechanism for p53 functions, which is tightly embedded in the conformational propensities of the protein structure. DNA-binding causes a population shifts toward states in which the residues of the S6-S7 loop lose most of the interaction with the disordered N-terminal region and are recruited for interactions with other protein regions that causes an "occluded" conformation of the S6-S7 loop, in turn affecting the possibility to recruit biological partners such as Ku70, Ark1, and Nb139. Our predictions suggest that these interactions are favored in the p53 DBD unbound states [14].

Of particular interest, the Ku70 interaction with p53 is necessary to release and activate Bax to initiate the apoptosis pathway [16]. It will become interesting to validate the interaction between p53 DBD and Ku70 experimentally, also considering that another crucial regulator of p53 nontranscriptional function as Bcl-xL and Bcl-2 also binds to the p53 DBD at a different region, i.e., competing with the DNA-binding interface directly [22].

Our results suggest that a conformational selection is in play so that p53 DBD is protected by interactions that are essential to mediate transcription-independent functions, such as the p53 apoptotic cytosolic functions, as long as the DNA is bound to the protein and p53 needs to act as a transcriptional activator. We also observed that Aurora kinases-mediated phosphorylation at Ser215 [96, 97] is another modulator of the p53 S6-S7 loop state and it is

likely to be only the tip of the iceberg of other complex and interconnected layers of regulation.

The long-range communication from DNA binding loops to distal regions of the protein that act as cofactor recruitment hotspots can be a broader mechanism, as we recently showed also in the DBD of the ARID family of transcription factors [98]. Also, p53 is not the unique case of known transcription factors for which transcription-independent cytosolic or membrane-associated functions have been identified [99] so that the conformational selection mechanisms proposed for p53 could be an example of a more general scenario.

More in general, the methodology illustrated here can be extended to other proteins to understand how long-range communication occurs from distal sites and which regions interested by the allosteric effects can act as cofactor recruitment interface. As an example, we recently applied the same tools to the study of the MZF1 SCAN domain [76].

2 Materials

The software and tools listed below are needed with a Linux/Unix working environment and access to HPC resources is recommended for the classical and metadynamics simulations.

1. *Pymol* (www.pymol.org) or similar software for structure visualization, superimposition, and manipulation.
2. *Gromacs* version 4 or higher to run unbiased MD simulations [100–102].
3. *Gromacs* version 4 or higher patched with *Plumed* 1.3 or 2 to perform the metadynamics simulations [103, 104].
4. *PyInteraph* [68] or *Wordom* [105] to calculate the PSN and the paths of long-range communication.
5. *Xpyder* plugin [106] for *Pymol* for visualization of the PSN results and *PyKnife* to assess convergence of the PSN properties or automatize the collection of the network [75].
6. *PRISM* [107, 108] to predict the protein-protein complexes.

3 Methods

For the sake of clarity, we will illustrate the example of the analyses for p53 DBD in its wild-type variant comparing DNA-bound and DNA-unbound states, inspired by our recent work [14]. The protocol can be translated and adjusted to any other case of study and also not necessarily to study DNA-induced effects, but more broadly long-range effects induced by a biological partner that binds to a biomolecule of interest.

3.1 Structure Selection and Preparation

It is essential in any MD study, to accurately select the starting structure for simulations. In the case of the p53 DBD, through the comparison of different PDB structures available in an MD framework, we realized the importance of including the residues of the N-terminal disordered extremity, as explained in the paragraphs above. In this example, it is suggested to use the X-ray structure of the p53 DBD in the PDB entry 2XWR (chain A, residues 91–289) [95]. This is a structure of the unbound p53 DBD including four residues from the N-terminal tail. To model the p53 DBD DNA-bound state, we could use the PDB structure 1TSR (chains B 95-289 of p53 DBD, E and F, [109]) and carry out a structural alignment of the two p53 DBDs, and then retain in the PDB file for simulations only the chain A of 2XWR and remove chain B from 1TSR. As an alternative, the missing residues could be modeled into the 1TSR structure using 2XWR as a template. We suggest using the chain B of 1TSR in the modelling since it is centrally placed with respect to the DNA molecule in the crystal structure.

To define the protonation state of the histidine, we used the *Propka* server [110], whereas the cysteine and histidine residues for Zn^{2+} coordination were kept in their unprotonated state. The N-terminus and C-terminus were also modeled at pH = 7 as positively and negatively charged moieties, respectively.

3.2 Preparatory Steps for Classical MD and Metadynamics

A force field has to be selected for the simulations, and we recommend to carefully read the literature with force field benchmarking and to use those force fields that would be more suitable as a combination to model both a protein and a DNA molecule. In our previous publication, we tested both CHARMM22-CMAP [111] or CHARMM22* [112] and CHARMM27 DNA parameters [113]. Nevertheless, MD force fields are continuously adjusted or developed so that in new applications, for example, a more recent force field could be used [114].

The systems have to be solvated in a dodecahedral box of water molecules at 150 mM NaCl using periodic boundary conditions with a minimum distance of at least 1.3 nm between the protein and the box edges. In the case of the CHARMM family of force fields, the usage of the TIP3P solvent model [115, 116] is recommended. Topology and box preparation, as well as solvation of the system, can be carried out using the *pdb2gmx, editconf,* and *genbox* tools from *Gromacs*. respectively.

Each system needs to be initially relaxed by 10,000 steps of energy minimization with the steepest descent algorithm. The optimization step was followed by 0.5 ns or one ns of solvent equilibration at 300 K while restraining the protein atomic positions using a harmonic potential. Each system can be then equilibrated to the target temperature (300 K) and pressure (1 bar) through thermalization and pressurization simulations in the

NVT or NPT ensemble, respectively for some nanoseconds. The final step before the production run is an equilibration in the NVT ensemble of at least 5 ns. We suggest to perform the productive MD simulations using LINCS algorithm [117] to constrain heavy-atom bonds, allowing for a 2 fs time-step. Long-range electrostatic interactions can be described by the Particle-mesh Ewald summation Scheme [118]. Van der Waals and Coulomb interactions can be truncated at 0.9 nm.

3.3 Classical MD Simulations

Classical MD simulations should be carried out for some hundreds of ns up to microseconds if possible, for the first exploration of distal coupling between different protein regions. They can be carried out at 300 K using, for example, a velocity-rescale thermostat [119].

We suggest evaluating the radius of gyration and the secondary structure content to ensure that they do not dramatically deviate from the corresponding values in the known experimental structures used as starting structures for the simulations. These are best practices and common sense in the field. It is also important to monitor the main-chain Root Mean Square Deviation (RMSD) of the folded regions of the protein with respect to the initial structure, and if some ns are required to reach convergence of this property, the first part of the simulation can be discarded from further analyses. In a case such as p53, where a metal ion is bound to the structure, it is also important to evaluate the distances between the metal atom and its coordinating groups over the simulation time, as well as the coordination geometry. In a case such as the one published by us, we carried out multiple MD replicates of each p53 variant to better explore the conformational space. This can be achieved, for example, using different random initial velocities associated with the atoms at the beginning of each simulation. If multiple replicates are used, the equilibrated portions of each trajectory can then be concatenated in a unique macro-trajectory for the analyses.

3.4 Dimensionality Reduction Methods

In this example, we will illustrate the usage of Principal Component Analysis (PCA) [120], but also other methods can be used for the same purpose such as higher-order statistics methods, or other metrics for ensemble comparison that have been cited in the paragraphs above. The *Gromacs*tools for PCA such as *g_covar* and *g_anaeig* can be applied for this step. The PCA of MD trajectories allows to identify the eigenvectors (also called principal components) of the mass-weighted covariance matrix of the atomic positional fluctuations, and it is generally carried out using the Cα atoms. If different variants are under comparison, it is important to concatenate the trajectories of the different system so that the comparison can be made in the same PCA subspace. With a proper MD sampling, the first three PCs generally accounts alone for more

than 50% of the fluctuations of the system. They can be used as "reaction coordinates" to generate two-dimensional (2D) probability distribution plots and identify putative conformational substates of different regions of the protein that need to be predefined on the base of the structural knowledge on the system, the question to be addressed or also a first visual exploration of the MD trajectory. It is also useful to carry out all-atom PCA for a specific region of interest upon fitting on the main-chain or CA of the protein, to better appreciate the different states that these regions can assume. The PCA results can also be mapped on the 3D structure of the protein, and they allow to identify a potential coupling between different sites, as we observed for the conformation of the DNA-binding loop L1 (and its K120 residue especially) and the distal loop S6-S7 [14].

3.5 Paths of Communications between Distal Sites

To further verify the coupling between the distal regions of interest, i.e., the L1 and S6-S7 loops in our example, we can apply graph theory to the MD structural ensembles. There are multiple solutions to achieve this goal available. Our group recently implemented a *Python* suite of tools that can handle different MD trajectory formats, i.e., *PyInteraph* [68]. We also provided a plugin for *Pymol* to handle *PyInteraph* output formats for graphical visualization of the analyses [106]. As an alternative, the *Wordom* package could be used which also allow integrating the PSN with information from correlated motions [105]. PSN approaches such as the ones provided by *PyInteraph* and *Wordom* can be used not only for calculations of shortest communication pathways between two residues but also to calculate other important properties of a network (such as hubs, connected components, cliques, etc.). We here focus only on the path analyses since it is the suitable one for the processes that this protocol aims at dissecting (i.e., communication among distal sites).

We will here enter into the details of the *PyInteraph* approach, which is currently used in our group. At first, we need to generate a PSN based on contact between the centers of mass of residues side chains using the *pyinteraph* tool of the package. A distance cutoff of 5 Å has been recently shown as the best solution for this kind of network in a benchmarking of different proteins simulated with different force fields [75]. A *Python* tool, *PyKnife,* is also available to identify the cutoffs and monitor the properties of the network with a JackKnife resampling method [75]. The distance is calculated between the center of mass of the residues side chains, expect for glycines. To obtain the PSN, it is also recommended to retain those edges that are only populated in the 20% of the simulation frames [68, 73, 121] to remove noise in the final network. For each pair of nodes of interest in the PSN graph (which could be, for example, residues of L1 and residues of the S6-S7 loop), a variant of the depth-first search algorithm is used in the current *PyInteraph*

```
#!/bin/bash
# with a working PyInteraph Installation
# pre-processed trajectory to remove PBC issues
xtc="traj.mol.ur.nojump.xtc"
#gro="sim.gro" or "model0.pdb" from the trajectory
#pdb=Protein Data Bank file - only if there are numeration issues
gro="model0.pdb"
pdb="model0.pdb"
dat="hc-graph.dat"
datfilt="hc-graph_filt.dat"

# distance cutoff of 5 Ang.
pyinteraph -s $gro -t $xtc -r $pdb -f --hc-graph hc-graph.dat  --ff-masses charmm27  --hc-co 5 --
hc-residues ALA,ILE,LEU,VAL,PHE,TRP,TYR,MET,PRO,ARG,HIS,LYS,GLU,ASP,ASN,GLN,SER,THR,CYS

#filtering and calculation of  hubs and connected components
# hub degree and cluster ID. are stored in the B-factor column of the pdb files
filter_graph -d $dat -o $datfilt -t 20.0
graph_analysis -a $datfilt -r $pdb -u -ub hubs.pdb
graph_analysis -a $datfilt -r $pdb -c -cb con_comp.pdb

#path analyses
graph_analysis -a $datfilt -r $pdb -r1 SYSTEM-120LYS -r2 SYSTEM-208ASP -l 10 -p -d

~
```

Fig. 3 A bash wrapper to run the *PyInteraph* pipeline for contact-based PSN. The tools *pyinteraph*, *filter_graph,* and *graph_analysis* need to be used sequentially. For details on each option of the command lines, we recommend to refer to the *PyInteraph* official documentation

implementation to identify the shortest paths of communication between two sites. Indeed, communication in PSN is expected to work more efficiently through the shortest paths between two distal sites [122]. The PSN analysis sheds light on the atomic details behind the distal communication and also identifies the most important components in the mechanism. In the p53 example, it allowed identifying the residues in the N-terminal tail as important in modulating the conformation of the S6-S7 loop together with the DNA binding loops. An example of a script for *PyInteraph* is provided in Fig. 3.

3.6 Metadynamics Simulations

Once the first exploration of classical MD trajectories is carried out, and the regions of interest have been identified, it is essential to design collective variables for the metadynamics step which better describe the mechanism of interest or the working hypothesis. In the case of the S6-S7 loop, we monitored over the MD simulations more than 50 parameters in terms of side-chain and backbone dihedral angles, distances between different residues and angles formed by the loop motions. This allowed us to identify which variables could better describe the different states of the loop that we wanted to explore and calculate in more details with metadynamics. It is also fundamental that for the process of interest, the slowest degrees of freedom are identified and included in the metadynamics collective variables to have a proper and accurate exploration of the free energy landscape and avoid artificial results or phenomena such as hysteresis [58].

Once the collective variables of interest are identified, multiple metadynamics-based solutions can be explored. Indeed, the methodological metadynamics-oriented research evolved so much in the last 10 years and with so many fruitful contributions that it could be now seen as a field on its own. The one here provided is just one example to give a first guidance and also suitable if the HPC computational resources available are rather limited or to run with GPU support.

The approach is the metadynamics coupled with parallel tempering (PT, i.e., a sort of replica exchange in the temperature space) [123] in the well-tempered ensemble (WTE) [124] to overcome the usage of many close temperatures in the PT simulations when systems with a large number of atoms need to be used. Indeed, it is critical in a PT-metadynamics to achieve a sufficient energy overlap between adjacent replicas and sufficient exchange rates, which need to be carefully checked on short exploratory runs before even moving further in the sampling or analyses of the results. In the WTE approach, a constant bias on the energy is added to each replica to increase the width of the energy distribution so that a suitable exchange rate is ensured even when a lower number of replicas are used and the separation in the temperature space is larger.

To give a practical example, in the case of p53 DBD, after the classical MD exploratory analyses a working hypothesis that can be generated is that upon DNA binding or phosphorylation at Ser215 the conformational state of a distal loop, i.e., the S6-S7 loop (residues 207-213), can be conformationally modulated [14]. After exploring different reaction coordinates in the available unbiased MD runs, we concluded that the combination of at least four Cα-Cα key distances between residues of the S6-S7 loop and its surroundings is descriptive of the conformations that the loop assumes: Asp208-Arg156 (CV 1), Arg158-Phe212 (CV 2), Arg209-Glu221 (CV 3), and Arg209-Glu258 (CV 4). In contrast, CVs such as the radius of gyration are not of interest for this specific process since there is no remarkable change in the shape of the molecule upon opening and closing of the loop. Such as CV could become relevant for larger conformational rearrangements when more extended and disordered loops or entire domains change their reciprocal orientation.

Once the CVs are selected, we also need a proper definition of the temperatures for each replica, which should reasonably span from low to high temperatures but where the highest temperature should not encounter the risk to unfold our protein in the simulation time needed to reach convergence. Indeed, we want to simulate a conformational change occurring in a folded protein and not its unfolding/folding mechanism. We can thus run unbiased simulations at high temperatures of at least some hundreds of ns to monitor the stability of the protein architecture and identify the

minimum highest temperature to employ. In the case of our p53 example, a good scheme could have eight replicas (at 296, 298, 300, 308, 320, 332, 345, and 358 K) where the width of the energy distribution of all the replicas was increased except for a "neutral" 298 K replica [62].

All replicas are further subjected to an additional biasing force through metadynamics in which a Gaussian of width 0.1 nm in all the CV dimensions (i.e., the four distances in our example) is deposited in the collective variable space every 4 ps with an initial height of 0.12 kcal/mol and a bias factor of 6. All these parameters, i.e., the deposition time, the initial height of the Gaussian and the bias factor could be tuned according to the process of interest. We used quite mild biases since the changes in free energy among the minor and major states were not expected to be high, and we aimed to reproduce them accurately.

In our example (Fig. 4), the simulations were run for at least 300 ns per replica, checking the evolution of the monodimensional free energy surface (FES) along each collective variable and also the evolution of two-dimensional FES over the simulation time using the *sumhills* preprocessing tool of *Plumed*. In an ideal scenario, we could interrupt the simulation when we are confident that we are sampling the changes in the collective variable space multiple times (i.e., we observe multiple events of opening/closing of the loop). This is required to achieve the sufficient statistical power. Another criterion is to verify that the FES does not change remarkably over time, and the relevant minima have been explored.

3.7 Identification of Biological Partners Recruited at the Long-Range DNA- Modulated Sites

Once the structural mechanism behind long-range communication or allostery has been unveiled, in a case as p53, it becomes crucial to give a biological rationale to it. Does this conformational change have any meaning from the biological point of view? Alternatively, it is just an unrelated event to protein function?

In the p53 case, which needs to interact with multiple partners, an obvious working hypothesis could be that the distal region acts as an interface for recruitment of other biological partners and that the conformational change can either "activate" or "inactivate" this function.

Thus, we can first retrieve the available information on p53 partners using databases where experimental (or predicted but to be taken with caution) protein-protein interactions are annotated. For example, the *I2D* database can be used since it acts as a "metaserver" integrating annotations from different sources, included the literature. The pool can always be enriched by manual annotation from recent literature or other databases. The target list can also be pruned according to the *CRAPome* definition from hits that are likely to be artifacts in proteomics [125]. Once the target proteins have been identified, it becomes crucial to retain only those for which at least one experimental structure is available in

```
HILLS HEIGHT 0.5  W_STRIDE 2000

WELLTEMPERED SIMTEMP 300 BIASFACTOR 6

PTMETAD

PRINT W_STRIDE 1000

#ALIGN_ATOMS LIST <all>


# PASSIVE CV (cv=5) : ENERGY BIAS
ENERGY
# read the energy bias from grid files
EXTERNAL NCV 1 CV 1 FILENAME grid_200_BIAS

#CV2 DIST CA-D208 CA-R156

DISTANCE LIST 1830  987 SIGMA 0.05

#CV3 DIST CA-F212 CA-R158

DISTANCE LIST 1894 1027 SIGMA 0.05

#CV4 DIST CA-R209 CA-E221

DISTANCE LIST 1842 2049 SIGMA 0.05

#CV5 DIST CA-R209 CA-E258

DISTANCE LIST 1842 2604 SIGMA 0.05

UWALL CV 2 LIMIT 2.2 KAPPA 100.0
LWALL CV 2 LIMIT 1.0 KAPPA 100.0
UWALL CV 3 LIMIT 1.8 KAPPA 100.0
LWALL CV 3 LIMIT 1.1 KAPPA 100.0
UWALL CV 4 LIMIT 3.2 KAPPA 100.0
LWALL CV 4 LIMIT 1.9 KAPPA 100.0
UWALL CV 5 LIMIT 2.6 KAPPA 100.0
LWALL CV 5 LIMIT 0.8 KAPPA 100.0

ENDMETA
```

Fig. 4 An example of the *Plumed* input file for PT/WTE metadynamics. An example is shown for inspiration and it refers to the *Plumed* 1.3 syntax that we used in our publication on p53 [14]

the PDB. Then, we can use the *PRISM* approach, also including if possible a benchmarking against known complexes of the protein of interest and other proteins to define a suitable docking energy threshold, as we did in the p53 case [14]. *PRISM* is a template-based approach to predict protein-protein interactions. *PRISM* uses a rigid-body structural comparison of target proteins to known templates of protein-protein interfaces and a refinement using flexible docking. Moreover, it is useful also to calculate the Interface Similarity Score to assess models of protein complexes [126, 127] as an additional quantitative parameter.

In our example, since the interest is to understand the function of different states of the S6-S7 loop of p53, we can use the p53 experimental structures of the DBD but also conformations of it from out MD simulations with the loop in occluded or open states (subjected to energy minimization before the *PRISM* analyses). We identified a group of p53 interactors that are selectively bound to the DBD with the S6-S7 loop in its open solvent accessible state, as stated above and that have intriguing functional implication related to transcription-independent function activities [14].

In summary, the protocol here suggested provides a rich portfolio of information that can be used for the design of experiments to validate the modeling, and to disclose the mechanisms better. To cite a few examples, the interactions predicted by *PRISM* could be validated in vitro for example using chemical shift NMR perturbation experiments on the isolated domain of the protein or other biophysical techniques suitable for studying protein complexes. Moreover, the exchange by minor and major states that are modulated by DNA, PTMs, or mutations (such as the occluded and open states of the S6-S7 loop) can be explored using NMR relaxation dispersion measurements, also estimating from the exchange rate the population of the states and directly comparing them to the calculated ones. We could also exploit the protocol to design variants that can entrap one of the two states and experimentally test our structural hypotheses.

4 Notes

1. In the case a CHARMM force field is used in *Gromacs*, it has to be paid attention to the solvent model to select. The TIP3P tailored for the CHARMM family [116] of the force field is not necessarily the recommended option in some of the releases and, it is the one labeled as "TIPS3P". The simulations can become slower, but if one is interested in solvent accessible regions of the protein, the solvation is treated in a more suitable way for the CHARMM family.
2. It is also always important to verify that there are very transient or virtually none contacts between the periodic images, a suitable tool in this context is the *g_mindist Gromacs* tool.
3. If PCA is carried out using all atoms of the system, the covariance matrix of atomic fluctuations needs to be correctly mass-weighted.
4. A *PyInteraph* mass database compatible with the force field of interest needs to be selected since different MD force fields have different mass definitions.

Acknowledgements

This work was supported by the ISCRA-CINECA HPC Grants (HP10BLFPW4 and HP10C8LO8N) and the EU-PRACE DECI project DyNet. I would like to thank Matteo Lambrughi for fruitful inputs in the writing of this protocol.

References

1. Zhang Y, Lozano G (2017) p53: multiple facets of a rubik's cube. Annu Rev Cancer Biol 1:185–201. https://doi.org/10.1146/annurev-cancerbio-050216-121926
2. Vousden KH, Prives C (2009) Blinded by the light: the growing complexity of p53. Cell 137:413–431
3. Vogelstein B, Lane D, Levine AJ (2000) Surfing the p53 network. Nature 408:307–310
4. Aylon Y, Oren M (2016) The paradox of p53: what, how, and why? Cold Spring Harb Perspect Med 6(10):a026328
5. Fischer M (2017) Census and evaluation of p53 target genes. Oncogene 36:3943–3956
6. Luo Q, Beaver JM, Liu Y et al (2017) Dynamics of p53: a master decider of cell fate. Genes 8:66
7. White E (2016) Autophagy and p53. Cold Spring Harb Perspect Med 6:1–10
8. Pant V, Lozano G (2014) Limiting the power of p53 through the ubiquitin proteasome pathway. Genes Dev 28:1739–1751
9. Kandoth C, McLellan MD, Vandin F et al (2013) Mutational landscape and significance across 12 major cancer types. Nature 502:333–339
10. Leroy B, Anderson M, Soussi T (2014) TP53 mutations in human cancer: database reassessment and prospects for the next decade. Hum Mutat 35:672–688
11. Brosh R, Rotter V (2009) When mutants gain new powers: news from the mutant p53 field. Nat Rev Cancer 9:701–713
12. Joerger AC, Fersht AR (2016) The p53 pathway: origins, inactivation in cancer, and emerging therapeutic approaches. Annu Rev Biochem 85:375–404. https://doi.org/10.1146/annurev-biochem-060815-014710
13. Wasylishen AR, Lozano G (2016) Attenuating the p53 pathway in human cancers: many means to the same end. Cold Spring Harb Perspect Med 6(8):a026211
14. Lambrughi M, De Gioia L, Gervasio FL et al (2016) DNA-binding protects p53 from interactions with cofactors involved in transcription-independent functions. Nucleic Acids Res 44:9096–9109
15. Green DR, Kroemer G (2009) Cytoplasmic functions of the tumour suppressor p53. Nature 458:1127–1130
16. Speidel D (2010) Transcription-independent p53 apoptosis: an alternative route to death. Trends Cell Biol 20:14–24
17. Tasdemir E, Maiuri MC, Galluzzi L et al (2008) Regulation of autophagy by cytoplasmic p53. Nat Cell Biol 10:676–687
18. Vaseva AV, Moll UM (2009) The mitochondrial p53 pathway. Biochim Biophys Acta Bioenerg 1787:414–420
19. Kokontis JM, Wagner AJ, O'Leary M et al (2001) A transcriptional activation function of p53 is dispensable for and inhibitory of its apoptotic function. Oncogene 20:659–668
20. Leu JI-J, Dumont P, Hafey M et al (2004) Mitochondrial p53 activates Bak and causes disruption of a Bak–Mcl1 complex. Nat Cell Biol 6:443–450
21. Chipuk JE, Green DR (2008) How do BCL-2 proteins induce mitochondrial outer membrane permeabilization? Trends Cell Biol 18:157–164
22. Follis AV, Llambi F, Ou L et al (2014) The DNA-binding domain mediates both nuclear and cytosolic functions of p53. Nat Struct Mol Biol 21:535–543
23. Zhang X, Li CF, Zhang L et al (2016) TRAF6 restricts p53 mitochondrial translocation, apoptosis, and tumor suppression. Mol Cell 64:803–814
24. Giorgi C, Bonora M, Sorrentino G et al (2015) p53 at the endoplasmic reticulum regulates apoptosis in a Ca^{2+} –dependent manner. Proc Natl Acad Sci 112:1779–1784
25. Kroemer G, Bravo-San Pedro JM, Galluzzi L (2015) Novel function of cytoplasmic p53 at the interface between mitochondria and the endoplasmic reticulum. Cell Death Dis 6: e1698
26. Joerger AC, Fersht AR (2007) Structural biology of the tumor suppressor p53 and

cancer-associated mutants. Adv Cancer Res 97:1–23

27. Dai C, Gu W (2010) P53 post-translational modification: deregulated in tumorigenesis. Trends Mol Med 16:528–536
28. Petty TJ, Emamzadah S, Costantino L et al (2011) An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. EMBO J 30:2167–2176
29. Cho Y, Gorina S, Jeffrey PD et al (1994) Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science 265:346–355
30. Khoo KH, Andreeva A, Fersht AR (2009) Adaptive evolution of p53 thermodynamic stability. J Mol Biol 393:161–175
31. Soussi T, Curie M (2014) The TP53 gene network in a postgenomic era. Hum Mutat 35(6):641–642
32. Soussi T, Wiman KG (2015) TP53: an oncogene in disguise. Cell Death Differ 22:1239–1249
33. Abramo MD, Besker N, Desideri A et al (2015) The p53 tetramer shows an induced-fit interaction of the C-terminal domain with the DNA-binding domain. Oncogene 35:3272–3281
34. Chillemi G, Davidovich P, D'Abramo M et al (2013) Molecular dynamics of the full-length p53 monomer. Cell Cycle 12:3098–3108
35. Terakawa T, Takada S (2015) p53 dynamics upon response element recognition explored by molecular simulations. Sci Rep 5:17107
36. Demir Ö, Ieong PU, Amaro RE (2017) Full-length p53 tetramer bound to DNA and its quaternary dynamics. Oncogene 36:1451–1460
37. Saha T, Kar RK, Sa G (2015) Structural and sequential context of p53: a review of experimental and theoretical evidence. Prog Biophys Mol Biol 117(2-3):250–263
38. Lu Q, Tan YH, Luo R (2007) Molecular dynamics simulations of p53 DNA-binding domain. J Phys Chem B 111:11538–11545
39. Lukman S, Lane DP, Verma CS (2013) Mapping the structural and dynamical features of multiple p53 DNA binding domains: insights into loop 1 intrinsic dynamics. PLoS One 8:e80221
40. Pan Y, Nussinov R (2010) Lysine120 interactions with p53 response elements can allosterically direct p53 organization. PLoS Comput Biol 6:e1000878
41. Pan Y (2008) p53-induced DNA bending: the interplay between p53. J Phys Chem B 112:6716–6724
42. Fraser JA, Madhumalar A, Blackburn E et al (2010) A novel p53 phosphorylation site within the MDM2 ubiquitination signal II. A model in which phosphorylation at SER 269 induces a mutant. J Biol Chem 285:37773–37786
43. Henzler-Wildman KA, Lei M, Thai V et al (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature 450:913–916
44. Tang C, Schwieters CD, Clore GM (2007) Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. Nature 449:1078–1082
45. Baldwin AJ, Kay LE (2009) NMR spectroscopy brings invisible protein states into focus. Nat Chem Biol 5:808–814
46. Papaleo E, Saladino G, Lambrughi M et al (2016) The role of protein loops and linkers in conformational dynamics and allostery. Chem Rev 116:6391–6423
47. Cui Q, Karplus M (2008) Allostery and cooperativity revisited. Protein Sci 17:1295–1307
48. Tsai C-J, Nussinov R (2014) A unified view of "how allostery works". PLoS Comput Biol 10:e1003394
49. Ribeiro AAST, Ortiz V (2016) A chemical perspective on allostery. Chem Rev 116:6488–6502
50. Bray D, Duke T (2004) Conformational spread: the propagation of allosteric states in large multiprotein complexes. Annu Rev Biophys Biomol Struct 33:53–73
51. Nussinov R, Tsai C-J (2015) Allostery without a conformational change? Revisiting the paradigm. Curr Opin Struct Biol 30:17–24
52. Feher VA, Durrant JD, Van Wart AT et al (2014) Computational approaches to mapping allosteric pathways. Curr Opin Struct Biol 25:98–103
53. Dror RO, Dirks RM, Grossman JP et al (2012) Biomolecular simulation: a computational microscope for molecular biology. Annu Rev Biophys 41:429–452
54. Papaleo E (2015) Integrating atomistic molecular dynamics simulations, experiments, and network analysis to study protein dynamics: strength in unity. Front Mol Biosci 2:28
55. Torchia DA (2015) NMR studies of dynamic biomolecular conformational ensembles. Prog Nucl Magn Reson Spectrosc 84–85:14–32
56. O'Rourke KF, Gorman SD, Boehr DD (2016) Biophysical and computational methods to analyze amino acid interaction networks in proteins. Comput Struct Biotechnol J 14:245–251

57. Fraser JS, Clarkson MW, Degnan SC et al (2009) Hidden alternative structures of proline isomerase essential for catalysis. Nature 462:669–673
58. Laio A, Gervasio FL (2008) Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Rep Prog Phys 71:126601
59. Abrams C, Bussi G (2013) Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. Entropy 16:163–199
60. Luitz M, Bomblies R, Ostermeir K et al (2015) Exploring biomolecular dynamics and interactions using advanced sampling methods. J Phys Condens Matter 27:323101
61. Marino KA, Sutto L, Gervasio FL (2015) The effect of a widespread cancer-causing mutation on the inactive to active dynamics of the B-Raf kinase. J Am Chem Soc 137:5280–5283
62. Sutto L, Gervasio FL (2013) Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. Proc Natl Acad Sci 110:10616–10621
63. Papaleo E, Sutto L, Gervasio FL et al (2014) Conformational changes and free energies in a proline isomerase. J Chem Theory Comput 10:4169–4174
64. Wang Y, Papaleo E, Lindorff-Larsen K (2016) Mapping transiently formed and sparsely populated conformations on a complex energy landscape. elife 5:e17505
65. Palazzesi F, Barducci A, Tollinger M et al (2013) The allosteric communication pathways in KIX domain of CBP. Proc Natl Acad Sci U S A 110:14237–14242
66. Csermely P, Korcsmáros T, Kiss HJM et al (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol Ther 138:333–408
67. Angelova K, Felline A, Lee M et al (2011) Conserved amino acids participate in the structure networks deputed to intramolecular communication in the lutropin receptor. Cell Mol Life Sci 68:1227–1239
68. Tiberti M, Invernizzi G, Lambrughi M et al (2014) PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. J Chem Inf Model 54:1537–1551
69. Papaleo E, Lindorff-larsen K, De Gioia L (2012) Paths of long-range communication in the E2 enzymes of family 3: a molecular dynamics investigation. Phys Chem Chem Phys 14:12515–12525
70. Whitley MJ, Lee AL (2009) Frameworks for understanding long-range intra-protein communication. Curr Protein Pept Sci 10:116–127
71. Di Paola L, De Ruvo M, Paci P et al (2013) Protein contact networks: an emerging paradigm in chemistry. Chem Rev 113:1598–1613
72. Ribeiro AAST, Ortiz V (2015) Energy propagation and network energetic coupling in proteins. J Phys Chem A 119:1835–1846
73. Papaleo E, Renzetti G, Tiberti M (2012) Mechanisms of intramolecular communication in a hyperthermophilic acylaminoacyl peptidase: a molecular dynamics investigation. PLoS One 7:e35686
74. Di Paola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48
75. Salamanca Viloria J, Allega MF, Lambrughi M et al (2016) An optimal distance cutoff for contact-based protein structure networks using side chain center of masses. Sci Rep 7:2838
76. Nygaard M, Terkelsen T, Olsen AV et al (2016) The mutational landscape of the oncogenic MZF1 SCAN domain in cancer. Front Mol Biosci 3:78
77. Ng JWK, Lama D, Lukman S et al (2015) R248Q mutation—beyond p53-DNA binding. Proteins 83:2240–2250
78. Thayer KM, Quinn TR (2016) p53 R175H hydrophobic patch and H-bond reorganization observed by MD simulation. Biopolymers 105:176–185
79. Calhoun S, Daggett V (2011) Structural effects of the L145Q, V157F, and R282W cancer-associated mutations in the p53 DNA-binding core domain. Biochemistry 50:5345–5353
80. Bista M, Freund SM, Fersht AR (2012) Domain-domain interactions in full-length p53 and a specific DNA complex probed by methyl NMR spectroscopy. Proc Natl Acad Sci U S A 109:15752–15756
81. Bethuyne J, De Gieter S, Zwaenepoel O et al (2014) A nanobody modulates the p53 transcriptional program without perturbing its functional architecture. Nucleic Acids Res 42:12928–12938
82. Tsai CJ, del Sol A, Nussinov R (2008) Allostery: absence of a change in shape does not imply that allostery is not at play. J Mol Biol 378:1–11

83. Lange OF, Grubmüller H (2006) Can principal components yield a dimension reduced description of protein dynamics on long time scales? J Phys Chem B 110:22842–22852
84. Daidone I, Amadei A (2012) Essential dynamics: foundation and applications. Wiley Interdiscip Rev Comput Mol Sci 2:762–770
85. Lindorff-Larsen K, Ferkinghoff-Borg J (2009) Similarity measures for protein ensembles. PLoS One 4:e4203
86. Tiberti M, Papaleo E, Bengtsen T et al (2015) ENCORE: software for quantitative ensemble comparison. PLoS Comput Biol 11: e1004415
87. Martín-García F, Papaleo E, Gomez-Puertas P et al (2015) Comparing molecular dynamics force fields in the essential subspace. PLoS One 10:e0121114
88. Ramanathan A, Savol AJ, Langmead CJ et al (2011) Discovering conformational sub-states relevant to protein function. PLoS One 6:e15827
89. Savol AJ, Burger VM, Agarwal PK et al (2011) QAARM: quasi-anharmonic autoregressive model reveals molecular recognition pathways in ubiquitin. Bioinformatics 27: i52–i60
90. Wriggers W, Stafford KA, Shan Y et al (2009) Automated event detection and activity monitoring in long molecular dynamics simulations. J Chem Theory Comput 5:2595–2605
91. Kohlhoff KJ, Robustelli P, Cavalli A et al (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. J Am Chem Soc 131:13894–13895
92. Sahakyan AB, Vranken WF, Cavalli A et al (2011) Structure-based prediction of methyl chemical shifts in proteins. J Biomol NMR 50:331–346
93. Li DW, Brüschweiler R (2012) PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. J Biomol NMR 54:257–265
94. Li D, Brüschweiler R (2015) PPM_One: a static protein structure based chemical shift predictor. J Biomol NMR 62:403–409
95. Natan E, Baloglu C, Pagel K et al (2011) Interaction of the p53 DNA-binding domain with its n-terminal extension modulates the stability of the p53 tetramer. J Mol Biol 409:358–368
96. Liu Q, Kaneko S, Yang L et al (2004) Aurora-A abrogation of p53 DNA binding and transactivation activity by phosphorylation of serine 215. J Biol Chem 279:52175–52182
97. Fraser JA, Vojtesek B, Hupp TR (2010) A novel p53 phosphorylation site within the MDM2 ubiquitination signal: I. phosphorylation at SER269 in vivo is linked to inactivation of p53 function. J Biol Chem 285:37762–37772
98. Invernizzi G, Tiberti M, Lambrughi M et al (2014) Communication routes in ARID domains between distal residues in helix 5 and the DNA-binding loops. PLoS Comput Biol 10:e1003744
99. Lim CP, Cao X (2006) Structure, function, and regulation of STAT proteins. Mol BioSyst 2:536
100. Abraham MJ, Murtola T, Schulz R et al (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2:19–25
101. Hess B, Kutzner C, van der Spoel D et al (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4:435–447
102. Pronk S, Páll S, Schulz R et al (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29:845–854
103. Bonomi M, Branduardi D, Bussi G et al (2009) PLUMED: a portable plugin for free-energy calculations with molecular dynamics. Comput Phys Commun 180:1961–1972
104. Tribello GA, Bonomi M, Branduardi D et al (2014) PLUMED 2: new feathers for an old bird. Comput Phys Commun 185:604–613
105. Seeber M, Felline A, Raimondi F et al (2011) Wordom: a user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces. J Comput Chem 32:1183–1194
106. Pasi M, Tiberti M, Arrigoni A et al (2012) xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. J Chem Inf Model 279:1–6
107. Baspinar A, Cukuroglu E, Nussinov R et al (2014) PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. Nucleic Acids Res 42:W285–W289
108. Tuncbag N, Gursoy A, Nussinov R et al (2011) Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat Protoc 6:1341–1354
109. Cho Y, Gorina S, Jeffrey PD et al (1994) Crystal structure of p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science 265(5170):346–355

110. Dolinsky TJ, Czodrowski P, Li H et al (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res 35:522–525
111. Mackerell AD, Feig M, Brooks CL (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 25:1400–1415
112. Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? Biophys J 100:L47–L49
113. Mackerell AD, Banavali NK (2000) All atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. J Comput Chem 21:105–120
114. Parsons DW, Li M, Zhang X et al (2011) The genetic landscape of the childhood cancer medulloblastoma. Science 331:435–439
115. Jorgensen WL, Chandrasekhar J, Madura JD et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926
116. Bjelkmar P, Larsson P, Cuendet MA et al (2010) Implementation of the CHARMM force field in GROMACS: analysis of protein stability effects from correction Maps, virtual interaction sites, and water models. J Chem Theory Comput 6:459–466
117. Hess B, Bekker H, Berendsen H et al (1993) LINCS: a linear constraint solver for molecular simulations. J Comput Chem 12:1463–1472
118. Essmann U, Perera L, Berkowitz ML et al (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577
119. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. J Chem Phys 126:014101
120. Amadei A, Linssen AB, Berendsen HJ (1993) Essential dynamics of proteins. Proteins 17:412–425
121. Papaleo E, Pasi M, Tiberti M et al (2011) Molecular dynamics of mesophilic-like mutants of a cold-adapted enzyme: insights into distal effects induced by the mutations. PLoS One 6:e24214
122. Atilgan AR, Akan P, Baysal C (2004) Small-world communication of residues and significance for protein dynamics. Biophys J 86:85–91
123. Bussi G, Gervasio FL, Laio A et al (2006) Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. J Am Chem Soc 128:13435–13441
124. Bonomi M, Barducci A, Parrinello M (2009) Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. J Comput Chem 30:1615–1621
125. Mellacheruvu D, Wright Z, Couzens AL et al (2013) The CRAPome: a contaminant repository for affinity purification – mass spectrometry data. Nat Methods 10:730–736
126. Kuzu G, Gursoy A, Nussinov R et al (2014) Exploiting conformational ensembles in modeling protein-protein interactions on the proteome scale. J Proteome Res 12:2641–2653
127. Gao M, Skolnick J (2011) New benchmark metrics for protein-protein docking methods. Proteins 79:1623–1634

Chapter 14

Molecular Dynamics Simulation Techniques as Tools in Drug Discovery and Pharmacology: A Focus on Allosteric Drugs

Chiara Bianca Maria Platania and Claudio Bucolo

Abstract

Allosteric drugs are ligands that when bound to an allosteric site modify the conformational state of the pharmacological target, leading then to a modification of functional response upon binding of the endogenous ligand. Pharmacological targets are defined as biological entities, to which a ligand/drug binds and leads to a functional effect. Pharmacological targets can be proteins or nucleic acids. Computational approaches such as molecular dynamics (MD) sped up discovery and identification of allosteric binding sites and allosteric ligands. Classical all-atom and hybrid classical/quantum MD simulations can be generalized as simulation techniques aimed at analysis of atoms and molecular motion. Main limitations of MD simulations are related to high computational costs, that in turn limit the conformational sampling of biological systems. Indeed, other techniques have been developed to overcome limitations of MD, such as enhanced sampling MD simulations. In this chapter, classical MD and enhanced sampling MD simulations will be described, along with their application to drug discovery, with a focus on allosteric drugs.

Key words Allosteric drugs, Molecular dynamics, Drug discovery, Pharmacology

1 Introduction

1.1 Allosteric Drugs

Pharmacological target function could be modulated by drugs that bind to an orthosteric pocket or an allosteric pocket. Orthosteric drugs competitively bind to the same pocket of the endogenous ligand, working as competitive agonists or antagonists; while allosteric drugs bind to a pocket different from the orthosteric one, leading to conformational changes and positive or negative modulation of receptor activity, upon binding of the endogenous ligand. Positive allosteric modulators (PAM) increase the receptor functional response or decrease the EC50 of agonists or endogenous ligand (increased endogenous ligand efficiency) (Fig. 1). The negative allosteric modulators (NAM) decrease receptor functional response and increase agonists or endogenous ligand EC50 (decreased endogenous ligand efficiency) (Fig. 1). Russinov &

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_14,

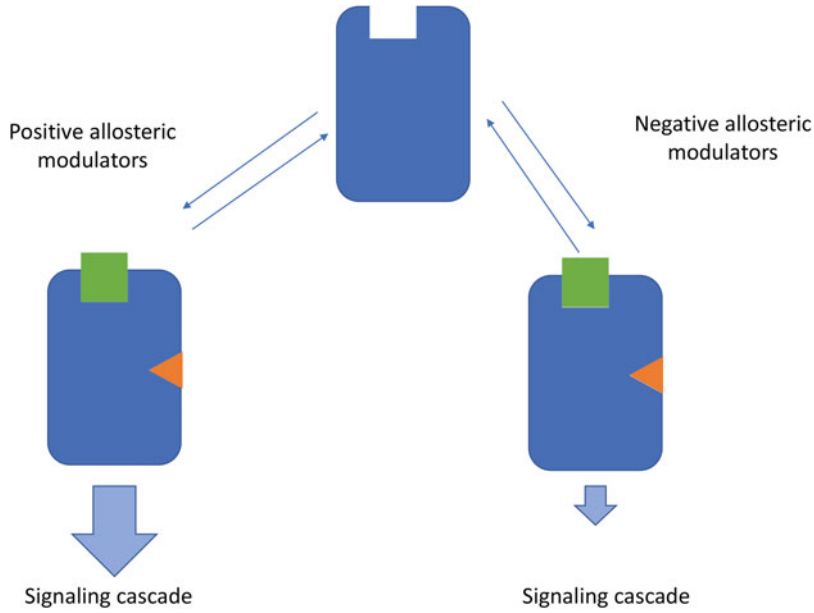

Fig. 1 Allosteric drugs

Tsai in 2013 reviewed the impact of allosteric drugs in the field of drug discovery; this review highlighted that most of the allosteric drugs work through non-covalent mechanisms; e.g., positive allosteric modulators of the γ-aminobutyric acid receptor A (GABA-A) [1]. GABA-A is an ubiquitous receptor in the central nervous system (CNS) and orthosteric ligands of GABA-A are not used in therapy, because of their poor pharmacodynamic and safety profile. In fact, GABA-A competitive agonists, such as muscimol, are potent psychoactive drugs (sedative-hypnotic, depressant, and allucinogen), while GABA-A antagonists such as bicucullin and picrotoxin are convulsivant drugs. Bicucullin is used as pharmacological tool, while picrotoxin was used for treatment of barbiturate acute toxicity. First developed positive allosteric modulators of GABA-A were barbiturates, used as anxiolytic and hypnotic drugs, but due to their narrow therapeutic index (low safety), they were largely substituted by benzodiazepines, that bear a good therapeutic index and are also positive allosteric modulators of GABA-A [2] (*see* **Note 1**). Similarly to the GABA-A receptor, the impact of allosteric modulation of the N-methyl-D-aspartate receptors (NMDARs) would be huge, since NMDAR drugs showed poor clinical outcome and serious side-effects [3]. Additionally, allosteric ligands have been found to modulate the activity of G protein-coupled receptors, such as the cannabinoid type 1 receptor [4].

1.2 Classical All-Atom MD, Elastic Network Model, and Enhanced Sampling MD

Protein conformational energy landscapes are characterized by a series of local minima (Fig. 1) [5], and transition between one minimum to another corresponds to a high-energy transition state. In a classical MD simulation, several minimization steps are required to reach a local minimum; after that MD production runs are started. However, given the characteristics of their energy landscape, proteins can be trapped in non-relevant local minima, even during long molecular dynamics simulations [6, 7]. These findings do not mean that classical MD simulations are meaningless, but MD simulations need to be addressed to specific aims, such as: minimization of protein models (e.g., membrane proteins) [8], normal mode analysis of protein motions [9], Molecular Mechanics—Poisson Boltzmann Surface Area calculations [9, 10], and protein contact network analysis [11]. The protein contact network (PCN) formalism could be addressed to identification of hotspot residues involved, for example, in ligand binding, protein–protein interactions, or allosteric modulation of protein motion [12–15]. However, PCN analysis of classical all-atom MD frames, for example, belonging to an energy minimum (equilibrated MD), can be characterized by non-relevant correlations of PCN metrics parameters vs time (e.g., degree, average shortest path, closeness centrality, betweenness centrality etc.,) [9]. Thus, PCN analysis can be applied to two different protein conformations or to protein normal modes generated with Gaussian Network Model (GNM) [16] or Anisotropic Network Model (ANM) [11]. GNM and ANM are both based on Elastic Network Model (ENM) formalism, in which the macromolecule is treated as a network (coarse-grained model), where nodes are atoms, nucleotides or amino acids, that are linked by edges treated as harmonic restraints on displacement from the structure at equilibrium. ENM provides a fast calculation of low-frequency normal modes [17]. Indeed, considering that the ENM results are comparable to MD simulation analysis, ENM approaches can provide valuable structural information at low computational costs (minutes vs days) [17]. However, enhanced sampling MD simulations are commonly used when large protein motions with high-energy barriers (Fig. 2) are studied [18]. Additionally, enhanced sampling approaches are particularly suited for search and identification of putative allosteric pockets that can be hidden in an apo crystal structure [19]. In this perspective, large-scale biased MD simulations can explore low-populated conformations, where hidden allosteric pockets might be unveiled [20]. Accelerated MD simulations are characterized by application of a positive potential in the protein potential energy surface, leading to overcoming of high-energy barriers; with this approach, novel allosteric pockets were identified in the IL-1R1 receptor [21]. A technique closely related to accelerated MD is metadynamics, where a bias potential enforces the exploration of novel unexplored conformations [22], this technique recently

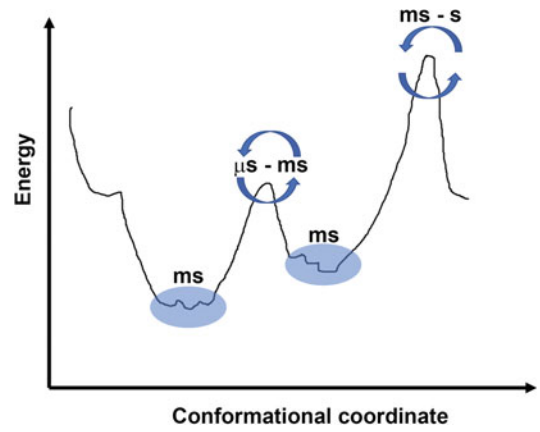

Fig. 2 Protein conformational energy landscape and simulation time-frame

characterized the binding of a negative allosteric modulator of the purinergic P2X3 receptor [22]. Moreover, the Monte Carlo simulation technique (Forced-Biased Metropolis Monte Carlo simulated annealing) was aimed at identification of allosteric binding site of pregnenolone, an allosteric modulator of the cannabinoid 1 receptor (CB_1) [4].

2 Methods and Applications

2.1 Case Studies "Protein Contact Network" Analysis

2.1.1 Anti-VEGF Agents

PCN analysis has been applied in order to study equilibrium structures of VEGF-A, anti-VEGF agents (bevacizumab, ranibizumab, and aflibercept), and related VEGF-A/anti-VEGF complexes [9]. The correlation analysis between topological descriptors and time for three independent replicas of simulated complexes was carried out. The dGsolv (delta of solvation free energy) did not change over the time, given that the correlation dGsolv vs. time was not significant. Few topological descriptors showed correlations with time and/or with each other; for example, the average shortest path (asp) positively correlated with time. Clustering of PCN was also carried out and then complexes were represented as functional models. After partitioning of the protein contact network into clusters, the structure of the complexes was represented as functional modules [9]. The VEGFA dimer was divided into two clusters that were found to be highly interconnected (Fig. 3). VEGFR1d2_R2d3 (aflibercept binding domain)/VEGFA was divided into four clusters; this cluster partitioning of the aflibercept/VEGFA complex revealed a conserved network for VEGFA and two distinct domains corresponding to R1d2 and R2d3, forming long-range interactions with VEGFA. The partition into four clusters of Fab-bevacizumab/VEGFA and ranibizumab/VEGFA revealed for bound VEGFA a different cluster patterning,

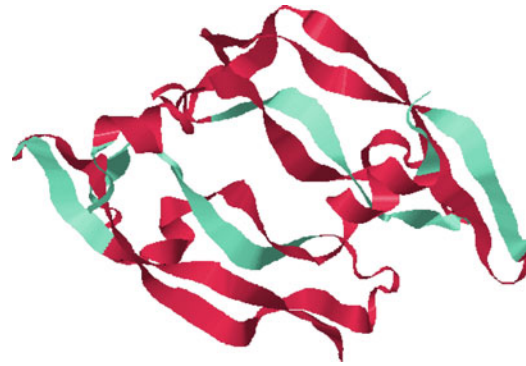

Fig. 3 Cluster partitioning of the VEGFA dimer (red and cyan clusters)

compared to unbound VEGFA. Furthermore, some whiskers projected from VEGFA to Fab-bevacizumab and ranibizumab modules. PCN analysis has shown a greater number of long-range interactions between ranibizumab and VEGFA in comparison to the Fab-bevacizumab/VEGFA complex. In conclusion, PCN analysis was found to be a helpful integrative tool for analysis of MD simulations of protein-protein complexes [9].

2.1.2 P2X7 Receptor

The P2X7 receptor is a purinergic homotrimeric channel receptor. PCN analysis was carried out in order to highlight hot-spot allosteric residues involved in channel opening [11]. PCN was not carried out on MD snapshots; because channel opening was simulated by access to the ANM Pathway server, which uses a two-state anisotropic network model [23]. ANMpathway generates snapshots (.pdb files) of transition between two structural endpoints, a two-state potential is built from two elastic network models (ENMs), which are representative of the endpoint structures: closed (apo-P2X7) and open conformation (ATP-bound P2X7).

PCN analysis has been applied after conversion of each snapshot into an indirect, unweighted graph, whose nodes are the α-carbons and edges, linking two residues, describe long-range interactions [11]. Calculation of centrality metrics (closeness and betweenness) was used for identification and characterization of residues belonging to orthosteric or to allosteric sites. Correlation analysis of closeness and betweenness centrality upon channel opening revealed that both parameters anticorrelated with residue (node) displacement, an index of protein flexibility. During P2X7 channel opening, the betweenness centrality correlated less with displacement, compared to the closeness centrality, suggesting that residues with high betweenness are involved in signal transmission upon channel opening, indeed high betweenness centrality residues were considered as allosteric residues [11]. In particular, high closeness residues were located in the core of the extracellular

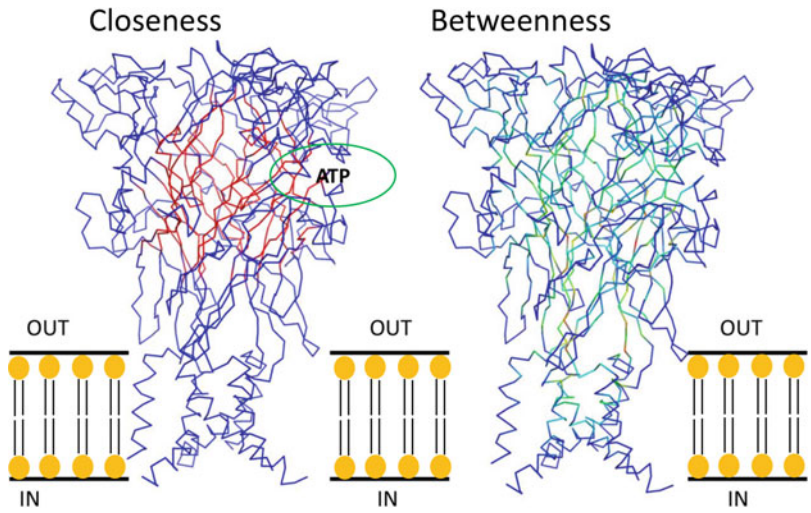

Fig. 4 Closeness centrality and betweenness centrality values mapped onto the P2X7 structure model [11]

region, while high betweenness residues (Fig. 4) were in the inner part of the channel, including the extracellular and transmembrane domains of the trimer. The allosteric region, as defined by protein contact network analysis, corresponded to a pocket recognized by the SiteMap tool of Schrödinger, as well as to data reported by Karasawa & Kawate (2016) [24].

2.2 Case Studies "Enhanced Sampling Approaches"

As described above, PCN analysis is a simple still valid method to identify and validate topological and functional properties of protein residues. However, other molecular modeling approaches, with high demanding computational resources, can unveil details of allosteric pockets, especially if few structural information is available, such as open and closed (or active and inactive) receptor conformations.

2.2.1 Monte Carlo Simulations

In 2014, Vallée et. al's paper unveiled the pharmacological properties and binding site of pregnenolone, a precursor of a series of neurosteroids, that are synthetized in the brain and exert several neuromodulatory functions (Fig. 5).

For a long-time, pregnenolone was considered as an inactive precursor, till findings of Vallée et al. [4]. In particular, pregnenolone levels were found to be increased in the rat brain after tetrahydrocannabinol (THC) treatment; therefore, a negative feedback for THC-induced pregnenolone synthesis was proven. In fact, pregnenolone counteracted hypolocomotion, hypothermia, catalepsy, and analgesia induced by THC administration. Additionally, pregnenolone decreased food intake and memory impairment induced by THC. Pregnenolone was able to decrease cannabinoid agonist self-administration in CD1 mice. Given that pregnenolone has been shown to decrease levels of p-ERK1/2^{MAPK} and

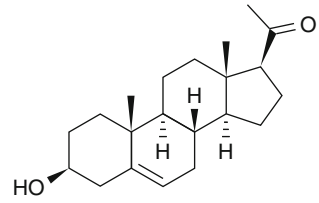

Fig. 5 Pregnenolone (pregn-5-en-3β-ol-20-one) structure

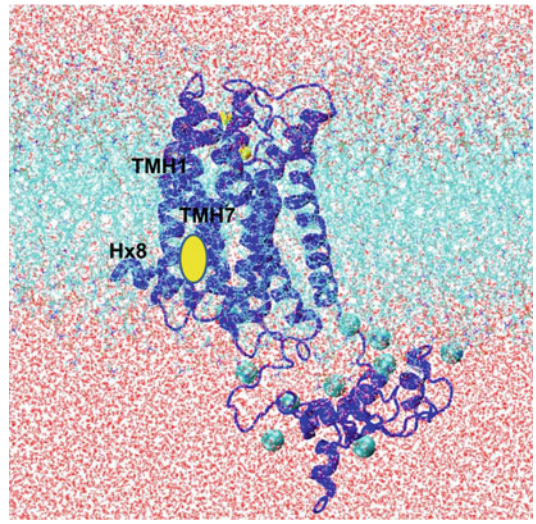

Fig. 6 Binding pocket of pregnenolone in the CB1 receptor

mitochondrial respiration, negative allosteric modulation of CB1 was investigated and identification of binding pocket was carried out with Forced-Biased Metropolis Monte Carlo simulated annealing calculation [4]. The binding pocket of pregnenolone was localized at the CB1 receptor lipid interface and this pocket faces the TMH1/TMH7/Hx8 region of the receptor (Fig. 6). Because E1.49 was identified as the key residue for pregnenolone by Monte Carlo simulations, the pregnenolone pocket was validated using a mutant hCB1 receptor where an aspartate residue in the helix 1 (E1.49) was mutated. Then, pregnenolone lost its effects as NAM of CB1, when cells expressing hCB1 mutant receptors were treated with THC [4] (*see* **Note 2**).

2.2.2 *Metadynamics*

In the recent paper of Wang et al. [22] a new allosteric pocket was identified in the purinergic channel receptor P2X3, which is different from the one identified in the P2X7 receptor [24]. Figure 7 shows the superimposition of the P2X3 receptor bound to a negative allosteric ligand, AF-219 (magenta cartoon and sticks), with the P2X7 receptor model [11]. In Fig. 7, the P2X7 receptor is represented as a ribbon and residues are colored as function of

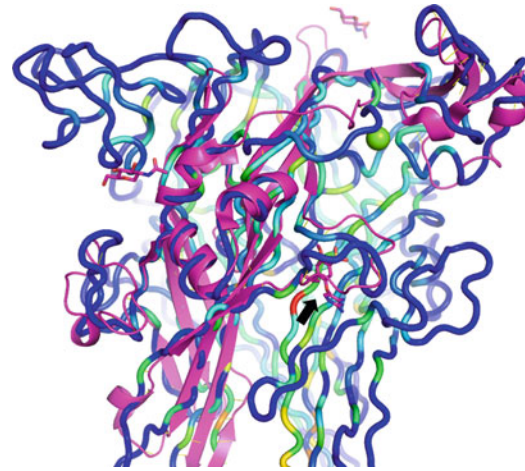

Fig. 7 Allosteric binding pocket (black arrow) in the P2X3 receptor (magenta cartoon) and structure superimposition with P2X7 receptor (ribbon, residues colored on basis of betweennes centrality)

betweenness values, as calculated and reported in Fig. 4. Hot-spot allosteric residues of the P2X7 receptor are located in the extracellular and intracellular region of the channel cavity (Fig. 4). Figure 7 shows that whenever high-betweenness residues would be conserved in the P2X3 and P2X7 receptors, the AF-219 pocket could be identified also in the P2X7 receptor. Therefore, further studies on the P2X7 receptor will be carried out on the basis of experimental evidence, recently reported by Wang et al. [22].

Wang et al. reported that only the x-ray structure of the P2X3/AF-219 complex was solved, but authors have also proven with metadynamics that the NAM AF-353 binds to P2X3, with a pose similar to AF-219 [22]. In conclusion, the recent experimental data on allosteric modulation of P2X3 will burst research of allosteric modulators of P2X receptors, considering that allosteric pocket of the P2X3 receptor should not be excluded also in the P2X7 receptor (Fig. 7). However, further studies with Monte Carlo simulation or metadynamics should be carried out on other P2X receptors (*see* **Note 3**).

3 Notes

1. Allosteric ligands are valuable pharmacological tools, and approved allosteric drugs changed the landscape of CNS diseases treatments.
2. Identification of the allosteric pockets and, indeed, design and discovery of the allosteric ligands is challenging.

3. Protein contact network analysis, along with classical MD simulations or enhanced sampling methods, can burst the identification of allosteric pockets and design of allosteric modulators.

References

1. Nussinov R, Tsai C-J (2013) Allostery in disease and in drug discovery. Cell 153:293–305. https://doi.org/10.1016/j.cell.2013.03.034
2. Olsen RW (2018) GABAA receptor: positive and negative allosteric modulators. Neuropharmacology 136:10–22. https://doi.org/10.1016/j.neuropharm.2018.01.036
3. Perszyk R, Katzman BM, Kusumoto H, Kell SA, Epplin MP, Tahirovic YA, Moore RL, Menaldino D, Burger P, Liotta DC, Traynelis SF (2018) An NMDAR positive and negative allosteric modulator series share a binding site and are interconverted by methyl groups. elife 7:e34711. https://doi.org/10.7554/eLife.34711
4. Vallee M, Vitiello S, Bellocchio L, Hebert-Chatelain E, Monlezun S, Martin-Garcia E, Kasanetz F, Baillie GL, Panin F, Cathala A, Roullot-Lacarriere V, Fabre S, Hurst DP, Lynch DL, Shore DM, Deroche-Gamonet V, Spampinato U, Revest J-M, Maldonado R, Reggio PH, Ross RA, Marsicano G, Piazza PV (2014) Pregnenolone can protect the brain from cannabis intoxication. Science 343:94–98. https://doi.org/10.1126/science.1243985
5. Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 48:545–600. https://doi.org/10.1146/annurev.physchem.48.1.545
6. Bergonzo C, Henriksen NM, Roe DR, Swails JM, Roitberg AE, Cheatham TE 3rd (2014) Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide. J Chem Theory Comput 10:492–499. https://doi.org/10.1021/ct400862k
7. Marsili S, Signorini GF, Chelli R, Marchi M, Procacci P (2010) ORAC: a molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level. J Comput Chem 31:1106–1116. https://doi.org/10.1002/jcc.21388
8. Platania CBM, Salomone S, Leggio GM, Drago F, Bucolo C (2012) Homology modeling of dopamine D2 and D3 receptors: molecular dynamics refinement and docking evaluation. PLoS One 7:e44316. https://doi.org/10.1371/journal.pone.0044316
9. Platania CBM, Di Paola L, Leggio GM, Romano GL, Drago F, Salomone S, Bucolo C (2015) Molecular features of interaction between VEGFA and anti-angiogenic drugs used in retinal diseases: a computational approach. Front Pharmacol 6:248. https://doi.org/10.3389/fphar.2015.00248
10. Corrada D, Colombo G (2013) Energetic and dynamic aspects of the affinity maturation process: characterizing improved variants from the bevacizumab antibody with molecular simulations. J Chem Inf Model 53:2937–2950. https://doi.org/10.1021/ci400416e
11. Platania CBM, Giurdanella G, Di Paola L, Leggio GM, Drago F, Salomone S, Bucolo C (2017) P2X7 receptor antagonism: implications in diabetic retinopathy. Biochem Pharmacol 138:130–139. https://doi.org/10.1016/j.bcp.2017.05.001
12. De Ruvo M, Giuliani A, Paci P, Santoni D, Di Paola L (2012) Shedding light on protein-ligand binding by graph theory: the topological nature of allostery. Biophys Chem 165–166:21–29. https://doi.org/10.1016/j.bpc.2012.03.001
13. Di Paola L, Giuliani A (2015) Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 31:43–48. https://doi.org/10.1016/j.sbi.2015.03.001
14. Di Paola L, Platania CBM, Oliva G, Setola R, Pascucci F, Giuliani A (2015) Characterization of protein-protein interfaces through a protein contact network approach. Front Bioeng Biotechnol 3:170. https://doi.org/10.3389/fbioe.2015.00170
15. Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci U S A 103:8577–8582. https://doi.org/10.1073/pnas.0601602103
16. Hu G, Di Paola L, Liang Z, Giuliani A (2017) Comparative study of elastic network model and protein contact network for protein complexes: the hemoglobin case. Biomed Res Int 2017:2483264. https://doi.org/10.1155/2017/2483264
17. Doruker P, Atilgan AR, Bahar I (2000) Dynamics of proteins predicted by molecular

dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. Proteins 40:512–524

18. Bernardi RC, Melo MCR, Schulten K (2015) Enhanced sampling techniques in molecular dynamics simulations of biological systems. Biochim Biophys Acta 1850:872–877. https://doi.org/10.1016/j.bbagen.2014.10.019
19. Lu S, Ji M, Ni D, Zhang J (2018) Discovery of hidden allosteric sites as novel targets for allosteric drug design. Drug Discov Today 23:359–365. https://doi.org/10.1016/j.drudis.2017.10.001
20. Dror RO, Pan AC, Arlow DH, Borhani DW, Maragakis P, Shan Y, Xu H, Shaw DE (2011) Pathway and mechanism of drug binding to G-protein-coupled receptors. Proc Natl Acad Sci U S A 108:13118–13123. https://doi.org/10.1073/pnas.1104614108
21. Yang C-Y (2015) Identification of potential small molecule allosteric modulator sites on IL-1R1 ectodomain using accelerated conformational sampling method. PLoS One 10:e0118671. https://doi.org/10.1371/journal.pone.0118671
22. Wang J, Wang Y, Cui W-W, Huang Y, Yang Y, Liu Y, Zhao W-S, Cheng X-Y, Sun W-S, Cao P, Zhu MX, Wang R, Hattori M, Yu Y (2018) Druggable negative allosteric site of P2X3 receptors. Proc Natl Acad Sci U S A 115:4939–4944. https://doi.org/10.1073/pnas.1800907115
23. Das A, Gur M, Cheng MH, Jo S, Bahar I, Roux B (2014) Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model. PLoS Comput Biol 10:e1003521. https://doi.org/10.1371/journal.pcbi.1003521
24. Karasawa A, Kawate T (2016) Structural basis for subtype-specific inhibition of the P2X7 receptor. elife 5:e22153. https://doi.org/10.7554/eLife.22153

Chapter 15

Cooperativity and Allostery in RNA Systems

Alla Peselis and Alexander Serganov

Abstract

Allostery is among the most basic biological principles employed by biological macromolecules to achieve a biologically active state in response to chemical cues. Although initially used to describe the impact of small molecules on the conformation and activity of protein enzymes, the definition of this term has been significantly broadened to describe long-range conformational change of macromolecules in response to small or large effectors. Such a broad definition could be applied to RNA molecules, which do not typically serve as protein-free cellular enzymes but fold and form macromolecular assemblies with the help of various ligand molecules, including ions and proteins. Ligand-induced allosteric changes in RNA molecules are often accompanied by cooperative interactions between RNA and its ligand, thus streamlining the folding and assembly pathways. This chapter provides an overview of the interplay between cooperativity and allostery in RNA systems and outlines methods to study these two biological principles.

Key words RNA cooperativity, Thermodynamics, Conformational change

1 Introduction

The vast majority of functional macromolecules are a result of primary sequences properly folded into secondary, tertiary, and quaternary structures. In order to achieve functional conformations, biopolymers such as proteins and nucleic acids must proceed through a folding pathway(s), which could initially yield a large pool of partially folded conformations and nonfunctional states. Often times, in order to adopt the active state, macromolecules have to sample a broad number of possible states and finalize their conformation by making a proper set of intermolecular interactions or binding to specific ligands [1]. Interactions with ligands can depend on their presence in the cells and environment, and, therefore, can impose a regulatory effect on the function of a macromolecule. A structural change of a molecule in response to ligand binding has been defined as an allosteric modulation.

Like proteins, some RNA needs to fold into three-dimensional structures and undergo structural transitions to carry out biological function [2]. RNA does not have many functional groups and is,

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8_15,

therefore, a poorer catalyst than proteins. As a result, only a handful of protein-free RNA-based enzymes have evolved [3], which most often cleave the RNA backbone without the regulatory input of other molecules. Interestingly, these enzymes, called ribozymes, do not typically display allosteric changes as a result of binding an effector molecule outside of the active site; rather, they are able to act independently of interacting partners or ligands. However, conformational transitions accompany the assembly of practically all RNA-ligand complexes, and since they involve ligand-induced changes, they can be broadly defined as allosteric changes. In contrast to proteins, RNA folding involves many structural adaptations upon binding to Mg^{2+} cations, which are essential for neutralizing the negative charge of phosphate moieties in the RNA backbone and which make possible formation of secondary and tertiary RNA structures. Each of these cation-mediated folding transitions could also be considered as allosteric modulations of the RNA structure.

Many researchers focus their efforts on studying the interplay between RNA structure transitions and ligand binding, providing an enormous number of examples of allosteric modulations in RNA (reviewed in [4, 5]). However, one aspect closely related to allostery remains poorly understood due to methodological difficulties and the complexity of studied systems. This aspect pertains to coordination of molecular events in order to overcome the time constraints imposed by sampling the vast number of various conformations and quickly narrowing down the options to the final functional state. Such a coordination can involve interdependence of molecular events, a phenomenon observed in many biological systems and defined by the thermodynamic term "cooperativity" [5]. In RNA systems, cooperative binding is often based on allosteric changes introduced by initial ligand binding and is therefore of high importance for understanding the functions of many RNA-containing assemblies.

Cooperativity and allostery have long been known to researchers as the most basic biological principles. Cooperativity was first described as the change in ligand-binding affinity observed upon the binding of another, identical ligand (reviewed in [6]). One of the earliest examples of this phenomenon was the tetrameric hemoglobin molecules [7] composed of four identical monomers each capable of binding to a single oxygen molecule. Upon the binding of the first oxygen molecule, the protein undergoes an allosteric change that increases the binding affinity for oxygen in the other monomers, allowing each subsequent oxygen molecule to bind more easily, represented by a sigmoidal binding curve. Over the years, the meaning of the term cooperativity has broadened from the classical description of ligands binding to multiple sites of an oligomeric protein, as in hemoglobin, to interdependent

conformational transitions in folding and function of proteins and nucleic acids, resulting in the formation of multicomponent complexes [8].

In this review, we primarily focus on various manifestations of cooperativity and accompanying allosteric modulations in RNA. We provide a brief explanation of thermodynamics related to cooperativity, and outline some helpful methods for interrogating RNA folding and cooperativity. We further provide examples of cooperativity and allostery in RNA, including formation of secondary and tertiary interactions, multiple binding of small ligands, and assembly of RNA-protein complexes.

2 Methods of Studying Cooperativity and Allosteric Modulations

2.1 Thermodynamic Basis of Cooperativity

Despite the variety of instances of cooperativity, they all have a thermodynamic quality, which actually defines the term cooperativity. The simplest system to explain cooperativity is a binding reaction with 3 components, the RNA (R), ligand (L), and either an identical or different ligand molecule (M), to yield a ternary complex R:L:M from individual components (Fig. 1a) [8]. The reaction could proceed through two pathways involving initial interactions between R and L followed by addition of M, or with initial binding of R to M followed by interactions with L. Cooperativity is observed when the binding of L and M to R depends on each other. The thermodynamic construction that illustrates this principle is known as a thermodynamic cycle. Each binary binding reaction is described by its own equilibrium constants K_1 and K_2, while formation of ternary complexes gives additional constants K_3 and K_4. All reactions have their own free energy terms ΔG°_1, ΔG°_2, ΔG°_3, and ΔG°_4. The overall thermodynamics of forming the ternary complex does not depend on the assembly pathway, therefore $K_1K_3 = K_2K_4$ and $\Delta G^\circ_1 + \Delta G^\circ_3 = \Delta G^\circ_2 + \Delta G^\circ_4$. If binding of L does not stimulate binding of M to the R:L complex, each binding event is independent, the free energy of the formation of binary and ternary complexes is identical, $\Delta G^\circ_1 = \Delta G^\circ_4$, and the system does not display cooperativity. However, if binding of L enhances binding of M to the R:L complex, $\Delta G^\circ_1 > \Delta G^\circ_4$, and the system has positive cooperative binding. If binding of M is hindered by binding of L, then $\Delta G^\circ_1 < \Delta G^\circ_4$ and the binding of L and M has negative cooperativity. Similar statements can be made for vertical reactions in Fig. 1a. Thermodynamically, the extent of cooperativity could be expressed by the coupling free energy, $\Delta\Delta G = \Delta G^\circ_1 - \Delta G^\circ_4 = \Delta G^\circ_2 - \Delta G^\circ_3$. The system has positive and negative cooperativity if $\Delta\Delta G > 0$ and $\Delta\Delta G < 0$, respectively. Thus, in order to dissect cooperativity in the biological system that involves interactions between RNA and small ligands such as

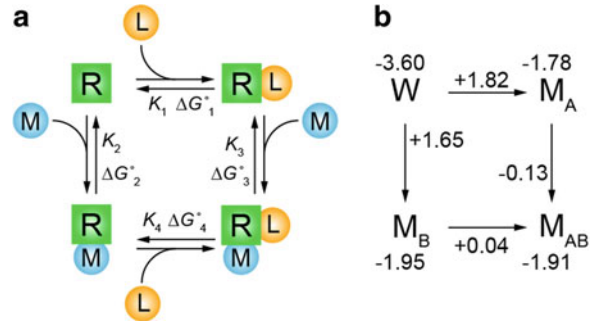

Fig. 1 Thermodynamic cycles illustrating cooperativity. (**a**) Hypothetical thermodynamic cycle for binding RNA (R) to two ligands (L and M) through two different pathways [8]. Formation of each complex is described by the equilibrium constant (K) and free energy (ΔG°). (**b**) Thermodynamic cycle for probing hydrogen bonding in a GCA triloop hairpin [9]. All values are in kcal/mol. Corners depict a wild-type (W), single mutants (M_A and M_B) and a double-mutant (M_{AB}) constructs of the molecule with thermodynamic stability values measured by UV melting studies. Values next to arrows indicate free energy of structural transition associated with each mutation and calculated by subtraction of values from appropriate corners of the box

metabolites or ions, one has to determine free energies of the binding reactions and calculate $\Delta\Delta G$.

The same principle of breaking a complex reaction into a series of smaller, experimentally accessible reactions whose sum leads to a final reaction product could be applied to dissect the mechanisms of cooperativity in areas such as the assembly of RNA-protein complexes and RNA folding. Dissection of mechanisms of cooperativity often requires special tricks, typically aimed at disrupting one interaction and probing the thermodynamic worth of another. This approach is applicable to study large multicomponent systems as well as to reveal fine structural features, for example determining whether two functional groups are involved in hydrogen bonding that stabilizes a secondary structure element.

Illustration of the approach is given by a thermodynamic box in Fig. 1b, which shows four related molecules, each located at a corner of the box [10]. The first molecule is the wild-type (WT) construct (depicted as W), the two molecules adjacent to the WT are the single mutants of the functional groups of interest, M_A and M_B, and across from the WT is the double mutant (M_{AB}). If A and B indeed form a hydrogen bond to each other, thermodynamic stability of the double mutant should decrease and mutations in the functional groups should show strong interdependence or positive cooperativity. In practical terms, the conclusion can be drawn after experimental determination of thermodynamic stability of all four molecules (values at each corner of the box) by, for example, the UV-melting method. These values are then used to

calculate the free energy of transition to each state (values along the arrows) and subsequent calculations of $\Delta\Delta G$. The results of these calculations indicate that the first change along either pathway has significant thermodynamic worth, the second change has essentially no thermodynamic worth, and the coupling free energy indicates positive cooperativity. Thus, these data suggest that functional groups A and B participate in the same hydrogen bonding, since their individual mutations remove the hydrogen bond, and once the bond is removed it can no longer be used by the second functional group.

2.2 Methods to Determine Cooperativity and Detect Allosteric Modulations

Over 100 years of research has expanded methodology to study cooperativity and allostery, from measuring hemoglobin saturation with oxygen as a function of the partial pressure of oxygen to many more approaches, including detailed structural studies. However, spectroscopy techniques remain the oldest and most popular methods for determining nucleic acid thermodynamics because of simplicity and low cost. In RNA spectroscopy, the signal is proportional to the advance of the reaction as the molecules undergo structural transitions brought on by either ligand binding or changes in temperature. Conformational rearrangements are accompanied by changes in intrinsic UV absorbance, which depend on the formation of base pairs [11]. Another method for spectroscopically detecting conformational changes is through fluorescence [12, 13], for example, by internally incorporating fluorescent nucleotide analog 2-aminopurine [14]. As an RNA molecule folds upon ligand binding, the attached chromophore is exposed to a different microenvironment, which alters fluorescence intensity and the emission spectrum. Fluorescent assays typically have high sensitivity since the chromophore's properties strongly depend on the microenvironment. A more sophisticated spectroscopic method to study conformational transitions in RNA is nuclear magnetic resonance (NMR) (reviewed in [15]). Since NMR signals depend on the microenvironment of atoms, they can provide information about protonation, interactions, and local structure, and thus they can be effectively used to study a complex behavior such as cooperativity [16] .

While spectroscopy is extremely useful, technological advances over the last several decades have given rise to the use of isothermal titration calorimetry (ITC) as a means to measure the heat associated with interactions between biological molecules [17]. In this technique, the solution of a macromolecule is located inside the sample cell directly adjacent to a reference cell and the ligand solution in the injector syringe. The ligand solution is injected periodically into the sample cell, and each injection triggers the binding reaction and formation of the complex. As the sequence of injections proceeds, the heat associated with each injection is proportional to the increase in complex concentration.

Applying the appropriate model and nonlinear regression in data analysis, ITC can determine the association constant or binding affinity, K_a, the binding enthalpy, ΔH, the stoichiometry, n, the entropy change, ΔS, and the Gibbs free energy of binding, ΔG, in a single experiment. In contrast, spectroscopy experiments must be performed at various temperatures to determine the enthalpy and entropy of the reaction. It should be noted that the heat measured upon RNA-ligand binding does not result only from the formation of the direct RNA-ligand interactions but also includes heat generated from other binding-associated events, such as allosteric structural transitions and desolvation of the molecules. While several methodological advances allowed ITC to be used for determining cooperativity in protein systems [18, 19], the use of calorimetry to directly determine cooperativity of ligand binding to RNA is difficult [20], which makes ITC more appropriate for studying the mechanisms of RNA folding [21] and macromolecular assemblies [22, 23].

Spectroscopic and ITC methods typically provide bulk measurements and are invaluable for determining macroscopic characteristics of RNA systems but cannot directly visualize conformational transitions in RNA folding. The allosteric modulations can, however, be traced at the level of individual molecules by the so-called single-molecule techniques such as single-molecule fluorescence resonance energy transfer (smFRET) [24]. This method involves visualization of fluorescently labeled individual RNA molecules under the microscope. Special fluorescent labels attached to different regions of RNA induce FRET when coming in close proximity upon ligand binding [25]. Analysis of individual traces of molecules provides a comprehensive picture of allosteric re-arrangements in the RNA.

2.3 Determination of Cooperative Ligand Binding "On the Fly"

Throughout the twentieth century, researchers have developed various models to describe the binding of multiple ligands to oligomeric proteins, and many such models are applicable to manifestations of cooperativity in RNA. One of the most used models describing cooperative binding of ligands was developed by Hill and named after him [26]. The equation produces a "Hill coefficient" n, which is more than 1 when the system exhibits positive cooperativity and less than 1 if the system exhibits negative cooperativity, with the total number of ligand-binding sites being an upper limit for the coefficient. Although not ideal for evaluation of cooperativity [27, 28], the Hill coefficient is broadly used by biochemists for detecting cooperativity in the systems that involve binding of multiple ligands to RNA molecules. For example, the Hill coefficient value of 1.6 was the basis for conclusions about positive cooperativity in the truncated version of the dual glycine-sensing RNA [29].

Determination of the Hill coefficient relies on measuring binding affinity between RNA and ligands and does not require deep knowledge of the system's thermodynamics or special knowledge from the experimenter beyond the ability to employ a binding technique. The Hill coefficient can be determined by a variety of approaches including both the techniques used for studying protein-ligand interactions and the methods that exploit unique chemical and structural properties of RNA. The former includes various spectroscopic methods and ITC, as discussed earlier. Among the latter, it is worth mentioning techniques developed to probe ligand-induced changes in the conformation and stability of RNA molecules, such as in-line probing [29], nuclease cleavage [20], and chemical probing [30, 31]. Although all techniques aim to detect ligand-induced allosteric changes in RNA, in-line probing is probably the most robust and easiest method that does not require incubation with specific probes (nucleases or chemicals) and special treatments to stop the cleavage or modification reaction. In-line probing specifically exploits inherent instability of RNA in water solutions, which is more pronounced in flexible regions and greatly accelerated by divalent Mg^{2+} cations and elevated pH [32]. The method involves incubation of end-labeled RNA molecules in the absence of the ligand and at various ligand concentrations, and detection of changes in the RNA cleavage pattern after separating RNA fragments electrophoretically on a denaturing polyacrylamide gel. The extent of changes upon ligand titration can be used as a measure of binding affinity and the Hill coefficient. The method is naturally restricted to probe the RNA regions that change stability by binding to a ligand or becoming involved in intermolecular interactions as a result of ligand binding.

3 Folding of RNA into Its Secondary Structure

Although RNA is capable of forming intricate tertiary structures paralleled to those formed by proteins, the folding of both macromolecules involves different forces and results in dissimilar structural features. In contrast to proteins that have twenty amino acids, RNA is composed of only four similar chemical blocks with 50% more atoms than a protein having the same number of residues. Along with a larger size, RNA contains more dihedral bonds where rotations that introduce a greater potential for alternative conformations can occur. Unlike proteins, which are mostly composed of α-helices and β-sheets merged into a compact structure by hydrophobic interactions of side chains, RNA predominantly adopts a single secondary structure element, the double-stranded helix. This structural element contains many negative charges made up of phosphate groups on its periphery, thereby restricting the assembly of a hydrophobic core. Formation of helices in RNA is mostly

driven by stacking interactions between nucleobases and by the formation of complementary base pairs within a single RNA molecule when the RNA chain folds back on itself. In this case, the resulting double-stranded structure contains a loop that links the oppositely directed strands. If the loop is small, the structure is called a hairpin; when the double-stranded region is closed by a large loop, it is known as a stem-loop.

Significant negative charges along the RNA phosphodiester backbone and base stacking stiffen RNA helices and restrict the ability of RNA to form complex structures. Disruptions within the helix and neutralization of negative charges by counter ions relieve stiffness and facilitate formation of tertiary structure and adaptation of unique functional conformations [33]. Fundamental RNA properties such as favorable thermodynamics of stacking and hydrogen bonding, rapid kinetics of secondary, relative to tertiary, structure formation, and directional in vivo folding [34, 35] dictate a hierarchical manner by which RNA adopts its complex structure. This hierarchical folding, initiated by the formation of secondary structure elements from consecutively transcribed regions, proceeds by joining independently formed elements to form tertiary structures.

The hairpin is the most common secondary structural elements in RNA, which can function on its own as a ligand-binding region or nucleate formation of a complex RNA structure through RNA-RNA interactions [34, 36]. The stem of a hairpin is comprised mainly of Watson-Crick base pairs formed between two antiparallel stretches of RNA, and ranges in length from 1 base pair (bp) to more than 10, with an average length of 3–4 bp [37]. Due to steric repulsion, a loop connecting the strands contains a minimum of three nucleotides. RNA hairpin folding can often be described as a cooperative event, especially for short hairpins in which the thermodynamic worth of each base pair is more significant and the formation of a loop brings more disorder than in hairpins with longer helices [38].

To gain a better understanding of hairpin folding, RNA loops and their closing base pairs have been thermodynamically dissected for various systems, including for most prevalent four-nucleotide loops called tetraloops [39, 40]. Atomic resolution structures of phylogenetically common tetraloops UNCG, GNRA, and CUUG (where N represents any nucleotide, and R represents A or G) [41–43] have shown that these motifs undergo base stacking and extensive hydrogen bonding that make them extremely stable. Thermal studies revealed that some tetraloops undergo a two-state all-or-none folding, indicative of high cooperativity in the system [38]. Other studies have pointed out that in relationship to DNA, RNA is less cooperative in its folding, reflecting a smaller thermodynamic effect upon mutating 1–3 nucleotides of a loop [44]. This may be a beneficial feature of RNA allowing for a diverse primary

sequence while maintaining the secondary structure stability and function.

To understand the basic concepts in folding of more complex structures that contain more than one structural element, several studies interrogated folding of the P5abc stem-loop structure from the *Tetrahymena thermophila* ribozyme [45–47]. P5abc is a long stem-loop structure that contains three irregularities: a nucleotide bulge, an A-rich internal loop, and an additional hairpin, P5c, branching off of the stem (Fig. 2a). Elimination of the internal loop and the junctional hairpin in the P5ab construct showed a two-state formation of P5ab, indicative of a highly cooperative mechanism of the hairpin folding. Addition of the P5c hairpin in P5abcΔA retained two-state folding although folding and unfolding rates decreased. This reduction in transition between states likely results from nucleation of two hairpins instead of one; therefore, kinetic barriers for each substructure must be crossed prior to completion of folding. Finally, in the presence of both the P5c hairpin and A-rich bulge, the construct P5abc folds and unfolds with intermediates. Thus, introduction of irregularities into the regular stem-loop structure changes kinetics of folding and breaks highly cooperative formation of the helix providing the structural and kinetic foundation for tertiary interactions in the functional domain.

4 RNA Tertiary Structure Formation

Despite its fundamental importance, cooperativity does not necessarily contribute to all aspects of protein folding as some regions of proteins can fold and unfold as independent units. RNA molecules, like proteins, must fold into three-dimensional structures to carry out biological functions. However, RNA can form stable secondary structures in the absence of tertiary structure, thus posing a question of whether cooperativity needs to be employed in tertiary RNA folding. The P4-P6 domain of the *Tetrahymena* group I ribozyme represents a model RNA system that has been extensively studied to determine the extent of cooperativity in tertiary RNA folding [48–51]. The crystal structure of the P4-P6 domain [41] revealed the side-by-side packing of two helical structures (Fig. 2b) connected by the J5/5a turn and stabilized by two long-distance tertiary contacts involving the metal core/metal core receptor and the tetraloop/tetraloop receptor [41, 52, 53]. Thus, P4-P6 domain folding may require cooperativity between two tertiary contacts.

To gain insight into the cooperative folding mechanism, the P4-P6 domain folding was studied using a thermodynamic box similar to the one we described for the study of hairpin folding [48]. Each of the tertiary contacts was disrupted by mutations, one site at a time, and each RNA was internally labeled by two

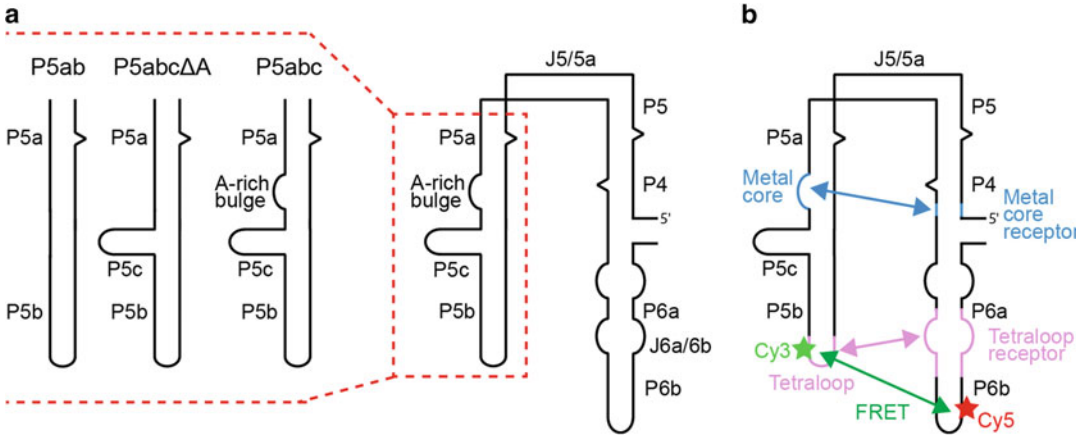

Fig. 2 Cooperative interactions in folding of the P4-P6 domain of the *Tetrahymena* group I ribozyme. (**a**) Folding of the P5 stem-loop structure [45]. Molecular constructs used for assessing the folding of the RNA are on the left and the entire domain is on the right. (**b**) Molecular construct used to determine cooperativity of tertiary contact formation [48, 49]. Tertiary contacts are depicted in colors and connected by arrows. Internal fluorescent labels and FRET are shown with stars and a green arrow, respectively

fluorophores, Cy3 and Cy5, which produce FRET, when coming in close proximity. Measuring FRET at the single-molecule level allowed detection of the folded state and determination of equilibrium constants for structural transitions in the wild-type and two mutant RNAs. Quantification of tertiary cooperativity from these measurements revealed that in each case the tertiary contact formation is 240-fold more favorable subsequent to formation of the other tertiary contact. Therefore, RNA folding may involve cooperative formation of tertiary contacts, at least in the systems with close positioning of such contacts.

5 RNA-Ligand Interactions

5.1 Cooperativity and Allostery in Binding of Ligands to Riboswitches

Cooperativity in RNA interactions occurs beyond folding of secondary and tertiary structures. Many RNAs interact with various cations and some RNAs bind small molecules or macromolecules in a cooperative manner. Examples of such RNAs include riboswitches, noncoding RNA regions which are capable of undergoing large allosteric transitions to modulate expression of the adjacent genes in response to specific binding to cellular metabolites and ions (reviewed in [54]). Evolutionarily conserved metabolite-sensing domains of riboswitches adopt intricate three-dimensional structures that specifically recognize cognate metabolites and reject similar compounds.

The cognate ligands of the M-box riboswitch, found in *Bacillus subtilis mgtE* gene, are multiple Mg^{2+} cations [55]. Mg^{2+} cations are among the most abundant divalent cations in cells and are

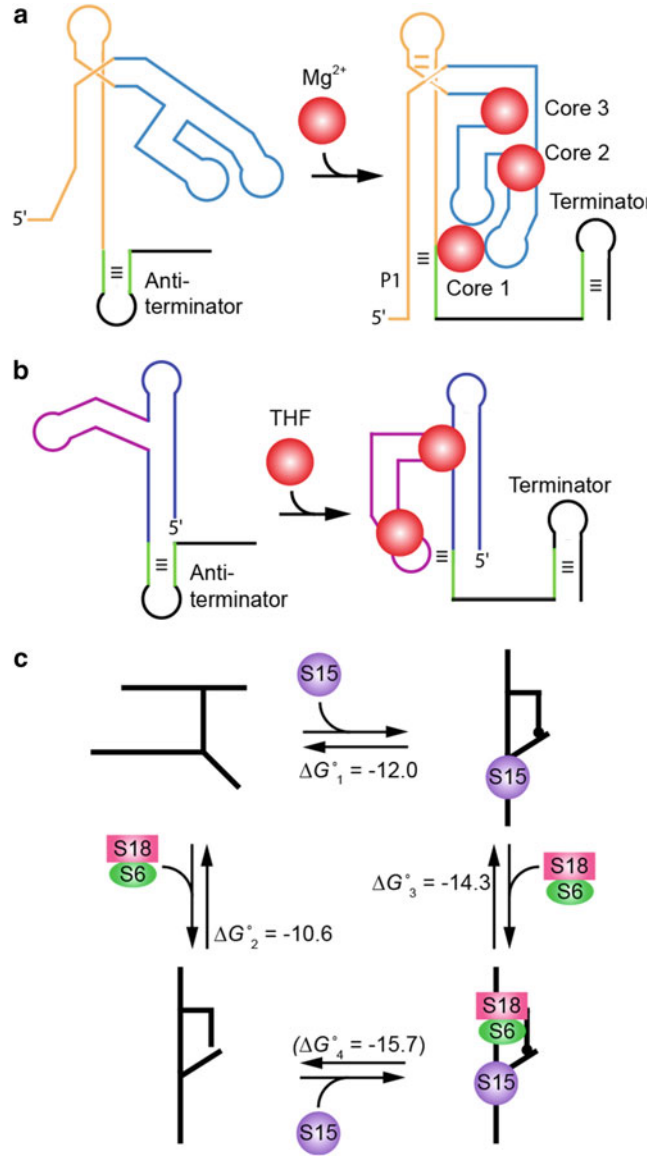

Fig. 3 Cooperativity and allosteric modulations in the assembly of RNA-ligand complexes. (**a**) Schematic of alternative folding of the M-box riboswitch [55]. Schematic depicts ligand-free active conformation of the riboswitch on the left and ligand-bound, repressed conformation on the right. Several Mg^{2+} cations predominantly bind RNA in core 2 and 3 regions and induce allosteric transition that brings the RNA structure, depicted in blue, closer to the 5′ end of the RNA. This conformational change induces tertiary RNA interactions, mediated by cations in the core 1 regions, and facilitates formation of the transcription terminator. This aborts transcription elongation and switches the gene off. Pairing interactions essential for switching the conformations are shown with thin black lines. (**b**) Schematic of alternative folding of the THF riboswitch. Cooperative binding of two ligands to the metabolite-sensing domain

extensively used for RNA folding to neutralize negative charges of the RNA backbone [33] and promote close contacts between RNA regions. The M-box riboswitch contains several Mg^{2+}-binding pockets, which facilitate the formation of the riboswitch structure. Three of these regions are particularly important for forming tertiary long-distance interactions and allosteric transitions that cause alternative folding of the riboswitch [56, 57] (Fig. 3a). Structural and biochemical studies suggest that two of these sites (cores 2 and 3) initially bind several Mg^{2+} ligands and induce a conformational change that brings together two RNA regions that form the third Mg^{2+} binding region, core 1, thus allowing for long-distance tertiary interactions to form. These interactions induce formation of the regulatory helix P1 of the riboswitch, thereby preventing formation of the transcription antiterminator hairpin and facilitating folding of the transcription terminator in the downstream region. Thus, Mg^{2+} binding to the riboswitch modulates transcription of the downstream gene, which is related to Mg^{2+} transport, and ensures an adequate amount of the Mg^{2+} transporter in the cell. Although cooperative binding has not been explicitly demonstrated for the M-box riboswitch, allosteric modulations upon Mg^{2+} binding undoubtedly indicate involvement of cooperativity in the formation of the ligand-bound state of this RNA.

Recent studies revealed that riboswitches can specifically recognize metals aside from Mg^{2+} cations. One of the most interesting metal-binding riboswitches resides upstream of a manganese (Mn^{2+}) efflux pump gene [59, 60]. This RNA forms two distant cation-binding sites, one for a Mg^{2+} cation and another for a Mn^{2+} cation. Since the concentration of Mg^{2+} cations in cells is high, the riboswitch initially binds a Mg^{2+} cation. This interaction induces an allosteric change in the RNA structure and facilitates binding of a Mn^{2+} cation, if the concentration of Mn^{2+} in the cell exceeds the threshold. Cooperative binding of two metals directs RNA folding such that the riboswitch adopts a conformation that precludes formation of the transcription terminator and allows transcription of the gene.

Metal cations are not the only ligands that are able to bind riboswitches in a cooperative fashion and induce large allosteric changes. Tetrahydrofolate (THF)-sensing riboswitch recognizes two THF molecules in a single domain using two very similar ligand-binding sites [58]. Although the two sites are separated by

Fig. 3 (continued) is required to form an adjacent transcription terminator instead of an antiterminator [58]. (**c**) Thermodynamic cycle for formation of the S15-S6-S18-rRNA complex in the central domain of the 30S ribosomal subunit [8]. The rRNA fragment is shown in black lines and proteins are shown in color. The schematic depicts structural transitions in RNA upon protein binding and these transitions are the basis for cooperativity in the system

a large distance, ligands bind to the RNA in cooperative manner under physiological concentrations of Mg^{2+} cations (Fig. 3b). Binding of both ligands is required for stabilization of tertiary interactions which ensure large allosteric changes resulting in the formation of a transcription terminator instead of an antiterminator in the downstream regions. Ablation of binding by mutagenesis in one site decreases binding affinity to the second site, thus suggesting that binding to one site facilitates ligand interactions with the other. Mutation analyses further indicates that one site is more important for genetic control than the other. High Mg^{2+} concentrations apparently provide extra stabilization to the structure and diminish cooperativity between ligand-binding sites.

The c-di-AMP riboswitch [61–63] is another example of double ligand binding to a single RNA domain. Interestingly, the riboswitch adopts a twofold pseudosymmetrical square that binds two molecules of c-di-AMP along opposite sides of the square in almost identical fashion. Although cooperativity has not been directly determined, mutagenesis studies have shown that tertiary structure formation requires binding of both ligands [61]. Elimination of one ligand-binding site reduces ligand binding to the second ligand, suggesting that formation of one binding pocket is required for long-distance allosteric change that causes the subsequent folding of the other pocket. These data further suggest cooperative binding of two ligands, a hypothesis awaiting confirmation by further biochemical and biophysical studies.

5.2 Allosteric Changes Define Cooperative Assembly of RNA-Protein Complexes

RNA participates in the formation of many ribonucleoprotein (RNP) complexes involved in mRNA processing, localization, transport, translation, and other cellular processes. One of the best studied RNPs is the ribosome, which consists of two subunits, 30S and 50S, formed by several dozen proteins and a few RNA molecules. The 30S subunit can be assembled from purified proteins and RNA; however, the reconstitution of the functional subunit requires a specific order for protein binding, suggestive of multiple allosteric modulations and a cooperative manner of assembly [64]. The central domain of the ribosome initially forms separately from the rest of the subunit and, therefore, represents an excellent model system to study cooperativity and associated allosteric modulations. Structural and biochemical studies revealed that the domain assembly involves highly cooperative binding of ribosomal proteins [22, 65–68]. The process is initiated by binding of the ribosomal protein S15 to a ~200 nt region of 16S rRNA (Fig. 3c). The binding re-arranges and stabilizes conformations of two three-helix junctions. The top junction constitutes the binding site for the dimer of proteins S6 and S18; therefore, S15 binding facilitates further binding of the S6:S18 complex [22, 68]. Thermodynamic studies showed that binding of S15 and S6:S18 heterodimer is highly cooperative, with coupling free energy of

-3.7 kcal mol^{-1}. Since S15 does not interact with S6:S18, the basis for cooperativity is allosteric changes in the rRNA structure upon S15 binding. Formation of the central domain is an example that highlights complex allosteric transitions in RNA and multiple cooperative binding events that lead to binding of over 20 proteins to a ~1500 nt RNA in the assembly of the 30S subunit and the entire ribosome.

6 Concluding Remarks

The role of allostery and cooperativity in RNA systems is difficult to underestimate. Virtually all structured RNAs and their macromolecular assemblies employ both of these biological principles for adopting functional conformations. RNA folding critically depends on interactions with metal cations, especially Mg^{2+} cations, which facilitate the formation of secondary structure elements. Allosteric modulations further define folding pathways for the formation of a tertiary structure and various complexes. The folding pathway is further assisted by allosteric modulations induced by binding of other ligands, small molecules or proteins, and most often involves cooperative effects, either in RNA folding or in ligand binding. Despite the essential contribution of cooperativity and allostery for the timely formation of biologically relevant RNA structures and their various activities, determination of the mechanism of cooperativity and extent of allosteric changes remains a difficult task and is limited to several well-behaving systems. The major setbacks in these studies are the lack of sufficient structural information for various states of RNA molecules and difficulties in conducting detailed thermodynamic analysis of the conformational transitions in complexly folded RNAs. Although we begin to understand folding and macromolecular interactions in small systems, progress in studies of large RNPs is mostly limited to the ribosome, whose structures are available in various states and in complex with various effectors. The mechanisms of many large RNPs, such as the spliceosome, are still poorly understood, despite tremendous structural and biochemical efforts. Developments in single-molecule approaches and structural methods, especially in cryogenic electron microscopy, will hopefully address these deficiencies in the near future.

Acknowledgments

This work was supported by the NIH grants GM112940 and MH112165 (A.S.) and the NIH fellowship F31GM119357 (A.P.).

References

1. Cruz JA, Westhof E (2009) The dynamic landscapes of RNA architecture. Cell 136:604–609. https://doi.org/10.1016/j.cell.2009.02.003
2. Dethoff EA, Chugh J, Mustoe AM, Al-Hashimi HM (2012) Functional complexity and regulation through RNA dynamics. Nature 482:322–330. https://doi.org/10.1038/nature10885
3. Breaker RR (2002) Engineered allosteric ribozymes as biosensor components. Curr Opin Biotechnol 13:31–39. https://doi.org/10.1016/S0958-1669(02)00281-1
4. Alemán EA, Lamichhane R, Rueda D (2008) Exploring RNA folding one molecule at a time. Curr Opin Chem Biol 12:647–654. https://doi.org/10.1016/j.cbpa.2008.09.010
5. Peselis A, Gao A, Serganov A (2015) Cooperativity, allostery and synergism in ligand binding to riboswitches. Biochimie 117:100–109. https://doi.org/10.1016/j.biochi.2015.06.028
6. Royer WE, Knapp JE, Strand K, Heaslet HA (2001) Cooperative hemoglobins: conserved fold, diverse quaternary assemblies and allosteric mechanisms. Trends Biochem Sci 26:297–304. https://doi.org/10.1016/S0968-0004(01)01811-4
7. Bohr C, Hasselbalch K, Krogh A (1904) Concerning a biologically important relationship-the influence of the carbon dioxide content of blood on its oxygen binding. Skand Arch Physiol 16:401–412
8. Williamson JR (2008) Cooperativity in macromolecular assembly. Nat Chem Biol 4:458–465. https://doi.org/10.1038/nchembio.102
9. Moody EM, Bevilacqua PC (2003) Folding of a stable DNA motif involves a highly cooperative network of interactions. J Am Chem Soc 125:16285–16293. https://doi.org/10.1021/ja038897y
10. Siegfried NA, Bevilacqua PC (2009) Chapter 13 Thinking inside the box. Designing, implementing, and interpreting thermodynamic cycles to dissect cooperativity in RNA and DNA folding. Methods Enzymol 455:365–393. https://doi.org/10.1016/S0076-6879(08)04213-4
11. Laing LG, Draper DE (1994) Thermodynamics of RNA folding in a conserved ribosomal RNA domain. J Mol Biol 237:560–576. https://doi.org/10.1006/jmbi.1994.1255
12. Tuschl T, Gohlke C, Jovin TM et al (1994) A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. Science 266:785–789
13. 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. https://doi.org/10.1038/9316
14. Haller A, Soulière MF, Micura R (2011) The dynamic nature of RNA as key to understanding riboswitch mechanisms. Acc Chem Res 44:1339–1348. https://doi.org/10.1021/ar200035g
15. Bothe JR, Nikolova EN, Eichhorn CD et al (2011) Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. Nat Methods 8:919–931. https://doi.org/10.1038/nmeth.1735
16. Tochtrop GP, Richter K, Tang C et al (2002) Energetics by NMR: site-specific binding in a positively cooperative system. Proc Natl Acad Sci 99:1847–1852. https://doi.org/10.1073/pnas.012379199
17. Velazquez-Campoy A, Freire E (2005) ITC in the post-genomic era...? Priceless. Biophys Chem 115:115–124
18. Brown A (2009) Analysis of cooperativity by isothermal titration calorimetry. Int J Mol Sci 10:3457–3477. https://doi.org/10.3390/ijms10083457
19. Houtman JCD, Brown PH, Bowden B et al (2007) Studying multisite binary and ternary protein interactions by global analysis of isothermal titration calorimetry data in SEDPHAT: application to adaptor protein complexes in cell signaling. Protein Sci 16:30–42. https://doi.org/10.1110/ps.062558507
20. Huang L, Serganov A, Patel DJ (2010) Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch. Mol Cell 40:774–786. https://doi.org/10.1016/j.molcel.2010.11.026
21. Salim NN, Feig AL (2009) Isothermal titration calorimetry of RNA. Methods 47:198–205. https://doi.org/10.1016/j.ymeth.2008.09.003
22. Recht MI, Williamson JR (2004) RNA tertiary structure and cooperative assembly of a large ribonucleoprotein complex. J Mol Biol 344:395–407. https://doi.org/10.1016/j.jmb.2004.09.009
23. Feig AL (2009) Studying RNA-RNA and RNA-protein interactions by isothermal titration calorimetry. Methods Enzymol

468:409–422. https://doi.org/10.1016/S0076-6879(09)68019-8

24. Kobitski AY, Nierth A, Helm M et al (2007) Mg^{2+} -dependent folding of a Diels-Alderase ribozyme probed by single-molecule FRET analysis. Nucleic Acids Res 35:2047–2059. https://doi.org/10.1093/nar/gkm072
25. Shaw E, St-Pierre P, McCluskey K et al (2014) Using sm-FRET and denaturants to reveal folding landscapes. Methods Enzymol 549:313–341. https://doi.org/10.1016/B978-0-12-801122-5.00014-3
26. Hill AV (1910) The possible effects of the aggregation of the molecules of hæmoglobin on its dissociation curves. J Physiol 40:iv–vii. https://doi.org/10.1113/jphysiol.1910.sp001386
27. Sherman EM, Esquiaqui J, Elsayed G, Ye J-D (2012) An energetically beneficial leader-linker interaction abolishes ligand-binding cooperativity in glycine riboswitches. RNA 18:496–507. https://doi.org/10.1261/rna.031286.111
28. Kladwang W, Chou FC, Das R (2012) Automated RNA structure prediction uncovers a kink-turn linker in double glycine riboswitches. J Am Chem Soc 134:1404–1407. https://doi.org/10.1021/ja2093508
29. Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, Breaker RR (2004) A glycine-dependent riboswitch that uses cooperative binding to control gene expression. Science 306:275–279. https://doi.org/10.1126/science.1100829
30. Lipfert J, Das R, Chu VB et al (2007) Structural transitions and thermodynamics of a glycine-dependent riboswitch from Vibrio cholerae. J Mol Biol 365:1393–1406. https://doi.org/10.1016/j.jmb.2006.10.022
31. Cheng CY, Chou FC, Kladwang W et al (2015) Consistent global structures of complex RNA states through multidimensional chemical mapping. elife 4:e07600. https://doi.org/10.7554/eLife.07600
32. Regulski EE, Breaker RR (2008) In-line probing analysis of riboswitches. Methods Mol Biol 419:53–67. https://doi.org/10.1007/978-1-59745-033-1_4
33. Misra VK, Draper DE (2002) The linkage between magnesium binding and RNA folding. J Mol Biol 317:507–521. https://doi.org/10.1006/jmbi.2002.5422
34. Tinoco I, Bustamante C (1999) How RNA folds. J Mol Biol 293:271–281. https://doi.org/10.1006/jmbi.1999.3001
35. Heilman-Miller SL, Woodson SA (2003) Effect of transcription on folding of the Tetrahymena ribozyme. RNA 9:722–733. https://doi.org/10.1261/rna.5200903
36. Varani G (1995) Exceptionally stable nucleic acid hairpins. Annu Rev Biophys Biomol Struct 24:379–404. https://doi.org/10.1146/annurev.bb.24.060195.002115
37. Gutell RR (1994) Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994. Nucleic Acids Res 22:3502–3507. https://doi.org/10.1093/nar/22.17.3502
38. Ma H, Proctor DJ, Kierzek E et al (2006) Exploring the energy landscape of a small RNA hairpin. J Am Chem Soc 128:1523–1530. https://doi.org/10.1021/ja0553856
39. Proctor DJ, Ma H, Kierzek E et al (2004) Folding thermodynamics and kinetics of YNMG RNA hairpins: specific incorporation of 8-bromoguanosine leads to stabilization by enhancement of the folding rate. Biochemistry 43:14004–14014. https://doi.org/10.1021/bi048213e
40. Fiore JL, Nesbitt DJ (2013) An RNA folding motif: GNRA tetraloop–receptor interactions. Q Rev Biophys 46:223–264. https://doi.org/10.1017/S0033583513000048
41. Cate JH, Gooding AR, Podell E et al (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. Science 273:1678–1685. https://doi.org/10.1126/science.273.5282.1678
42. Cheong C, Varani G, Tinoco I Jr (1990) Solution structure of an unusually stable RNA hairpin, 5′ GGAC(UUCG)GUCC. Nature 346:680–682. https://doi.org/10.1038/346680a0
43. Ennifar E, Nikulin A, Tishchenko S et al (2000) The crystal structure of UUCG tetraloop. J Mol Biol 304:35–42. https://doi.org/10.1006/jmbi.2000.4204
44. Moody EM, Feerrar JC, Bevilacqua PC (2004) Evidence that folding of an RNA tetraloop hairpin is less cooperative than its DNA counterpart. Biochemistry 43:7992–7998. https://doi.org/10.1021/bi049350e
45. Liphardt J, Onoa B, Smith SB et al (2001) Reversible unfolding of single RNA molecules by mechanical force. Science 292:733–737. https://doi.org/10.1126/science.1058498
46. Hyeon C, Thirumalai D (2005) Mechanical unfolding of RNA hairpins. Proc Natl Acad Sci U S A 102:6789–6794. https://doi.org/10.1073/pnas.0408314102
47. Hyeon C, Thirumalai D (2007) Mechanical unfolding of RNA: from hairpins to structures with internal multiloops. Biophys J

92:731–743. https://doi.org/10.1529/biophysj.106.093062
48. Sattin BD, Zhao W, Travers K et al (2008) Direct measurement of tertiary contact cooperativity in RNA folding. J Am Chem Soc 130:6085–6087. https://doi.org/10.1021/ja800919q
49. Bisaria N, Greenfeld M, Limouse C et al (2016) Kinetic and thermodynamic framework for P4-P6 RNA reveals tertiary motif modularity and modulation of the folding preferred pathway. Proc Natl Acad Sci 113: E4956–E4965. https://doi.org/10.1073/pnas.1525082113
50. Solomatin SV, Greenfeld M, Herschlag D (2010) Multiple native states reveal persistent ruggedness of an RNA folding landscape. Nature 463:681–684. https://doi.org/10.1038/nature08717
51. Greenfeld M, Solomatin SV, Herschlag D (2011) Removal of covalent heterogeneity reveals simple folding behavior for P4-P6 RNA. J Biol Chem 286:19872–19879. https://doi.org/10.1074/jbc.M111.235465
52. Murphy FL, Cech TR (1993) An independently folding domain of RNA tertiary structure within the Tetrahymena ribozyme. Biochemistry 32:5291–5300. https://doi.org/10.1021/bi00071a003
53. Murphy FL, Cech TR (1994) GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain. J Mol Biol 236:49–63. https://doi.org/10.1006/jmbi.1994.1117
54. Serganov A, Nudler E (2013) A decade of riboswitches. Cell 152:17–24. https://doi.org/10.1016/j.cell.2012.12.024
55. Dann CE, Wakeman CA, Sieling CL et al (2007) Structure and mechanism of a metal-sensing regulatory RNA. Cell 130:878–892. https://doi.org/10.1016/j.cell.2007.06.051
56. Wakeman CA, Ramesh A, Winkler WC (2009) Multiple metal-binding cores are required for metalloregulation by M-box riboswitch RNAs. J Mol Biol 392:723–735. https://doi.org/10.1016/j.jmb.2009.07.033
57. Ramesh A, Wakeman CA, Winkler WC (2011) Insights into metalloregulation by M-box riboswitch RNAs via structural analysis of manganese-bound complexes. J Mol Biol 407:556–570. https://doi.org/10.1016/j.jmb.2011.01.049
58. Trausch JJ, Ceres P, Reyes FE, Batey RT (2011) The structure of a tetrahydrofolate-sensing riboswitch reveals two ligand binding sites in a single aptamer. Structure 19:1413–1423. https://doi.org/10.1016/j.str.2011.06.019
59. Price IR, Gaballa A, Ding F et al (2015) Mn^{2+}-sensing mechanisms of yybP-ykoY orphan riboswitches. Mol Cell 57:1110–1123. https://doi.org/10.1016/j.molcel.2015.02.016
60. Dambach M, Sandoval M, Updegrove TB et al (2015) The ubiquitous yybP-ykoY riboswitch is a manganese-responsive regulatory element. Mol Cell 57:1099–1109. https://doi.org/10.1016/j.molcel.2015.01.035
61. Gao A, Serganov A (2014) Structural insights into recognition of c-di-AMP by the ydaO riboswitch. Nat Chem Biol 10:787–792. https://doi.org/10.1038/nchembio.1607
62. Jones CP, Ferré-D'Amaré AR (2014) Crystal structure of a c-di-AMP riboswitch reveals an internally pseudo-dimeric RNA. EMBO J 33:2692–2703. https://doi.org/10.15252/embj.201489209
63. Ren A, Patel DJ (2014) c-di-AMP binds the ydaO riboswitch in two pseudo-symmetry-related pockets. Nat Chem Biol 10:780–786. https://doi.org/10.1038/nchembio.1606
64. Held WA, Ballou B, Mizushima S, Nomura M (1974) Assembly mapping of 30 S ribosomal proteins from Escherichia coli. J Biol Chem 249:3103–3111
65. Nikulin A, Serganov A, Ennifar E et al (2000) Crystal structure of the S15-rRNA complex. Nat Struct Biol 7:273–277. https://doi.org/10.1107/S0108767300022558
66. Agalarov SC, Sridhar Prasad G, Funke PM, Stout CD, Williamson JR (2000) Structure of the S15,S6,S18-rRNA complex: assembly of the 30S ribosome central domain. Science 288:107–112. https://doi.org/10.1126/science.288.5463.107
67. Mulder AM, Yoshioka C, Beck AH et al (2010) Visualizing ribosome biogenesis: parallel assembly pathways for the 30S subunit. Science 330:673–677. https://doi.org/10.1126/science.1193220
68. Recht MI, Williamson JR (2001) Central domain assembly: thermodynamics and kinetics of S6 and S18 binding to an S15-RNA complex. J Mol Biol 313:35–48. https://doi.org/10.1006/jmbi.2001.5018

INDEX

Luisa Di Paola and Alessandro Giuliani (eds.), *Allostery: Methods and Protocols*, Methods in Molecular Biology, vol. 2253, https://doi.org/10.1007/978-1-0716-1154-8, © Springer Science+Business Media, LLC, part of Springer Nature 2021

N

P

R

S

T

W