

Advances in Experimental Medicine and Biology 1105

Haruki Nakamura · Gerard Kleywegt  
Stephen K. Burley · John L. Markley  
*Editors*

# Integrative Structural Biology with Hybrid Methods

 Springer

# **Advances in Experimental Medicine and Biology**

Volume 1105

## **Editorial Board**

IRUN R. COHEN, *The Weizmann Institute of Science, Rehovot, Israel*

ABEL LAJTHA, *N.S. Kline Institute for Psychiatric Research,  
Orangeburg, NY, USA*

JOHN D. LAMBRIS, *University of Pennsylvania, Philadelphia, PA, USA*

RODOLFO PAOLETTI, *University of Milan, Milan, Italy*

NIMA REZAEI, *Children's Medical Center Hospital, Tehran University of Medical  
Sciences, Tehran, Iran*

More information about this series at <http://www.springer.com/series/5584>

Haruki Nakamura • Gerard Kleywegt  
Stephen K. Burley • John L. Markley  
Editors

# Integrative Structural Biology with Hybrid Methods

 Springer

*Editors*

Haruki Nakamura  
Institute for Protein Research  
Osaka University  
Suita, Osaka, Japan

Stephen K. Burley  
Rutgers, The State University  
of New Jersey  
Piscataway,  
NJ, USA

Gerard Kleywegt  
European Bioinformatics Institute  
Cambridgeshire, UK

John L. Markley  
Biochemistry Department  
University of Wisconsin-Madison  
Madison, WI, USA

ISSN 0065-2598

ISSN 2214-8019 (electronic)

Advances in Experimental Medicine and Biology

ISBN 978-981-13-2199-3

ISBN 978-981-13-2200-6 (eBook)

<https://doi.org/10.1007/978-981-13-2200-6>

Library of Congress Control Number: 2018962157

© Springer Nature Singapore Pte Ltd. 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Contents

## Part I Introduction and Historical Background

- 1 Overall Introduction and Rationale, with View from Computational Biology** ..... 3  
Haruki Nakamura
- 2 Integrative/Hybrid Methods Structural Biology: Role of Macromolecular Crystallography** ..... 11  
Stephen K. Burley
- 3 View from Nuclear Magnetic Resonance Spectroscopy** ..... 19  
John L. Markley

## Part II New Experimental Tools Enabling Hybrid Methods

- 4 Complementary Use of Electron Cryomicroscopy and X-Ray Crystallography: Structural Studies of Actin and Actomyosin Filaments** ..... 25  
Takashi Fujii and Keiichi Namba
- 5 Current Solution NMR Techniques for Structure-Function Studies of Proteins and RNA Molecules**..... 43  
John L. Markley
- 6 The PA Tag: A Versatile Peptide Tagging System in the Era of Integrative Structural Biology** ..... 59  
Zuben P. Brown and Junichi Takagi
- 7 Small Angle Scattering and Structural Biology: Data Quality and Model Validation** ..... 77  
Jill Trehwella
- 8 Structural Investigation of Proteins and Protein Complexes by Chemical Cross-Linking/Mass Spectrometry** ..... 101  
Christine Piotrowski and Andrea Sinz

|   |  |     |
|---|--|-----|
| <b>9</b>  | <b>Prediction of Structures and Interactions from Genome Information</b> .....   | 123 |
|   | Sanzo Miyazawa   |     |
| <b>10</b>   | <b>A Hybrid Approach for Protein Structure Determination Combining Sparse NMR with Evolutionary Coupling Sequence Data</b> .....       | 153 |
|   | Yuanpeng Janet Huang, Kelly P. Brock, Chris Sander, Debora S. Marks, and Gaetano T. Montelione   |     |
| <b>11</b>   | <b>Harnessing the Combined Power of SAXS and NMR</b> .....   | 171 |
|   | A. M. Gronenborn   |     |
| <b>12</b>   | <b>2DHybrid Analysis</b> .....   | 181 |
|   | Atsushi Matsumoto and Kenji Iwasaki  |     |
| <b>Part III New Computational Tools Enabling Hybrid Methods</b> |  |     |
| <b>13</b>   | <b>Hybrid Methods for Macromolecular Modeling by Molecular Mechanics Simulations with Experimental Data</b> .....                      | 199 |
|   | Osamu Miyashita and Florence Tama  |     |
| <b>14</b>   | <b>Rigid-Body Fitting of Atomic Models on 3D Density Maps of Electron Microscopy</b> .....   | 219 |
|   | Takeshi Kawabata   |     |
| <b>15</b>   | <b>Hybrid Methods for Modeling Protein Structures Using Molecular Dynamics Simulations and Small-Angle X-Ray Scattering Data</b> ..... | 237 |
|   | Toru Ekimoto and Mitsunori Ikeguchi  |     |
| <b>Part IV Data Validation and Archives for Hybrid Methods</b>  |  |     |
| <b>16</b>   | <b>Archiving of Integrative Structural Models</b> .....  | 261 |
|   | Helen M. Berman, Jill Trehwella, Brinda Vallat, and John D. Westbrook  |     |

**Part I**  
**Introduction and Historical Background**



# Chapter 1

## Overall Introduction and Rationale, with View from Computational Biology



Haruki Nakamura

**Abstract** By integrating the experimental information given from the Hybrid/Integrative methods to determine the structures of large macromolecular machines, the static and dynamic molecular models in the atomic or semi-atomic resolution have been built with the aid of bioinformatics and computer simulations. Here, review of the recent progresses of such computational methods are made with discussion for the future direction.

**Keywords** Hybrid/integrative methods · Computational biology · Structural biology · X-ray · SAXS · NMR · Cryo-EM

### 1.1 Introduction

In recent years, the structures of large macromolecular machines in cells have been determined by combining observations from multiple, complementary experimental methods, such as X-ray crystallography, NMR spectroscopy, 3DEM (three-dimensional Electron Microscopy), X-ray and Neutron small-angle scattering (SAXS and SANS), FRET (Förster Resonance Energy Transfer), chemical crosslinking, and many others. In addition, by integrating such experimental information, the static and dynamic molecular models in the atomic or semi-atomic resolution have been built with the aid of bioinformatics and computer simulations. Currently, many structures determined by those so-called hybrid methods appear in high-impact-factor journals, and their atomic models are being deposited in the PDB (Protein Data Bank) (Berman et al. 2013, 2016) and the pilot site for the hybrid methods, PDB-dev (<https://pdb-dev.wwpdb.org/>) (Burley et al. 2017) which is managed by an international organization, the wwPDB (worldwide PDB: <https://wwpdb.org/>) (Berman et al. 2003, 2007; Markley et al. 2008)

---

H. Nakamura (✉)

PDBj, Institute for Protein Research, Osaka University, Suita, Osaka, Japan  
e-mail: [harukin@protein.osaka-u.ac.jp](mailto:harukin@protein.osaka-u.ac.jp)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,  
Advances in Experimental Medicine and Biology 1105,  
[https://doi.org/10.1007/978-981-13-2200-6\\_1](https://doi.org/10.1007/978-981-13-2200-6_1)

In October 2014, a task force wwPDB workshop was held to discuss how structural models derived from the integration of hybrid methods should be represented, validated and archived. News about this workshop was published in *Nature* (Ewen 2014), and the proceedings of the workshop were published in *Structure* (Sali et al. 2015). On October 3, 2015, the wwPDB Symposium “Integrative Structural Biology with Hybrid Methods” was held in Osaka, Japan.

This book will present the methods used to determine, validate, and archive structural models of large biomolecular complexes and cellular machines. Recent examples will be discussed along with current trends in molecular and cellular structural biology. Most of the initially proposed authors were speakers at the wwPDB symposium on October 3, 2015, and the book was first planned to serve as an updated summary of that meeting. However, the progress in this field has been much faster than what we planned first, and so we extended the Chapters covering the latest developments, which, we are sure, should be useful as one of the book series, *Advances in Experimental Medicine and Biology*.

Here, we review the recent progresses of such computational methods for several roles in the Hybrid/Integrative methods: (i) Analysis of genome information to obtain structural information at various levels, (ii) Integration of various methods to build the most probable atomic or semi-atomic resolution models, and (iii) Analysis of dynamic natures of complex structures. Finally, we discuss the future direction of the Hybrid/Integrative methods with the aid of the computational biology. There are other important issues, Validation of structures with the Hybrid/Integrative methods and Archiving of structural models determined by Hybrid/Integrative methods. Those will be described by other authors in this book, and we will not touch these issues here.

## 1.2 Analysis of Genome Information to Obtain Structural Information

There has been a long history to predict 3D protein structures from genome information, including comparative or homology modeling for the homologous proteins with sequence similarities larger than 30%, and *de novo* structural modeling with sequence similarities less than 30% to any known structures. Many methods have been developed and been matured during the blind contests, the Critical Assessment of Techniques for Protein Structure Prediction (CASP), since 1994 (Moult et al. 2016). Another blind contest, the Critical Assessment of Predicted Interactions (CAPRI), has also been established as the community-wide initiative since 2001, in order to develop reliable methodologies to predict protein-protein interactions and structures of protein assemblies (Wodak and Janin 2017).

In particular, by distinguishing true co-evolution couplings from the noisy observation for the evolutionary sequence variation, accurate predictions of residue-residue contacts have been made, and more reliable 3D protein structures are

predicted (Marks et al. 2011). This approach to use co-evolution information with the multiple sequence alignment for large family members has made great successes in not only soluble proteins (Marks et al. 2011), but also membrane proteins (Hopf et al. 2012) and even for protein complexes (Hopf et al. 2014). By combining with the atomic structure refinement, more precise 3D atomic structures have been built using metagenome sequence data as the genome wide analysis (Ovechinnikov et al. 2017). About this algorithm and method, Sanzo Miyazawa describes in details at Chap. 9 of this book.

Recent Hi-C (high-resolution chromosome conformation capture) experiments have revealed the chromatin organization in 3D (Lieberman-Aiden et al. 2009). In particular, from the single-cell Hi-C technology, individual chromosomes are shown to maintain domain organization at the megabase scale (Nagano et al. 2013), and 3D structural models reveal a radial architecture of chromosomal compartments with epigenome signature depending on the cell cycle (Nagano et al. 2017). It is also suggested that patchiness of DNA methylation correlates the 3D chromatin structure (Zhang et al. 2017). Those 3D chromatin structures have been constructed based on the distance geometry algorithm, essentially the same one, which was developed in the field of NMR structure determination described by John L. Markley (see Chap. 5) in this book. This field has just started to reveal many different dynamic chromatin structures, but massive genome sequencing will soon give us more detailed view of individual chromosomes at each cell-cycle and their relation to epigenetic signals.

### 1.3 Integration of Various Methods to Build the Atomic or Semi-atomic Resolution Models

The most important role of computation in the Hybrid methods is to build atomic or semi-atomic resolution models by integrating structure information obtained from various experimental methods.

When the cryo-EM does not give data with the atomic resolution, the ordinary way to make the atomic model is to fit the atomic structure already determined by X-ray crystallography or NMR. For the fitting method of such atomic models, Takeshi Kawabata describes a review in Chap. 14 including his own method of *gmfit* with Gaussian mixed modeling (Kawabata 2008).

In many cases, it is necessary to modify the amino-acid sequences by the comparative modeling mentioned in the above section. In addition, such atomic structures, which are only part of the huge complex structures in many cases, were determined in crystals or in solution, and so they may not completely fit the electron density maps because of structural changes. Polymorphic property of the structures captured by cryo-EM is also rather intrinsic. Thus, in order to solve those structural multiplicity, flexible fitting methods of atomic structural models into microscopy maps have been proposed using molecular dynamics (MD) simulations, which are reviewed by Florence Tama in Chap. 13. Usually, a pseudo potential function is

added to the conformation energy of the protein system, so as to fit the density map given by the cryo-EM experiment with that synthetically simulated from all-atom MD simulation (Orzechowski and Tama 2008). In particular, a program *MDF* developed by Klaus Schulten group has been frequently used (Trabuco et al. 2008; Alvarez et al. 2017). Other approach to cryo-EM structure refinement is to integrate the  $^{13}\text{C}$  chemical shifts from solution and solid-state NMR with MD simulation (Perilla et al. 2017).

## 1.4 Analysis of Dynamic Natures of Complex Structures

The other important role of computation is to analyze the dynamic natures of proteins. From the high-resolution maps by cryo-EM, it is now possible to directly determine various atomic structures by 2D- and 3D-image classification without using MD simulations (Zhao et al. 2015). However, such polymorphic structures are still difficult to be captured.

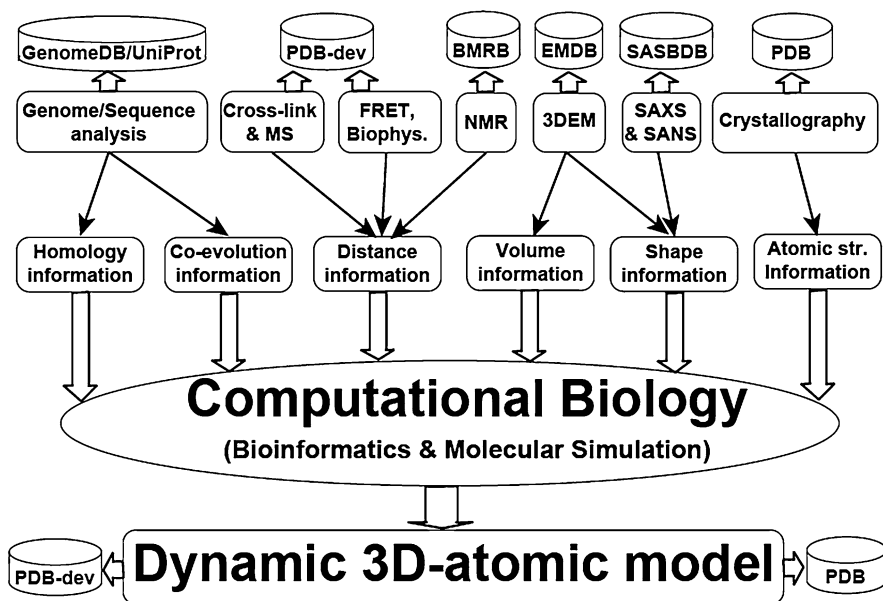
The solution NMR method is able to observe the dynamic property directly, but it is difficult to solve the structures having the molecular weights larger than 30,000 Da in an ordinary way. Tang et al. propose a new method, so called *EC-NMR* method, where the structural information measured by NMR is coupled with the co-evolution information with the multiple sequence alignment mentioned above (Tang et al. 2015). Gaetano T Montelione introduces the method in Chap. 10.

Because Small angle scattering experiments by X-ray (SAXS) or Neutron (SANS) are only available to give rough but dynamic structural image, MD simulations are powerful tool to build dynamic protein structural models in solution (Oroguchi and Ikeguchi 2011; Chen and Hub 2015). Mituhiro Ikeguchi describes the recent progresses of the method in Chap. 15. SAXS is frequently used to confirm the oligomeric state of protein systems. A hybrid NMR/SAXS approach integrated by computation has also been reported (Rossi et al. 2015), which is introduced by Angela M. Grogenborn in Chap. 11.

Finally, very large structural changes are observed by the intrinsically disordered regions, which are now understood to be very abundant in nuclei and cytoplasm of higher organism (Wright and Dyson 2015; Babu 2016). Because of their multi-modal nature, NMR and computer simulations can capture their putative structures as the ensemble. In particular, an enhanced structural sampling method is very powerful method to capture the multi-modal conformations (Kasahara et al. 2018). Recently, High-speed atomic force microscope (*HS-AFM*) can give us the images of the intrinsic disordered regions (Miyagi et al. 2008), in addition to the dynamic images of the actual movements of motor proteins and the rotational motion of  $\text{F}_1$ -ATPase (Ando 2014).

## 1.5 Conclusion

As shown in Fig. 1.1, computational biology offers a crucial tool for the Hybrid/Integrative methods, not only giving the initial putative models estimated from genome information, but also integrating information observed by various experiments, X-ray crystallography, NMR, cryo-EM, SAXS and so on. In particular, when the space resolution of each method is not very high, an atomic or semi-atomic resolution model can be built to satisfy the information given by the various methods. The dynamic natures of the protein systems can be revealed by molecular simulations. The role of bioinformatics and molecular simulation should become



**Fig. 1.1** Roles of computational biology in Hybrid/Integrative methods and data archives. Genomics by next-generation sequencer (NGS) produce huge information of genome sequences. Chemical cross-link with mass spectroscopy (MS) and Förster Resonance Energy Transfer (FRET) or any other biophysical measurements provide distance information among several particular atoms or atom groups in a molecule or supra-molecule, as well as NMR observation. Three-dimensional electron microscopy (3DEM) gives the volume map in a real space, and the atomic structure can be obtained when high-resolution electron density map is observed. Small angle X-ray scattering (SAXS) and that of neutron scattering (SANS) provide the shape information of molecules in solution. Many different kinds of experimental information are integrated by various methods of bioinformatics and molecular simulations. Raw experimental data are archived in the public databases: BMRB for NMR data, EMDB for 3DEM data, SASBDB for SAXS and SANS data, PDB-dev for distance data by FRET and other methods, and PDB for structure factors given by crystallography. The final three-dimensional atomic models, which have often dynamic features, are also archived by the wwPDB to PDB and PDB-dev

much more crucial for understanding the mechanisms and functions of molecular machines in cells.

**Acknowledgements** This work was supported by grants from the Database Integration Coordination Program from the National Bioscience Database Center (NBDC) – JST (Japan Science and Technology Agency), the Platform Project for Supporting in Drug Discovery and Life Science Research (Platform for Drug Discovery, Informatics, and Structural Life Science) from AMED, and JSPS KAKENHI [17K07364].

## References

- Alvarez FJD, He S, Perilla JR, Jang S, Schulten K, Engelman AN, Scheres SHW, Zhang P (2017) CryoEM structure of MxB reveals a novel oligomerization interface critical for HIV restriction. *Sci Adv* 3:e1701264
- Ando T (2014) High-speed AFM imaging. *Curr Opin Str Biol* 28:63–68
- Babu MM (2016) The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem Soc Trans* 44:1185–1200
- Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide protein data bank. *Nature Struct Biol* 10:980
- Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide protein data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucl Acids Res* 35:D301–D303
- Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2013) How community has shaped the protein data bank. *Structure* 21:1485–1491
- Berman HM, Burley SK, Kleywegt GJ, Markley JL, Nakamura H, Velankar S (2016) The archiving and dissemination of biological structure data. *Curr Opin Struct Biol* 40:17–22
- Burley SK, Kurisu G, Markley JL, Nakamura H, Velankar S, Berman HM, Sali A, Schwede T, Trewthella J (2017) PDB-dev: a prototype system for depositing integrative/hybrid structural models. *Structure* 25:1317–1318
- Chen P, Hub JS (2015) Interpretation of solution X-ray scattering by explicit-solvent molecular dynamics. *Biophys J* 108:2573–2584
- Ewen C (2014) Data bank struggles as protein imaging ups its game: hybrid methods to solve structures of molecular machines create a storage headache. *Nature* 514:416
- Hopf TA, Colwell LJ, Sheridan T, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149:1607–1621
- Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, Bonvin AMJJ, Marks DS (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *elife* 3:e03430
- Kasahara K, Shiina M, Higo J, Ogata K, Nakamura H (2018) Phosphorylation of an intrinsically disordered region of Ets1 shifts a multi-modal interaction ensemble to an auto-inhibitory state. *Nucl Acids Res* 46:2243–2251
- Kawabata T (2008) Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a Gaussian mixture mode. *Biophys J* 95:4643–4658
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragozcy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–293
- Markley JL, Ulrich EL, Berman HM, Henrick K, Nakamura H, Akutsu H (2008) BioMagResBank (BMRB) as a partner in the worldwide protein data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J Biomol NMR* 40:153–155

- Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6:e28766
- Miyagi A, Tsunaka Y, Uchihashi T, Mayanagi K, Hirose S, Morikawa K, Ando T (2008) Visualization of intrinsically disordered regions of proteins by high-speed atomic force microscopy. *Chem Phys Chem* 9:1859–1866
- Moult J, Fidelis K, Kryshchak A, Schwede T, Tramontano A (2016) Critical assessment of methods of protein structure prediction: progress and new directions in round XI. *Proteins* 84(Suppl 1):4–14
- Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502:59–64
- Nagano T, Lubling Y, Várnai C, Dudley C, Leung W, Baran Y, Cohen NM, Wingett S, Fraser P, Tanay A (2017) Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 547:61–67
- Oroguchi T, Ikeguchia M (2011) Effects of ionic strength on SAXS data for proteins revealed by molecular dynamics simulations. *J Chem Phys* 134:025102
- Orzechowski M, Tama F (2008) Flexible fitting of high-resolution X-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys J* 95:5692–5705
- Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D (2017) Protein structure determination using metagenome sequence data. *Science* 355:294–298
- Perilla JR, Zhao G, Lu M, Ning J, Hou G, Byeon I-JL, Gronenborn AM, Polenova T, Zhang P (2017) CryoEM structure refinement by integrating NMR chemical shifts with molecular dynamics simulations. *J Phys Chem B* 121:3853–3863
- Rossi P, Shi L, Liu G, Barbieri CM, Lee H-W, Grant TD, Luft JR, Xiao R, Acton TB, Snell EH, Montelione GT, Baker D, Lange OF, Sgourakis NG (2015) A hybrid NMR/SAXS-based approach for discriminating oligomeric protein interfaces using Rosetta. *Proteins* 83:309–317
- Sali A, Berman HM, Schwede T, Trewhella J, Kleywegt G, Burley SK, Markley JL, Nakamura H, Adams P, Bonvin AM, Chiu W, Peraro MD, Di Maio F, Ferrin TE, Grünewald K, Gutmanas A, Henderson R, Hummer G, Iwasaki K, Johnson G, Lawson CL, Meiler J, Marti-Renom MA, Montelione GT, Nilges M, Nussinov R, Patwardhan A, Rappsilber J, Read RJ, Saibil H, Schröder GF, Schwieters CD, Seidel CA, Svergun D, Topf M, Ulrich EL, Velankar S, Westbrook JD (2015) Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* 23:1156–1167
- Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat Methods* 12:751–754
- Trabuco LG, Villa E, Mitra K, Frank J, Schulten K (2008) Flexible fitting of atomic structures into Electron microscopy maps using molecular dynamics. *Structure* 16:673–683
- Wodak SJ, Janin J (2017) Modeling protein assemblies: critical assessment of predicted interactions (CAPRI) 15 years hence. *Proteins* 85:357–358
- Wright PE, Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* 16:18–29
- Zhang L, Xie WJ, Liu S, Meng L, Gu C, Gao YQ (2017) DNA methylation landscape reflects the spatial organization of chromatin in different cells. *Biophys J* 113:1395–1404
- Zhao J, Benlekbir S, Rubinstein JL (2015) Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase. *Nature* 521:241–245

# Chapter 2

## Integrative/Hybrid Methods Structural Biology: Role of Macromolecular Crystallography



Stephen K. Burley

**Abstract** Macromolecular crystallography has been central to the emergence and development of structural biology as a scientific discipline. Approximately 90% of the more than 138,000 three-dimensional structures currently available in the Protein Data Bank (PDB) archive, the single, global open access data resource for macromolecular structure data, were determined using X-ray crystallography. MX, the enormous variety of PDB structures of proteins, DNA, and RNA, and computational models derived therefrom will be central to the growth of integrative or hybrid (I/H) methods structural studies of macromolecular assemblies and other complex biological systems.

**Keywords** X-ray crystallography · Macromolecular crystallography · MX · Protein crystallography · 3D structure · Protein · DNA · RNA · Protein data Bank · PDB · Worldwide protein data Bank · wwPDB · Atomic coordinates · Structural biology · Integrative/hybrid methods · I/H methods

### 2.1 Introduction

Macromolecular crystallography or MX, also known as protein crystallography, first yielded atomic-level three-dimensional (3D) structures of small proteins in the 1950s and 1960s following the pioneering efforts by J.D. Bernal (London, UK), Dorothy Hodgkin (London, Oxford, UK), John Kendrew (Cambridge, UK), William N. Lipscomb, Jr. (Cambridge, US), Max F. Perutz (Cambridge, UK), David C. Phillips (London, Oxford, UK), Frederick M. Richards (New Haven, US), and their co-workers (many of them pioneers in their own right and too numerous to name in this chapter).

---

S. K. Burley (✉)

Rutgers, The State University of New Jersey, Piscataway, NJ, USA

e-mail: [stephen.burley@rcsb.org](mailto:stephen.burley@rcsb.org)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,

Advances in Experimental Medicine and Biology 1105,

[https://doi.org/10.1007/978-981-13-2200-6\\_2](https://doi.org/10.1007/978-981-13-2200-6_2)



In principle, the MX method is a simple one. The diffraction experiment is nothing more than an analog calculation in 3D of a discretely sampled, continuous Fourier transform of the shape of the electron rich portions of an ordered crystal made up of one or more macromolecules; followed by a digital calculation of a second Fourier transform; yielding a magnified 3D image of the electron rich portions of crystal, which can be interpreted as a 3D atomic-level structure of a macromolecule(s).

In practice, the experiment can be challenging, requiring highly purified preparations of biological macromolecules that will form a well-ordered 3D crystal; an intense, highly collimated source of monochromatic X-rays; a sample stage on which to position and eucentrically move the crystal within the X-ray beam; an electronic detector that accurately measures the intensity of the resulting X-ray diffraction pattern (i.e., directed spray of X-rays emerging from the crystal); an effective strategy for recovering the phase information for each diffracted X-ray beam that is sacrificed when the X-ray measurement are performed; a digital computer; an expert software system augmented by skilled a human to generate the 3D atomic coordinates of the non-hydrogen atoms comprising the macromolecule(s) that make up the crystal.

The very first X-ray structures of myoglobin, hemoglobin, lysozyme, carboxypeptidase A, ribonuclease S, and insulin literally took decades from the time that diffraction quality crystals were initially grown, requiring 100 s of person years of effort by large, multi-disciplinary teams. The situation was not much better in the early 1980s, when a single protein crystallographic structure determination typically required 20 person years. Today, it is not unusual for a 3D structure of a 50 kDa protein to be determined at near atomic resolution in less than 1 month by a trained individual, starting from a segment of double-stranded DNA that encodes the protein of interest.

Given the challenges workers in the field of protein crystallography faced through the decades of the 1950s and 1960s, what transpired in the summer of 1971 at Cold Spring Harbor Laboratory can be ascribed to enlightened self-interest. The famous quote from Benjamin Franklin, “We must, indeed, all hang together or, most assuredly, we shall all hang separately.” must have been top of mind. Protein crystallographers “hung together” by establishing the Protein Data Bank (PDB) as the first open access digital data resource in biology with just 7 X-ray structures (Protein Data Bank 1971). Doing so accelerated scientific and technical developments in the field, and the PDB now contains more than 138,000 structures of proteins, DNA, and RNA determined by MX, nuclear magnetic resonance spectroscopy (NMR), and electron microscopy (3DEM). Since 2003, the Worldwide PDB (wwPDB, wwPDB.org) organization has managed the PDB archive and ensured that PDB data are freely and publicly available to >1 million PDB *Data Consumers* around the globe (Berman et al. 2003). Locally-funded, regional PDB Data Centers in the US [RCSB Protein Data Bank, (Berman et al. 2000; Rose et al. 2017) and BioMagResBank (Ulrich et al. 2008)], Europe [Protein Data Bank in

Europe, (Velankar et al. 2016)], and Asia [Protein Data Bank Japan, (Kinjo et al. 2017)] safeguard and disseminate PDB structures using a common data dictionary (Fitzgerald et al. 2005) and a unified global system for data deposition-validation-biocuration by >30,000 PDB *Data Depositors* (Young et al. 2017).

It is not possible to do justice to the scientific underpinnings (e.g., chemistry, physics, mathematics, and statistics), the myriad technologies (X-ray sources and detectors, beam line engineering, computer hardware, data collection and analysis software, structure determination and refinement software, and molecular graphics hardware and software), and the power of MX as an experimental tool in a single book chapter or even an entire book. This chapter describes the roles that MX can play in I/H methods structure determination. Two topics are covered in some detail, including (i) MX Structure Data for I/H methods and (ii) Accessing Public-domain MX Structure Data for use in I/H methods.

## 2.2 MX Structure Data for I/H Methods

MX structure data are used for I/H methods structure determination in two ways.

First, but typically in only the most favorable situations, the macromolecular assembly of interest can be produced in sufficient amounts and with adequate purity that it will yield 3D crystals suitable for 3D structure determination from the X-ray diffraction experiment. As I/H methods target every larger experimental systems, crystalline samples will be fewer and farther between, and ever more challenging to work with. Rarely will they diffract strongly, or give diffraction data at high-enough resolution to succumb to structure determination using a single method and produce a 3D atomic level structure. When the method does work, the resulting structures are typically of only modest resolution (i.e., lower than 4 Å).

As of early 2017, the PDB archive contained >120,000 X-ray structures, of which 779 were obtained at 4 Å resolution or lower with 543 falling between 4 and 5 Å resolution (Fig. 2.1). The paucity of structures at very low resolution reflects the difficulty of phasing the diffraction pattern in the absence of higher resolution data. The lowest resolution MX structure in the PDB is that of Tropomyosin (PDB ID 2tma), obtained at 15 Å by Phillips and coworkers (Phillips 1986). Some of these lower resolution PDB structures lack atomic coordinates for amino acid side chains, treating the polypeptide chain as a polymer of Alanine residues.

Figure 2.2 illustrates the rate of addition of low-resolution X-ray structures to the PDB archive from 1971–2017. It is remarkable that >70 new low-resolution structures have been added to the archive each year since 2012. This acceleration reflects both the growing interest in macromolecular systems that do not yield high quality crystals and improvements in structure determination methods at lower resolution (Karmali et al. 2009; Brunger et al. 2009; Dyda 2010; DiMaio et al. 2013; Goh et al. 2016).

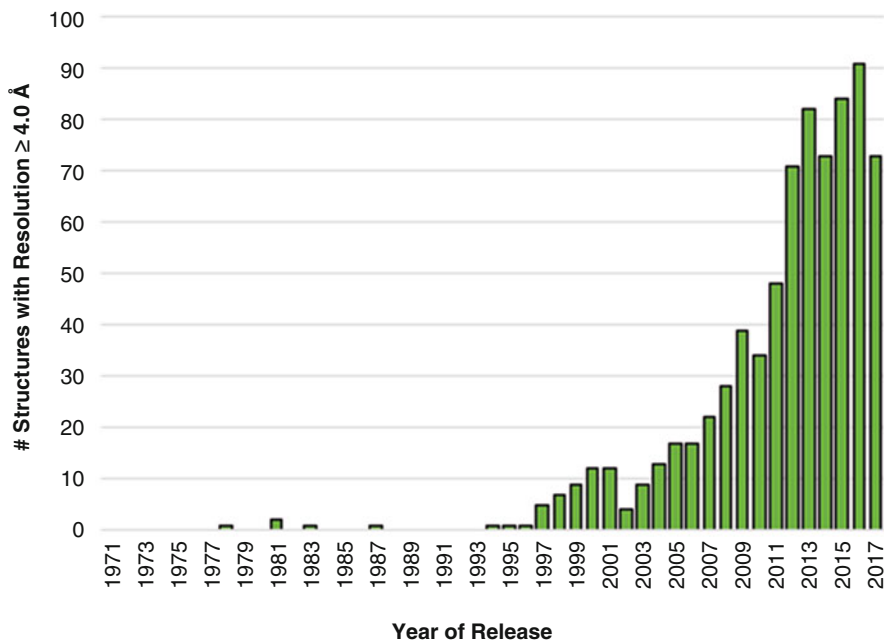


Fig. 2.1 Growth of PDB MX structures obtained at 4 Å resolution or lower

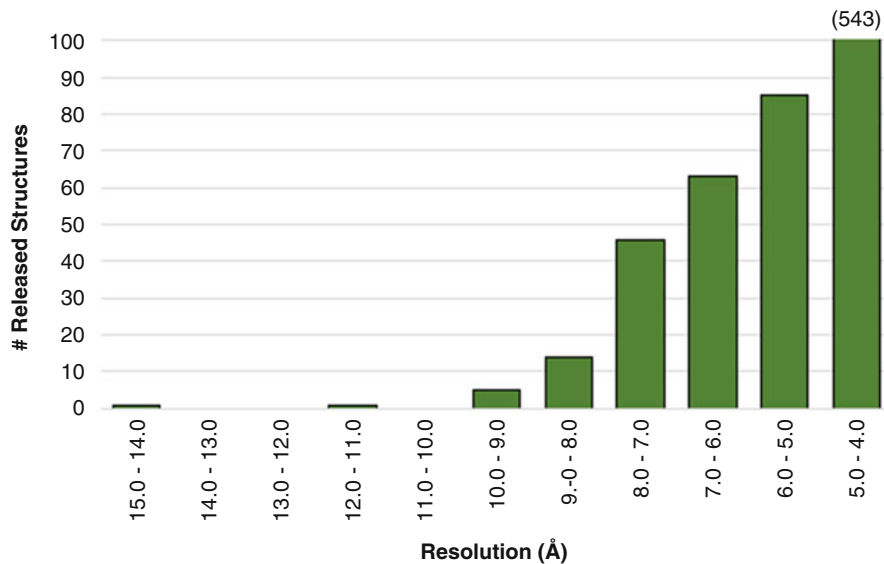


Fig. 2.2 Lower-resolution MX PDB structures *versus* resolution limit

Second, and more common, higher-resolution MX structures are used for I/H methods structure determination in piecemeal fashion. Much of the time, the overall size and shape of a macromolecular assembly can be determined by 3DEM or lower-resolution MX approaches. Then individual MX (and NMR) structures of components of the assembly can be positioned within the overall “envelope”, ideally by docking secondary structural elements of the component structure (typically  $\alpha$ -helices) into recognizable features identified in the lower-resolution MX electron map or the 3DEM mass density map. In the absence of an experimental structure of one or more individual components, it is often possible to use homology models computed from experimental structures of orthologous or paralogous proteins freely available from the PDB archive. Complementary data from chemical-crosslinking and fluorescence resonance energy transfer can also be used to refine placement of component structures (or homology models) within the envelope. The outcome of this more typical approach to I/H methods structure determination for a nuclear pore complex is illustrated in Fig. 2.3. Comprehensive reviews of integrative structure determination strategies have been published by Webb et al. (2018) and in Chaps. 4, 5, 6, 7, 8, 9, 10, 11 and 12 in Part 2 of this volume. At present, I/H methods structures can be deposited to PDB-Dev ([pdb-dev.wwpdb.org](http://pdb-dev.wwpdb.org)), a prototype deposition and archiving system (Burley et al. 2017).

### 2.3 Accessing Public-Domain Experimental and Computational Structure Data for I/H Methods

Experimental 3D structure data for biological macromolecules are made freely available to all without limitations on usage by the Protein Data Bank (PDB). Structure data are available from each of wwPDB partner websites [RCSB Protein Data Bank ([www.rcsb.org](http://www.rcsb.org)), Protein Data Bank Japan ([www.pdbj.org](http://www.pdbj.org)), the Protein Data Bank in Europe ([www.pdbe.org](http://www.pdbe.org))], which distribute identical archival data together with complementary information from other data resources. The quality of each incoming PDB structure is assessed at the time of deposition into the archive and then re-assessed annually (*versus* the entire archive). wwPDB structure validation reports for each structure are made available with the experimental data provided by each wwPDB partner (Gore et al. 2017). Every PDB structure is identified with a unique 4-character code (e.g., PDB 1vol). This code can be used to download the desired structure directly from the wwPDB ftp site (e.g., <ftp://ftp.wwpdb.org/pub/pdb/data/structures/divided/mmCIF/vo/1vol.cif.gz> for 1vol).

Computed homology models of biological macromolecules are available from various individual biodata resources. One of the most efficient ways to access homology models is to use the Protein Model Portal ([www.proteinmodelportal.org](http://www.proteinmodelportal.org)). This resource provides access to homology models computed *en masse* by SWISS-MODEL ([swissmodel.expasy.org](http://swissmodel.expasy.org)) and ModBase ([modbase.compbio.ucsf.edu](http://modbase.compbio.ucsf.edu)),

**Fig. 2.3** I/HM multi-scale structural model of the nuclear pore Nup84 complex (Shi et al. 2014)



together with model validation metrics. Individually archived homology models can be downloaded directly from the Model Archive ([www.modelarchive.org](http://www.modelarchive.org)) or from the Protein Model Portal.

**Acknowledgments** The RCSB PDB is jointly funded by the National Science Foundation, the National Institutes of Health, and the Department of Energy (NSF-DBI 1338415). We gratefully acknowledge help from Brian Hudson with figure preparation, and Nicole Oorbeek with manuscript preparation, and contributions from all members of the Research Collaboratory for Structural Bioinformatics Protein Data Bank and our Worldwide Protein Data Bank partners.

## References

- 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. <https://doi.org/10.1093/nar/28.1.235>
- Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide protein data Bank. *Nat Struct Biol* 10(12):980. <https://doi.org/10.1038/nsb1203-980>

- Brunger AT, DeLaBarre B, Davies JM, Weiss WI (2009) X-ray structure determination at low resolution. *Acta Crystallogr Ser D* 65(Pt 2):128–133. <https://doi.org/10.1107/S0907444908043795>
- Burley SK, Kurisu G, Markley JL, Nakamura H, Velankar S, Berman HM, Sali A, Schwede T, Trewella J (2017) PDB-dev: a prototype system for depositing integrative/hybrid structural models. *Structure* 25:1317–1318
- DiMaio F, Echols N, Headd JJ, Terwilliger TC, Adams PD, Baker D (2013) Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat Methods* 10(11):1102–1104. <https://doi.org/10.1038/nmeth.2648>
- Dyda F (2010) Developments in low-resolution biological X-ray crystallography. *F1000 Biol Rep* 2:80. <https://doi.org/10.3410/B2-80>
- Fitzgerald PMD, Westbrook JD, Bourne PE, McMahon B, Watenpaugh KD, Berman HM (2005) 4.5 macromolecular dictionary (mmCIF). In: Hall SR, McMahon B (eds) *International tables for crystallography G. Definition and exchange of crystallographic data*. Springer, Dordrecht, pp 295–443
- Goh BC, Hadden JA, Bernardi RC, Singharoy A, McGreevy R, Rudack T, Cassidy CK, Schulten K (2016) Computational methodologies for real-space structural refinement of large macromolecular complexes. *Annu Rev Biophys* 45:253–278. <https://doi.org/10.1146/annurev-biophys-062215-011113>
- Gore S, Sanz Garcia E, Hendrickx PMS, Gutmanas A, Westbrook JD, Yang H, Feng Z, Baskaran K, Berrisford JM, Hudson BP, Ikegawa Y, Kobayashi N, Lawson CL, Mading S, Mak L, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Peisach E, Sahni G, Sekharan MR, Sen S, Shao C, Smart OS, Ulrich EL, Yamashita R, Quesada M, Young JY, Nakamura H, Markley JL, Berman HM, Burley SK, Velankar S, Kleywegt GJ (2017) Validation of the structures in the protein data bank. *Structure* 25:1916–1927. <https://doi.org/10.1016/j.str.2017.10.009>
- Karmali AM, Blundell TL, Furnham N (2009) Model-building strategies for low-resolution X-ray crystallographic data. *Acta Crystallogr Ser D* 65(Pt 2):121–127. <https://doi.org/10.1107/S0907444908040006>
- Kinjo AR, Bekker GJ, Suzuki H, Tsuchiya Y, Kawabata T, Ikegawa Y, Nakamura H (2017) Protein data bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Res* 45(D1):D282–D288. <https://doi.org/10.1093/nar/gkw962>
- Phillips GN Jr (1986) Construction of an atomic model for tropomyosin and implications for interactions with actin. *J Mol Biol* 192(1):128–131
- Protein Data Bank (1971) Crystallography: protein data bank. *Nature New Biol* 233(42):223–223. <https://doi.org/10.1038/newbio233223b0>
- Rose PW, Prlc A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, Green RK, Goodsell DS, Hudson B, Kalro T, Lowe R, Peisach E, Randle C, Rose AS, Shao C, Tao YP, Valasatava Y, Voigt M, Westbrook JD, Woo J, Yang H, Young JY, Zardecki C, Berman HM, Burley SK (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* 45(D1):D271–D281. <https://doi.org/10.1093/nar/gkw1000>
- Shi Y, Fernandez-Martinez J, Tjioe E, Pellarin R, Kim SJ, Williams R, Schneidman-Duhovny D, Sali A, Rout MP, Chait BT (2014) Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol Cell Proteomics* 13(11):2927–2943. <https://doi.org/10.1074/mcp.M114.041673>
- Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. *Nucleic Acids Res* 36(Database issue):D402–D408. <https://doi.org/10.1093/nar/gkm957>
- Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P, Hendrickx PM, Lagerstedt I, Mir S, Fernandez Montecelo MA, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Sanz-Garcia E, Sen S, Slowley RA, Wainwright ME, Deshpande MS, Iudin A, Sahni G, Salavert Torres J, Hirshberg M, Mak L, Nadzirin N, Armstrong DR, Clark AR, Smart OS, Korir PK, Kleywegt GJ (2016) PDBe: improved

- accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res* 44(D1):D385–D395. <https://doi.org/10.1093/nar/gkv1047>
- Webb B, Viswanath S, Bonomi M, Pellarin R, Greenberg CH, Saltzberg D, Sali A (2018) Integrative structure modeling with the integrative modeling platform. *Protein Sci* 27(1):245–258. <https://doi.org/10.1002/pro.3311>
- Young JY, Westbrook JD, Feng Z, Sala R, Peisach E, Oldfield TJ, Sen S, Gutmanas A, Armstrong DR, Berrisford JM, Chen L, Chen M, Di Costanzo L, Dimitropoulos D, Gao G, Ghosh S, Gore S, Guranovic V, Hendrickx PMS, Hudson BP, Igarashi R, Ikegawa Y, Kobayashi N, Lawson CL, Liang Y, Mading S, Mak L, Mir MS, Mukhopadhyay A, Patwardhan A, Persikova I, Rinaldi L, Sanz-Garcia E, Sekharan MR, Shao C, Swaminathan GJ, Tan L, Ulrich EL, van Ginkel G, Yamashita R, Yang H, Zhuravleva MA, Quesada M, Kleywegt GJ, Berman HM, Markley JL, Nakamura H, Velankar S, Burley SK (2017) OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure* 25(3):536–545. <https://doi.org/10.1016/j.str.2017.01.004>

# Chapter 3

## View from Nuclear Magnetic Resonance Spectroscopy



John L. Markley

**Abstract** Nuclear magnetic resonance (NMR) spectroscopy is one of the three major approaches for determining the structures of biological macromolecules. Historically, NMR spectroscopy was number two after X-ray crystallography in the rate of depositions to the Protein Data Bank (PDB). However, electron cryomicroscopy (CryoEM) recently surpassed NMR in this regard. NMR frequently is used in conjunction with X-ray or CryoEM in structure determinations. NMR has advantages over the other structural approaches in studies of conformational dynamics and interconverting conformational states of proteins and nucleic acids in solution. NMR spectroscopy, itself, can be considered as collection of hybrid methods in that structure determinations rely on the results of several separate magnetic resonance experiments that measure connectivities of magnetic-resonance-active nuclei through covalent bonds or through space or determine relative orientations of magnetic dipoles. NMR results frequently are combined with data from small-angle X-ray scattering or chemical crosslinking in developing structural models. NMR spectroscopy and CryoEM are particularly synergistic in that neither requires crystallization.

**Keywords** NMR · Spectral assignment · Structure determination · Data visualization

Unlike X-ray crystallography, where a set of diffraction data collected with a single crystal can be sufficient to determine a structure, provided that the phase problem can be solved, NMR structural studies always require the collection of data sets from several different experiments from one sample, and frequently from multiple samples (Marion 2013). In this regard, NMR spectroscopy is, in itself, a hybrid method. Structures are determined through the combined analysis of results from a

---

J. L. Markley (✉)

Biochemistry Department, University of Wisconsin-Madison, Madison, WI, USA  
e-mail: [jmarkley@wisc.edu](mailto:jmarkley@wisc.edu)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,  
Advances in Experimental Medicine and Biology 1105,  
[https://doi.org/10.1007/978-981-13-2200-6\\_3](https://doi.org/10.1007/978-981-13-2200-6_3)

19



variety of NMR experiments. Structures derived from, or including, NMR data can benefit from their combination with information from other approaches. A recent review discusses the general derivation of structural information from a variety of types of experimental measurements, including NMR spectroscopy (van Gunsteren et al. 2016). Results from small angle X-ray scattering are frequently used now as an adjunct to NMR data, either to constrain the overall shape of the molecule or to position subunits, whose structures were determined by NMR, in an oligomeric structure, and these approaches with proteins have been reviewed recently (Mertens and Svergun 2017; Venditti et al. 2016; Prischi and Pastore 2016). RNA structures determined from NMR data are also benefitting from hybrid methods (Duss et al. 2015; Schlundt et al. 2017; Cornilescu et al. 2016). NMR structures of subunits or separately-folding fragments have been successfully incorporated into CryoEM images to improve the overall resolution as reviewed in (Cuniasse et al. 2017). In addition, NMR data can be used to phase crystallographic data (Zhang et al. 2014).

NMR spectroscopy can be carried out with samples in solution (solution NMR) or in the solid state (ssNMR). Well-developed protocols have been developed for determining solution NMR structures of proteins up to about 60 kDa (Cavanagh et al. 2010), RNA molecules up to 100 kDa (Barnwal et al. 2017), and large protein-RNA complexes (Yadav and Lukavsky 2016). Sparse structural and functional information can be obtained with proteins as large as 900 kDa (Sprangers et al. 2007; Fiaux et al. 2002). Peaks from solution NMR broaden with increased molecular weight as a consequence of slower molecular tumbling and become less well resolved as a consequence of the larger number of signals within the spectral window. These problems can be overcome, in part, by collecting NMR spectra in multiple dimensions, and/or by simplifying spectra by selective labeling with  $^2\text{H}$ ,  $^{13}\text{C}$ , and/or  $^{15}\text{N}$ . Typically, uniform labeling with  $^{13}\text{C}$  and  $^{15}\text{N}$  is used with proteins up to 20–25 kDa, and this labeling pattern is supplemented by  $^2\text{H}$  labeling of carbon-bound hydrogens for proteins above 25 kDa (Gardner and Kay 1998). Selective labeling of side-chain methyls of Ala, Ile, Leu, Met, Thr, and/or Val with  $^{-13}\text{CH}_3$  is a strategy used with still larger proteins (Tugarinov and Kay 2005). More elaborate labeling patterns can be achieved by segmental labeling (Liu et al. 2009), residue-selective labeling, alternate  $^{13}\text{C}$ - $^{12}\text{C}$  labeling (Takeuchi et al. 2010, 2011), or incorporation of amino acids with tailored stereospecific labeling optimized for NMR (Kainosho et al. 2006). Although the widths of ssNMR signals do not suffer from molecular weight dependence, spectral resolution can be improved by isotope labeling, such as fractional deuterium labeling or  $^{13}\text{C}$  labeling schemes that minimize directly bound  $^{13}\text{C}$ - $^{13}\text{C}$  pairs.

Assessing the information content of NMR spectra requires that signals be assigned to the individual nuclei ( $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ ) that generate them. Considerable progress has been made in simplifying this task in solution NMR through automated spectral analysis combined with computer graphics tools that permit the visualization of potential assignments along with the underlying data (Lee et al. 2016). Similar tools for solid state NMR are under development. Structural information comes from a variety of experimental parameters. The patterns of backbone and  $^{13}\text{C}^\beta$  chemical shifts are fairly accurate predictors of secondary structure ( $\alpha$ -helix

or  $\beta$ -strand).  $^1\text{H}$ - $^1\text{H}$  NOEs report on short interproton distances up to 5–6 Å, and residual dipolar couplings (RDCs) report on the directions of bond vectors.

Although, NMR spectra provide information about individual nuclei and their interactions, the resulting structures are underdetermined because the number of spectral parameters is always many fewer than those needed to specify atom positions. To cope with this problem, NMR structural models are represented as a family of conformers that are consistent with the available data and whose differences represent the uncertainty in specifying atomic positions.

## References

- Barnwal RP, Yang F, Varani G (2017) Applications of NMR to structure determination of RNAs large and small. *Arch Biochem Biophys* 628:42–56. <https://doi.org/10.1016/j.abb.2017.06.003>
- Cavanagh J, Fairbrother WJ, Palmer AG, Skelton NJ, Rance M (2010) *Protein NMR spectroscopy principles and practice*
- Cornilescu G, Didychuk AL, Rodgers ML, Michael LA, Burke JE, Montemayor EJ, Hoskins AA, Butcher SE (2016) Structural analysis of multi-helical RNAs by NMR-SAXS/WAXS: application to the U4/U6 di-snRNA. *J Mol Biol* 428 (5 Pt A):777–789. <https://doi.org/10.1016/j.jmb.2015.11.026>
- Cuniassé P, Tavares P, Orlova EV, Zinn-Justin S (2017) Structures of biomolecular complexes by combination of NMR and cryoEM methods. *Curr Opin Struct Biol* 43:104–113. <https://doi.org/10.1016/j.sbi.2016.12.008>
- Duss O, Yulikov M, Allain FH, Jeschke G (2015) Combining NMR and EPR to determine structures of large RNAs and protein-RNA complexes in solution. *Methods Enzymol* 558:279–331. <https://doi.org/10.1016/bs.mie.2015.02.005>
- Fiaux J, Bertelsen EB, Horwich AL, Wüthrich K (2002) NMR analysis of a 900K GroEL GroES complex. *Nature* 418(6894):207–211
- Gardner KH, Kay LE (1998) The use of  $^2\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$  multidimensional NMR to study the structure and dynamics of proteins. *Annu Rev Biophys Biomol Struct* 27:357–406
- Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Mei Ono A, Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature* 440(7080):52–57
- Lee W, Cornilescu G, Dashti H, Eghbalnia HR, Tonelli M, Westler WM, Butcher SE, Henzler-Wildman KA, Markley JL (2016) Integrative NMR for biomolecular research. *J Biomol NMR* 64(4):307–332. <https://doi.org/10.1007/s10858-016-0029-x>
- Liu D, Xu R, Cowburn D (2009) Segmental isotopic labeling of proteins for nuclear magnetic resonance. *Methods Enzymol* 462:151–175. [https://doi.org/10.1016/S0076-6879\(09\)62008-5](https://doi.org/10.1016/S0076-6879(09)62008-5)
- Marion D (2013) An introduction to biological NMR spectroscopy. *Mol Cell Proteomics* 12(11):3006–3025. <https://doi.org/10.1074/mcp.O113.030239>
- Mertens HDT, Svergun DI (2017) Combining NMR and small angle X-ray scattering for the study of biomolecular structure and dynamics. *Arch Biochem Biophys* 628:33–41. <https://doi.org/10.1016/j.abb.2017.05.005>
- Prischi F, Pastore A (2016) Application of nuclear magnetic resonance and hybrid methods to structure determination of complex systems. *Adv Exp Med Biol* 896:351–368. [https://doi.org/10.1007/978-3-319-27216-0\\_22](https://doi.org/10.1007/978-3-319-27216-0_22)
- Schlundt A, Tants JN, Sattler M (2017) Integrated structural biology to unravel molecular mechanisms of protein-RNA recognition. *Methods* 118–119:119–136. <https://doi.org/10.1016/j.ymeth.2017.03.015>
- Sprangers R, Velyvis A, Kay LE (2007) Solution NMR of supramolecular complexes: providing new insights into function. *Nat Methods* 4(9):697–703

- Takeuchi K, Frueh DP, Sun ZYJ, Hiller S, Wagner G (2010) CACA-TOCSY with alternate C-13-C-12 labeling: a C-13(alpha) direct detection experiment for main chain resonance assignment, dihedral angle information, and amino acid type identification. *J Biomol NMR* 47(1):55–63. <https://doi.org/10.1007/s10858-010-9410-3>
- Takeuchi K, Gal M, Takahashi H, Shimada I, Wagner G (2011) HNCA-TOCSY-CANH experiments with alternate  $^{13}\text{C}$ - $^{12}\text{C}$  labeling: a set of 3D experiment with unique supra-sequential information for mainchain resonance. *J Biomol NMR* 49(1):17–26. <https://doi.org/10.1007/s10858-010-9456-2>
- Tugarinov V, Kay LE (2005) Methyl groups as probes of structure and dynamics in NMR studies of high-molecular-weight proteins. *Chembiochem* 6(9):1567–1577
- van Gunsteren WF, Allison JR, Daura X, Dolenc J, Hansen N, Mark AE, Oostenbrink C, Rusu VH, Smith LJ (2016) Deriving structural information from experimentally measured data on biomolecules. *Angew Chem Int Ed Engl* 55(52):15990–16010. <https://doi.org/10.1002/anie.201601828>
- Venditti V, Egner TK, Clore GM (2016) Hybrid approaches to structural characterization of conformational ensembles of complex macromolecular systems combining NMR residual dipolar couplings and solution X-ray scattering. *Chem Rev* 116:6305–6322. <https://doi.org/10.1021/acs.chemrev.5b00592>
- Yadav DK, Lukavsky PJ (2016) NMR solution structure determination of large RNA-protein complexes. *Prog Nucl Magn Reson Spectrosc* 97:57–81. <https://doi.org/10.1016/j.pnmrs.2016.10.001>
- Zhang W, Zhang T, Zhang H, Hao Q (2014) Crystallographic phasing with NMR models: an envelope approach. *Acta Crystallogr D Biol Crystallogr* 70(Pt 7):1977–1982. <https://doi.org/10.1107/S1399004714009754>

**Part II**  
**New Experimental Tools Enabling Hybrid**  
**Methods**

# Chapter 4

## Complementary Use of Electron Cryomicroscopy and X-Ray Crystallography: Structural Studies of Actin and Actomyosin Filaments



Takashi Fujii and Keiichi Namba

**Abstract** Visualization of macromolecular structures is essential for understanding the mechanisms of biological functions because they are all determined by the structure and dynamics of macromolecular complexes. Electron cryomicroscopy (cryoEM) and image analysis has become a powerful tool for structural studies because of recent technical developments in microscope optics, cryostage control, image detection and the methods of sample preparation. In particular, the recent development of CMOS-based direct electron detectors with high sensitivity, high resolution and high frame rate has revolutionized the field of structural biology by making near-atomic resolution structural analysis possible from small amounts of solution samples. However, for some biological systems, it is still difficult to reach high resolution due to somewhat flexible nature of the structure, and a complementary use of cryoEM with X-ray crystallography is essential and useful to gain mechanistic understanding of the biological functions and mechanisms. We will describe our strategy for the structural analyses of actin filament and actomyosin rigor complex and the biological insights we gained from these structures.

**Keywords** Hybrid method for structural analysis · Electron cryomicroscopy · Image analysis · 3D reconstruction · X-ray crystallography · F-actin assembly · Treadmill · Actomyosin motor · Skeletal muscle contraction · Biased Brownian motion

---

T. Fujii · K. Namba (✉)

Graduate School of Frontier Biosciences, Osaka University, Suita, Osaka, Japan

Quantitative Biology Center, Riken, Osaka, Japan

e-mail: [keiichi@fbs.osaka-u.ac.jp](mailto:keiichi@fbs.osaka-u.ac.jp)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,

Advances in Experimental Medicine and Biology 1105,

[https://doi.org/10.1007/978-981-13-2200-6\\_4](https://doi.org/10.1007/978-981-13-2200-6_4)

## 4.1 Introduction

Biological functions and activities that support the life of every biological organism are diverse, and yet the basic mechanisms that determine and exert those biological functions are highly shared by diverse organisms, from microorganisms such as bacteria and yeast to multicellular organisms such as animals and plants. Even the complex human brain functions are not the exception. The basic mechanisms are highly shared because all these functions are designed and determined by the structures of proteins and nucleic acids with complex three-dimensional (3D) arrangements of so many atoms that comprise these molecules, with the number ranging from a few to tens and hundreds of thousands. Moreover, their structures are not solid unlike bulk materials of metals and ceramics but are very dynamic and flexible so that they can function by actively utilizing thermal fluctuations. One of the major challenges in life science is the elucidation of mechanisms that determine and exert these extremely diverse functions by looking into the 3D structures and dynamics of so many different biological macromolecules involved in those diverse biological functions. We also need to look at the structures of macromolecules in each of their functional states appearing in the entire process of their functional cycles. Therefore the number of 3D structures we need to solve would be extremely large, probably ranging at least from a few hundreds of thousands to a few million.

Thus, structural information of biological macromolecular machinery is essential for understanding the mechanisms by which they function, and various methods for structural analyses have been developed to obtain structural information at highest possible resolution. We have been studying the structures and functions of protein motor complexes, such as the bacterial flagellar motor and actomyosin, to understand the mechanisms of force generation and highly efficient energy conversion. We have developed various techniques in X-ray fiber diffraction, X-ray crystallography and electron cryomicroscopy (cryoEM) and used them in a complementary manner to build atomic models of the motor complexes by docking crystal structures of component proteins into 3D density maps obtained by X-ray fiber diffraction and/or cryoEM and refining the entire models against these maps (Namba et al. 1985; Namba and Stubbs 1985, 1986; Samatey et al. 2001, 2004; Yonekura et al. 2003; Fujii et al. 2009, 2010; Gayathri et al. 2012; Fujii and Namba 2017). Although cryoEM image analysis has become a powerful tool for the structural analysis of macromolecular complexes by the recent introduction of direct electron detecting CMOS cameras and is now capable of resolving the structures at near atomic detail to allow *de novo* atomic model building as described in the following section, there are still many cases where the resolution is limited by the flexible and/or dynamic nature of the specimens, and a complementary use of cryoEM for the entire complex and X-ray crystallography or NMR of component molecules is necessary and useful to build the entire atomic model to study the structure-function relationships in such cases. We will describe a few example cases to demonstrate the usefulness of the method.

## 4.2 Power of cryoEM Image Analysis in the Past and Present

CryoEM image analysis, especially single particle image analysis, is a potentially powerful method because there is no need for sample crystallization that is essential for X-ray crystallography and there is virtually no upper limit in the size of molecular complexes unlike NMR. The structures of macromolecular complexes can be directly visualized by cryoEM in various functional states. It would therefore be desirable that cryoEM can visualize the structures of the macromolecular complexes at atomic resolution. The 2017 Nobel Prize in Chemistry was awarded to Jacques Dubochet (University of Lausanne, Switzerland), Joachim Frank (Columbia University, USA), and Richard Henderson (MRC Laboratory of Molecular Biology, UK), for their pioneering works in 1970s and 1980s in the development of cryoEM image analysis techniques for the structural analysis of biological macromolecules. By the development of transmission electron cryomicroscopes (cryoTEM) over many years in 1980s and 90's, especially those done in Japan, such as the implementation of a liquid helium-cooled specimen stage to minimize the radiation damage (Fujiyoshi et al. 1991) and a field emission electron gun to use a highly-coherent electron beam (Mimori et al. 1995), as well as various improvements in the method of image analysis, it became possible to achieve near atomic resolution for 2D crystal structures of membrane proteins, such as bacteriorhodopsin and aquaporin (Kimura et al. 1997; Mitsuoka et al. 1999; Murata et al. 2000) and filamentous helical assemblies of proteins, such as the bacterial flagellar filament (Yonekura et al. 2003). It was encouraging to see the polypeptide backbone folding and large side chains of flagellin clearly resolved in the structure of the bacterial flagellar filament at around 4 Å resolution analyzed from a set of filament images corresponding to only 40,000 flagellin molecules. Since the image quality and signal to noise ratio (S/N) of frozen-hydrated biological macromolecules embedded in vitreous ice is quite poor due to an extremely low electron dose to avoid radiation damage, a high cryo-protection factor by lowering the specimen temperature down to 4 K by liquid helium gave us a substantial advantage for achieving unprecedented resolution. However, it was by no means a high-throughput work partly because we had to use photographic films as the image detector.

By further implementation of new technologies in cryoTEM in 2000s, such as the CCD camera to evaluate the image quality immediately after recording by Fourier transformation and in-column energy filter to eliminate inelastically scattered electrons that form a high background noise, and working at a slightly elevated specimen temperature to around 50 K to increase the electron conductivity of the ice embedded specimen to reduce its charge up that tends to blur the cryoEM images, the efficiency of high-quality image data collection was drastically improved, and the image analysis by the computer became much faster by the improvement in the software and semiconductor nanotechnologies. These improvements made previously several years of works be done within a few weeks, and the visualization of protein secondary structures became relatively easy and quick (Fujii et al. 2009; Fujii et al. 2010; Gayathri et al. 2012), demonstrating a potential of achieving near

atomic resolution within such a short period of time as far as the structure is well ordered and stable, such as tobacco mosaic virus.

Then, at the end of 2013, two milestone papers were published by Yifan Cheng and his colleagues on the structure of the TRPV1 receptor ion channel, a membrane receptor protein that responds to heat and spiciness, solved at 3.4 Å resolution by cryoEM image analysis of about 100,000 single particle images of the protein picked up from about 1000 cryoEM images obtained from a small amount of sample solution (Liao et al. 2013; Cao et al. 2013). They were involved in the development of a CMOS-based direct electron detector camera and fully utilized its capability to record images of 4 K × 4 K pixels at 400 frames per second to carry out single electron counting to minimize the detection noise called the Landau noise, which is an intrinsic noise of large distribution that any types of energy accumulating detectors, such as film and CCD, suffer for individual electron detection. They also devised a way to collect sharp high-quality cryoEM images of proteins by movie-mode imaging and motion correction to minimize the image blur caused by a mechanical drift of the specimen stage and the distortion of ice film caused by electron irradiation (Li et al. 2013). Together with the development of a user-friendly, yet sophisticated image analysis software package, RELION (Scheres 2012; Kimanius et al. 2016), cryoEM image analysis has now become a very powerful tool for structural biology, achieving near atomic resolution in the structural analysis of many different macromolecular complexes including membrane proteins to allow *de novo* atomic model building relatively easily.

However, there are still many cases where the resolution is limited by the flexible and/or dynamic nature of the specimens, and in such cases a complementary use of cryoEM for the entire complex and X-ray crystallography or NMR of component molecules is necessary and useful to study the structure-function relationships. We will describe our structural studies of the skeletal muscle F-actin and actomyosin complex to demonstrate the usefulness of the complementary method.

### 4.3 Structural Study of F-Actin

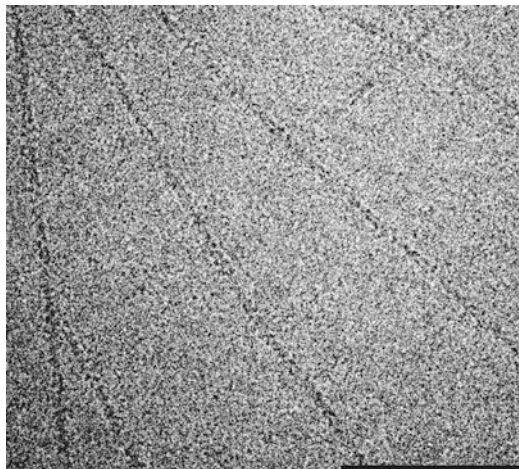
F-actin is a helical assembly of actin, is an essential component of muscle fibers for contraction and also plays crucial roles in various cellular processes as the most abundant component and regulator of cytoskeletons by dynamic assembly and disassembly processes (from G-actin to F-actin and vice versa), such as those called lamellipodia and filopodia (Pollard and Borisy 2003; Carlier and Pantaloni 2007). While actin is a ubiquitous protein and is involved in the various important biological functions and many crystal structures of actin were available over the years since the first crystal structure in complex with DNase-I (Kabsch et al. 1990), the definitive high-resolution structure of F-actin remained unknown until 2010 (Fujii et al. 2010). Steady technical advances in cryoEM image analysis over the years allowed near-atomic resolution structural analyses of many icosahedral viruses and helical assembly of macromolecules, such as the bacterial flagellar



filament, the tubular crystal of acetylcholine receptor and tobacco mosaic virus (TMV) in 2000s (Yonekura et al. 2003; Miyazawa et al. 2003; Sachse et al. 2007). But, it was possible to reach such high resolutions even by using photographic films as image detectors simply because their particle sizes or diameters were large enough to produce sufficiently high image contrast and S/N in their cryoEM images of ice-embedded frozen-hydrated specimens that allows accurate alignment and average of many particle images necessary to recover high-resolution structural information hidden under the noise. Since F-actin is a relatively thin filament with a flexible, twisted ribbon-like structure with the maximum diameter of only 10 nm, which is far thinner than TMV (18 nm) or the flagellar filament (23 nm), the image contrast of unstained, frozen-hydrated specimen is extremely low, making accurate image alignment extremely difficult and thereby high-resolution structural analysis elusive.

We used a cryoTEM (JEOL JEM-3200FSC) equipped with a field emission electron gun, a liquid helium-cooled specimen stage, an in-column  $\Omega$ -type energy filter, and a CCD camera (TIPVS F415MP) to collect cryoEM images of F-actin. We were able to obtain a remarkable gain ( $\sim 5$  times) in image contrast by the use of energy filtering, by controlling ice thickness, and by recording images at a specimen temperature of 50 K instead of 4 K (Fujii et al. 2009). Such improvement in image contrast allowed us to see the two-stranded helical features of F-actin in raw cryoEM images even at relatively small defocus levels close to 1  $\mu\text{m}$  (Fig. 4.1). Image recording by a CCD camera made high-quality data collection remarkably efficient. To avoid undesirable dumping of high-resolution contrast by a poor modulation transfer function of the CCD camera, we used a relatively high magnification of approximately  $172,000\times$  ( $0.87 \text{ \AA}/\text{pixel}$ ). We collected 490 cryoEM images manually in two days, picked up filament images and used a single particle image analysis method but still utilized the helical symmetry to make the image alignment as accurate as possible (Sachse et al. 2007; Fujii et al. 2009; Egelman

**Fig. 4.1** CryoEM image of F-actin in a frozen hydrated state recorded by CCD under a defocus value of 1500 nm. Scale bar, 100 nm

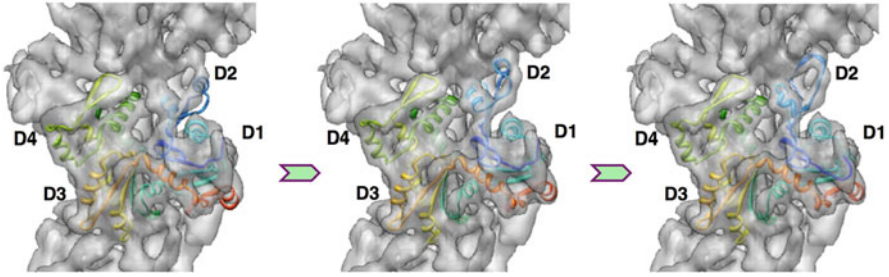


2000). Since the image analysis procedure was fully automated, it was completed within 2 days to reconstruct the final 3D image, and the resolution was 6.6 Å (at the Fourier shell correlation (FSC) = 0.143) (EMD-5168) (Fujii et al. 2010).

The resolution of the 3D map was high enough to clearly visualize the secondary structures, such as  $\alpha$ -helices,  $\beta$ -sheets and  $\beta$ -hairpins, and even some loops and the extended N-terminal chain that had never been seen in the crystal structures clearly showed up. So it was possible to build a complete atomic model of F-actin far more reliably than before. It was debated over long time that F-actin must have an intrinsic flexibility in its helical order and that is why the structures solved by cryoEM image analysis were limited to low resolution, but the fact that such a high resolution was achieved as described above by using over 90% of the image data we collected indicates that the flexibility is not so high as the previous studies suggested (Galkin et al. 2008).

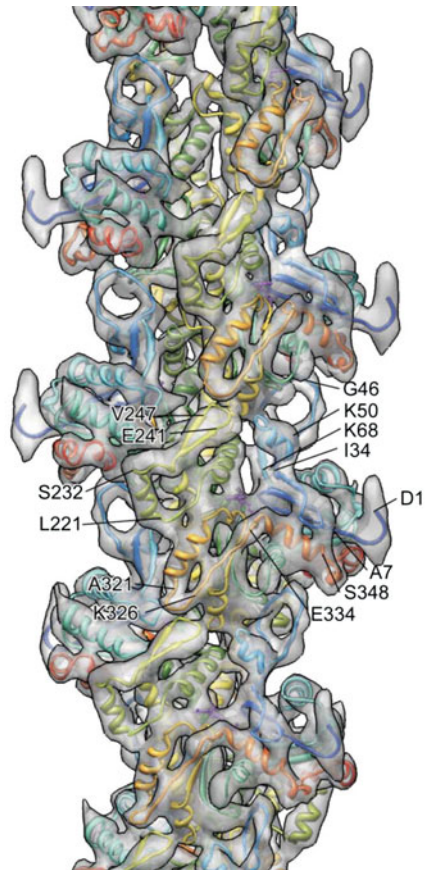
To build an atomic model of F-actin, we employed a program FlexEM (Topf et al. 2008), which refines the atomic model while fitting it into the EM density map by simulated annealing molecular dynamics with stereochemical and non-bonded interaction terms restrained. We used the crystal structure of uncomplexed actin (PDB code: 1J6Z) (Otterbein et al. 2001) as an initial model and divided it into four domains D1, D2, D3 and D4 to treat them as independent units because these four domains have well-defined hydrophobic cores. In the initial stage of the fitting process, we treated them as rigid bodies and allowed the joints of these domains to be flexible, but residues 1–8, 39–56, 221–234 and 337–375 were outside the density map. In the second stage, we allowed these residues to move flexibly to fit into the map under stereochemical restraints and then applied the helical symmetry of F-actin to this subunit model to build a complete F-actin model. We then minimized the conformational energy further by FlexEM to remove intermolecular clashes of atoms. The processes of the fitting and refinement are shown in Fig. 4.2, and the final refined model in Fig. 4.3 (PDB: 3FMP) (Fujii et al. 2010). The conformation of domains 1, 3 and 4 did not change so largely as indicated by the relatively small root-mean-squares (rms) displacements of C $\alpha$  atoms (domain 1: 0.3 Å; domain 3: 0.3 Å; domain 4: 0.8 Å). This is consistent with the fact that these three domains have stable conformations with well-defined hydrophobic cores and assures the reliability of the atomic model as well as the high quality of the cryoEM map. Domain 2 was, however, an exception. The 2-turn short  $\alpha$ -helix (residues 40–48) at the tip of the D-loop (the DNase I binding loop) in the actin crystal structure (Otterbein et al. 2001) became an extended loop (residues 38–53), reaching the bottom pocket between domains D1 and D4 of the above actin subunit (Fig. 4.3). Such a conformational change had been predicted from its variable conformations in the crystal structures of actin and its possible involvement in the axial intersubunit interactions (Oda et al. 2009), but this D-loop conformation was unique, indicating that it is totally dependent upon the molecule that it binds to.

Since the nature of conformational change from G-actin to F-actin is of immense importance for biological implications for actin functions, we carefully compared the F-actin structure with the crystal structure of G-actin. While the two major-domains were twisted in the crystal structures, they became flat in the F-actin



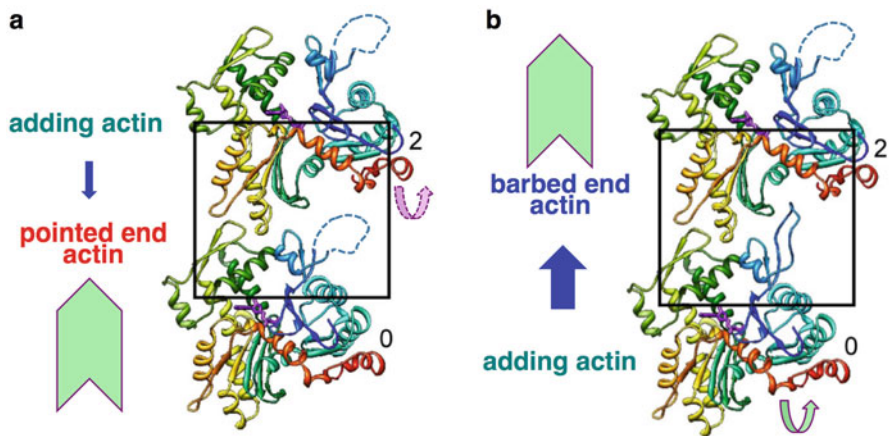
**Fig. 4.2** Process of docking and refinement of actin atomic model in the cryoEM density map from left to right. Four domains of actin are labeled D1, D2, D3 and D4. Left, G-actin crystal structure, presented as a  $C\alpha$  ribbon diagram, is docked into the cryoEM density map as a rigid body. Each domain is not well fitted to the density. Middle, each of the four domains is independently moved and rotated as a rigid body to fit to the density map. Domain D2 is still not fitted well. Right, the conformation of each domain is refined against the density map by flexible fitting

**Fig. 4.3** CryoEM density map of F-actin (EMD-5168) with a fitted and refined atomic model (PDB: 3MFP) (Fujii et al. 2010). The model is presented as a  $C\alpha$  ribbon diagram colored in rainbow from the N-terminus in blue to the C-terminus in red. Approximately seven subunits of actin are shown. Some amino acid residues are labeled as a guild to follow the chain



model in a similar manner to the relative domain motions described previously for the model that nicely reproduced X-ray fiber diffraction intensity data obtained from a highly oriented liquid-crystalline sol specimen of F-actin (Oda et al. 2009). However, the relative domain motions were more complex than those described previously. Together with the conformational change of the D-loop, these changes made the slightly bent domains 1–2 in G-actin significantly flatter in F-actin, allowing the D-loop to reach and bind to the bottom pocket between domains D1 and D4 of the above actin subunit. This is how the axial intersubunit interactions along the protofilament are made tight for F-actin polymerization as shown in Fig. 4.3. Including the interactions between protofilaments, the nature of intersubunit interactions between actin subunits is mostly electrostatic or hydrophilic, and this explains depolymerization of F-actin at concentrated salt solutions (Nagy and Jencks 1965).

Actin polymerization is known to have a distinct polarity, showing fast polymerization at the barbed end of F-actin while slow depolymerization from the pointed end under certain conditions (Fujiwara et al. 2007). This is called treadmilling and plays important roles in the formation of lamellipodia and filopodia for cell motility and morphogenesis (Pollard and Borisy 2003; Carlier and Pantaloni 2007). The conformational changes of actin between its monomeric G-actin form and polymerized F-actin form explains how this asymmetry is achieved (Fig. 4.4). Actin



**Fig. 4.4** Structural asymmetry of F-actin responsible for the difference in the assembly kinetics at the pointed and barbed ends. **(a)** An actin subunit shown above is being added to the pointed end of F-actin shown below. The flexible D-loop of actin at the pointed end is presented by dashed line. The domain motion of adding actin occurs but its F-actin conformation cannot be stabilized, as indicated by purple dashed arrow, due to the flexible D-loop of actin at the pointed end. **(b)** An actin subunit shown below is being added to the barbed end of F-actin shown above. Because the bottom pocket of actin at the barbed end is well ordered and has a stable F-actin conformation to act as the template for actin assembly, the D-loop of adding actin binds to the pocket and is stabilized to make the entire adding actin conformation stable in the F-actin form after domain motion, as indicated by green solid arrow

at the barbed end is stably in the F-actin conformation, forming the bottom pocket for the binding of actin subunit in the G form. The structure of the bottom pocket acts as the template for the formation of the D-loop with the above mentioned domain motions of adding actin to turn it into the F form, and this facilitates the polymerization of actin. On the other hand, because actin at the pointed end has domain D2 exposed to solution, the D-loop conformation cannot be stabilized at all. The exposed D-loop of domain D2 must be flexible and dynamic to make the binding of adding actin rather difficult because adding actin also has to change its conformation from the G to F form in order to bind to F-actin but no stable template structure is available for these conformational changes to occur and be stabilized. Thus, the asymmetry in the structure and conformational dynamics of actin at the barbed and pointed ends of F-actin is responsible for the distinct difference in the polymerization kinetics of actin at the both ends. The complementary use of cryoEM and X-ray crystallography allowed us to gain deep insights into this biologically important mechanism.

#### **4.4 Structural Study of Skeletal Muscle Actomyosin Rigor Complex**

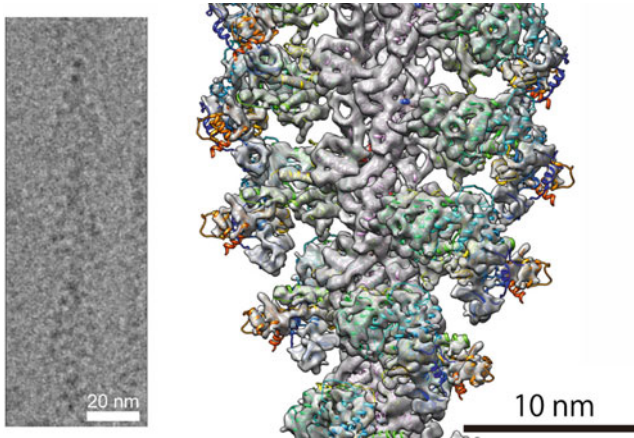
Muscle contraction occurs by mutual sliding of thick myosin filaments and thin actin filaments that shortens sarcomeres, the contractile units that regularly repeat along the entire muscle cells (Huxley 1969). The sliding force is generated via cyclic interactions of myosin heads, which are periodically projecting out from the thick filament towards surrounding thin actin filaments, with actin molecules of the thin filaments. Myosin head is an ATPase, and its ATP binding and hydrolysis regulates the cyclic association and dissociation of myosin with actin filament (Lymn and Taylor 1971). Upon binding of MgATP, myosin hydrolyses ATP relatively quickly but the hydrolysis products ADP and Pi stay in the nucleotide-binding pocket, and therefore its ATPase cycle does not proceed until myosin head binds to actin filament. Therefore, a conformational change of myosin head must occur upon binding to actin filament, and this should be responsible for this actin-activated ATPase, but structural information on the actomyosin rigor complex was limited to reveal this mechanism. X-ray crystal structures of the head domains of various myosins, such as myosin II, V and VI, in different nucleotide-binding states have suggested that myosin undergoes conformational changes during ATPase cycle in its lever arm domain to be in largely different angles within the plane of actin filament axis and that such changes represent a power stroke that drives the unidirectional movement of myosin against actin filament (Holmes et al. 2004; Sweeney and Houdusse 2004). However, since those myosin head structures obtained in atomic details were all in the absence of actin filament (Rayment et al. 1993; Dominguez et al. 1998; Bauer et al. 2000; Houdusse et al. 2000; Coureux et al. 2003; Reubold

et al. 2003, 2005; Mén  try et al. 2005, 2008; Yang et al. 2007), key piece of information was still missing.

The structure of the actomyosin rigor complex had been analyzed by electron cryomicroscopy (cryoEM) and image analysis (Holmes et al. 2003; Behrmann et al. 2012). However, the resolution and quality of the density maps were limited to reveal the conformational changes in sufficient detail, and it was still not so clear how ADP and Pi are released upon strong binding of myosin to actin filament and how ATP binding to myosin causes its dissociation from actin filament. We therefore solved the structure of actomyosin rigor complex of rabbit skeletal muscle by cryoEM image analysis. We obtained a 3D density map at 5.2   resolution (EMD-6664) and built an atomic model (PDB: 5H53) by using a method similar to that we used for F-actin as described in the previous section (Fig. 4.5) (Fujii and Namba 2017). We used the crystal structure of squid muscle myosin S1 fragment in the rigor-like state (PDB: 3I5G) (Yang et al. 2007) and the cryoEM structure of skeletal muscle F-actin (PDB: 3MFP) (Fujii et al. 2010) for docking and refinement. We employed DireX (Schroder et al. 2007) and FlexEM (Topf et al. 2008) to refine these models by flexible fitting while preserving stereochemistry. We carried out this model fitting refinement carefully to avoid overfitting, by imposing a relatively strong restraint to keep the conformations of individual domains with independent hydrophobic cores unchanged as much as possible and trying not to fit individual secondary structure elements separately. As a reliability measure of our model, the rms deviations of C  atoms for individual domains of myosin head of our rigor model from those of a crystal rigor-like model (PDB: 3I5G) (Yang et al. 2007) were calculated, and they were all with a range from 1.0 to 1.6  , which was comparable to those between crystal structures of myosin in different conformations, assuring that our model was refined without over fitting.

We then compared this structure with those of myosin in different nucleotide-binding states solved by X-ray crystallography and found a distinctly large conformational change of myosin head that widely opens up the nucleotide-binding pocket, even compared with the rigor-like structures of myosin head without nucleotide in the pocket (Fig. 4.6). It was obvious that this conformational change allows ADP and Pi to be quickly released from their binding sites upon myosin binding to actin filament. Myosin has been called a backdoor enzyme (Yount et al. 2007) because Pi leaves before ADP (Geeves et al. 1984) and a possible pathway for Pi release has been found only in the backside of the pocket in the myosin crystal structures (Yang et al. 2007; Llinas et al. 2015). However, the structure of actomyosin rigor state with such a widely open pocket (Fig. 4.6) suggests that Pi is likely to be released also from the front side. Although it is not obvious why Pi leaves before ADP, electrostatic repulsion by the negative charges of Pi or the way the ADP moiety is tightly bound by myosin may be responsible for this.

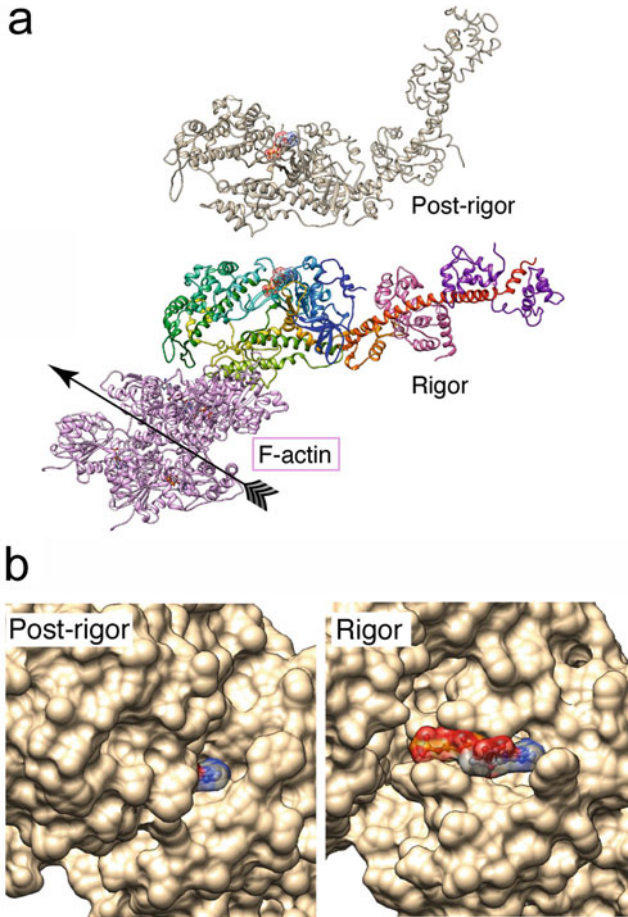
Recent publications on the structures of actomyosin rigor complexes by cryoEM image analysis revealed the structures of cytoplasmic myosins or smooth muscle myosin strongly bound to actin filament (von der Ecken et al. 2016; Wulf et al. 2016; Banerjee et al. 2017; Menten et al. 2018). They all show a similar conformational change of myosin head to those we observed for skeletal muscle myosin albeit in



**Fig. 4.5** CryoEM image and the reconstructed density map of actomyosin rigor complex (EMD-6664) with the model of actin and myosin after docking and refinement (PDB: 5H53) (Fujii and Namba 2017). The cryoEM image shows the typical arrowhead feature of the complex. About nine subunits of actin and myosin head are presented. Ribbon models of actin are colored purple and myosin in rainbow according to the sequence

much less extent, and those structures share the conformations with those of the crystal rigor-like structures with much less open nucleotide binding pocket that does not allow such a quick release of ADP and Pi as skeletal muscle myosin in the rigor state. The rates of ATP hydrolysis cycle of cytoplasmic and smooth muscle myosins are actually much slower than that of skeletal muscle myosin, and the same is true for the speed of myosin movement along actin filament. It appears that the structures of different types of myosins are optimally designed to move along actin filament at different speeds required for their physiological functions, and the rate of chemo-mechanical cycle is determined differently by their similar but distinct level of conformational changes.

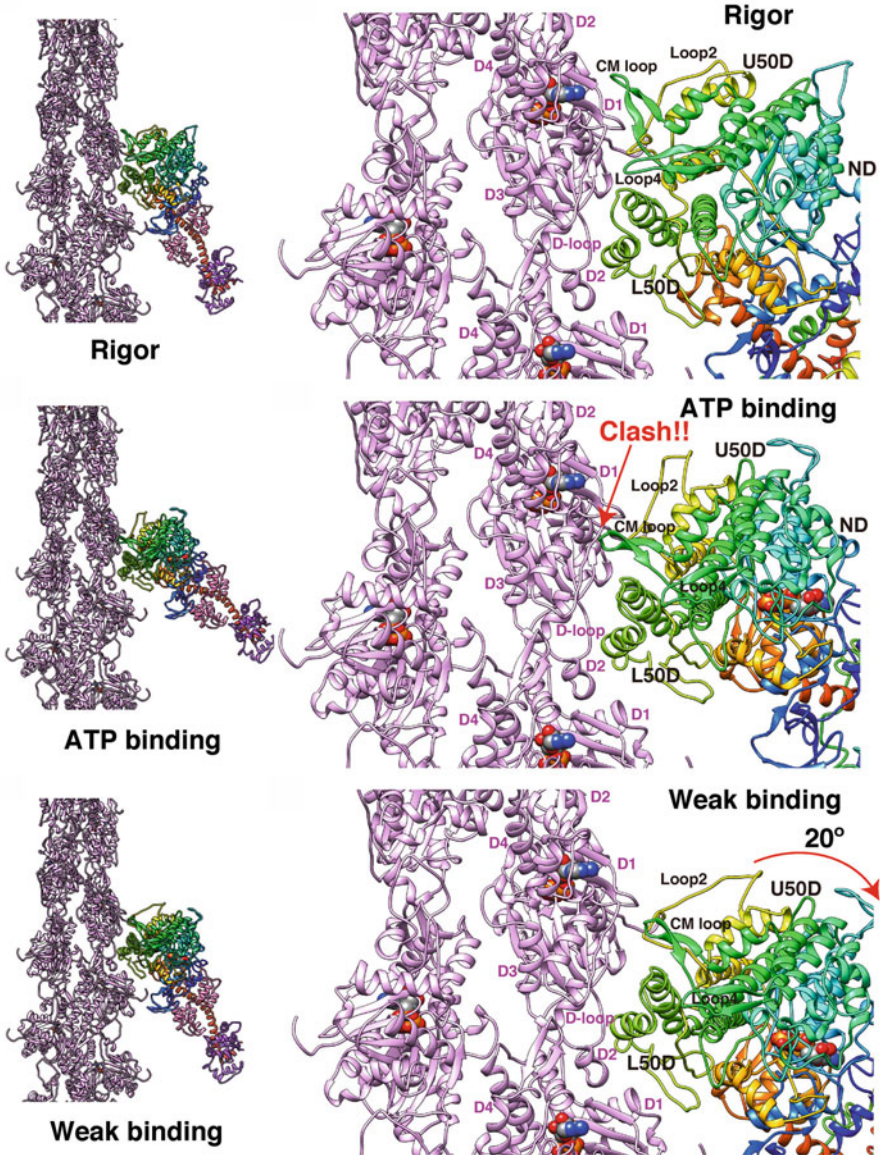
Structural comparison of our rigor model with an ATP bound post-rigor structure (Rayment et al. 1993) revealed how ATP binding may trigger dissociation of myosin from actin filament. We superposed myosin L50D domain (N473 – A593), which contains the helix-loop-helix tightly attached to two neighboring actin molecules along the protofilament (Fig. 4.7), to see what would occur in the actomyosin interactions upon ATP binding. We used L50D for superposition because this domain occupies the largest area of actomyosin interface. In the rigor structure, the CM loop and loop 4 are nicely fitted on and tightly bound to actin surface (domains D1 and D3, Fig. 4.7 top panel), but the post-rigor structure thus superimposed on the rigor structure shows a serious steric clash of the CM loop with domain D1 of actin (Fig. 4.7 middle panel). This clash is caused by U50D rotation nearly as a rigid body by  $21^\circ$  around the long axis of myosin head and appears to be the main cause of myosin dissociation from actin filament upon ATP binding. Assuming that L50D and loop 2 stay bound to both actin subunits with hydrophobic and electrostatic



**Fig. 4.6** Comparison of myosin structures in the actomyosin rigor state and a post-rigor state. **(a)** The post-rigor crystal structure of chicken muscle myosin (PDB: 2MYS) (Rayment et al. 1993) and the actomyosin rigor complex (PDB: 5H53) (Fujii and Namba 2017), viewed nearly in the axial direction of the filament from its barbed end. ATP is included in both models to indicate its binding position. **(b)** The nucleotide-binding sites of the two models in solid surface representation showing how widely the nucleotide-binding pocket is open when myosin head is bound strongly to actin filament in the rigor state

interactions, respectively, this CM loop clash against actin would push the CM loop back and cause a clockwise rotation of the entire motor domain by about  $20^\circ$  around its long axis to avoid the clash, and this results in a significant reduction in the interface area between myosin and two actin subunits to destabilize the actomyosin interactions (Fig. 4.7 bottom panel). This model would represent a possible structure of actomyosin in the weak binding state formed upon ATP binding, and this would be the state of myosin ready to dissociate from actin filament.





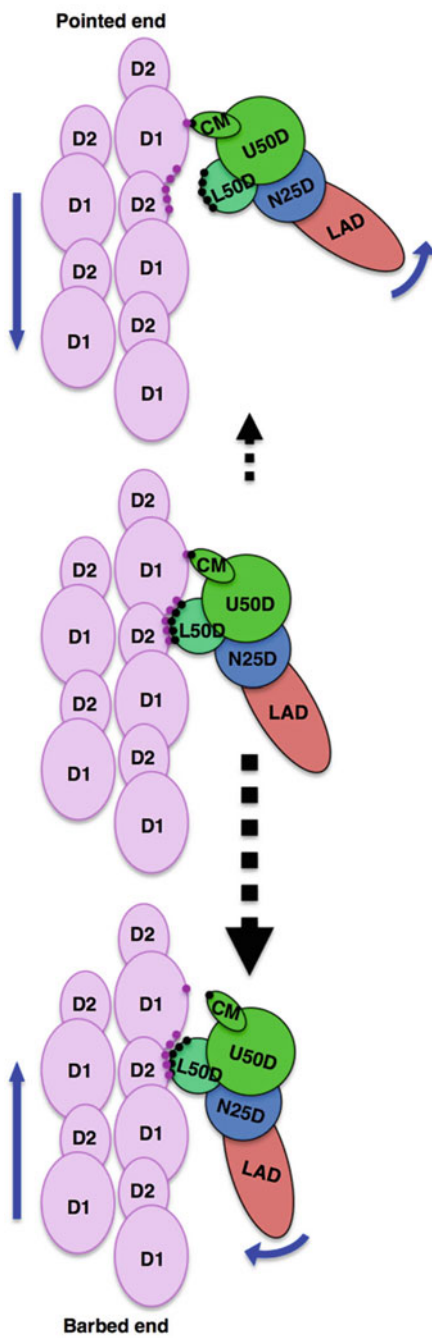
**Fig. 4.7** Conformational changes of rigor myosin head upon ATP binding and its possible consequence to form the weak binding state. Top panel shows the actomyosin rigor structure. The middle panel shows the myosin structure upon ATP binding with its L50D helix-loop-helix and loop 2 still attached to actin. The bottom shows myosin head after rotation to avoid the clash of CM loop with actin where L50D helix-loop-helix and loop 2 still attached to actin. Left panels are overviews, and the right panels are magnified. The N-terminal portion of loop 2 must be flexible enough to allow myosin head rotation while the lysine-rich C-terminal portion stays attached to the N-terminal region of actin to keep electrostatic interactions of the weak binding state. The crystal structure of chicken muscle myosin in the post-rigor state (PDB: 2MYS) was used to build the models shown in the middle and bottom panels by including loop 2 in different conformations to accommodate different distances between actin D1 and myosin U50D

A preferential binding of myosin to actin filament has been observed depending on the direction of relative motion and/or force (Iwaki et al. 2009). The asymmetry in the putative model of weakly bound actomyosin (Fig. 4.7, and schematically depicted in Fig. 4.8) can explain how such directionally preferential binding can be achieved. This structural asymmetry can also cause directionally preferential release of myosin upon ATP binding from actin filament, and the probability of dissociation is higher when actin filament moves forward to its pointed end than when actin filament moves backward to its barbed end. So the unidirectional sliding motions of myosin and actin filament could be achieved by just biasing their relative Brownian motions within each sarcomere by this directionally preferential release of myosin. This thermal-driven mechanism can explain why the sliding distance of myosin and actin filament in sarcomere is longer than 60 nm per one ATP hydrolysis cycle (Yanagida et al. 1985), which is much longer than the one predicted by the power stroke of myosin lever arm, and how a single myosin head can go through multiple steps of 5.3 nm along actin filament until myosin head strongly binds to actin by release of ADP and Pi when myosin is forced to stay near actin filament (Kitamura et al. 1999). These rather intriguing observations suggested the presence and involvement of a biased Brownian motion in the actomyosin motility mechanism, but how it can be achieved was elusive until we saw the molecular structures in detail. Thus, the complementary use of cryoEM and X-ray crystallography again played a very important role in revealing this biologically important mechanism.



**Fig. 4.8** Schematic diagram of actomyosin structure in the weak binding state, showing a possible mechanism of preferential transition to the strong binding state in the backward movement of actin filament (downward in this figure) and preferential release of myosin head from actin filament in the forward movement of actin filament (upward in this figure). Clockwise rotation of myosin by upward movement of actin filament (middle to bottom) can occur more easily than counterclockwise rotation by downward movement (middle to top), because the bonds between myosin and two actin subunits can be broken one after another by clockwise rotation, starting from those on the tip of CM loop (middle to bottom) but the tip of CM loop becomes the center or fulcrum of rotation by counterclockwise rotation and therefore many bonds between L50D and two actin subunits have to be broken simultaneously (middle to top). This results in a longer lifetime of the weak binding state, thereby a higher probability of transition to the strong binding state in the backward (downward) movement of actin filament and also in a directionally preferential release of myosin head in the forward movement of actin filament, causing a biased Brownian motion. Blue arrows indicate the directions of actin filament movement and myosin rotation, and dashed black arrows indicate the probabilities of transitions between the states by their sizes

Fig. 4.8 (continued)



**Acknowledgement** This work was supported by JSPS KAKENHI Grant number 25711010 to T.F and 25000013 to K.N.

## References

- Banerjee C, Hu Z, Huang Z, Warrington JA, Taylor DW, Trybus KM, Lowey S, Taylor KA (2017) The structure of the actin-smooth muscle myosin motor domain complex in the rigor state. *J Struct Biol* 200:325–333
- Bauer CB, Holden HM, Thoden JB, Smith R, Rayment I (2000) X-ray Structures of the Apo and MgATP-bound States of Dictyostelium discoideum Myosin Motor Domain. *J Biol Chem* 275:38494–38499
- Behrmann E, Müller M, Penczek PA, Manherz HG, Manstein D, Raunser S (2012) Structure of the rigor actin-tropomyosin-myosin complex. *Cell* 150:327–338
- Cao E, Liao M, Cheng Y, Julius D (2013) TRPV1 structures in distinct conformations reveal activation mechanisms. *Nature* 504:113–118
- Carlier MF, Pantaloni D (2007) Control of actin assembly dynamics in cell motility. *J Biol Chem* 282:23005–23009
- Coureux PD, Wells AL, Ménétrey J, Yengo CM, Morris CA, Sweeney HL, Houdusse A (2003) A structural state of the myosin V motor without bound nucleotide. *Nature* 425:419–423
- Dominguez R, Freyzon Y, Trybus KM, Cohen C (1998) Crystal structure of a vertebrate smooth muscle myosin motor domain and its complex with the essential light chain: visualization of the pre-power stroke state. *Cell* 94:559–571
- Egelman EH (2000) A robust algorithm for the reconstruction of helical filaments using single-particle methods. *Ultramicroscopy* 85:453–463
- Fujii T, Namba K (2017) Structure of actomyosin rigour complex at 5.2 Å resolution and insights into the ATPase cycle mechanism. *Nature Commun* 8:13969 (11pp)
- Fujii T, Kato T, Namba K (2009) Specific arrangement of  $\alpha$ -helical coiled coils in the core domain of the bacterial flagellar hook for the universal joint function. *Structure* 17:1485–1493
- Fujii T, Iwane AH, Yanagida T, Namba K (2010) Direct visualization of secondary structures of F-actin by electron cryomicroscopy. *Nature* 467:724–728
- Fujiwara I, Vavylonis D, Pollard TD (2007) Polymerization kinetics of ADP- and ADP-Pi-actin determined by fluorescence microscopy. *Proc Natl Acad Sci U S A* 104:8827–8832
- Fujiyoshi Y, Mizusaki T, Morikawa K, Yamagishi H, Aoki Y, Kihara H, Harada Y (1991) Development of a superfluid helium stage for high-resolution electron microscopy. *Ultramicroscopy* 38:241–251
- Galkin VE, Orlova A, Cherepanova O, Lebart MC, Egelman EH (2008) High-resolution cryo-EM structure of the F-actin-fimbrin/plastin ABD2 complex. *Proc Natl Acad Sci U S A* 105:1494–1498
- Gayathri P, Fujii T, Møller-Jensen J, van den Ent F, Namba K, Löwe J (2012) A bipolar spindle of antiparallel ParM filaments drives bacterial plasmid segregation. *Science* 338:1334–1337
- Geeves MA, Goody RS, Gutfrund H (1984) Kinetics of acto-S1 interaction as a guide to a model of the crossbridge cycle. *J Muscle Res Cell Motil* 5:351–356
- Holmes KC, Angert I, Kull FJ, Jahn W, Schröder RR (2003) Electron cryo-microscopy shows how strong binding of myosin to actin releases nucleotide. *Nature* 425:423–427
- Holmes KC, Schroder RR, Sweeney HL, Houdusse A (2004) The structure of the rigor complex and its implications for the power stroke. *Philos Trans R Soc B* 359:1819–1828
- Houdusse A, Szent-Gyögyi AG, Cohen C (2000) Three conformatinoal states of scallop myosin S1. *Proc Natl Acad Sci U S A* 97:11238–11243
- Huxley HE (1969) The mechanism of muscular contraction. *Science* 164:1356–1365
- Iwaki M, Iwane AH, Shimokawa T, Cooke R, Yanagida T (2009) Brownian search-and-catch mechanism for myosin-VI steps. *Nature Chem Biol* 5:403–405

- Kabsch W, Mannherz HG, Suck D, Pai EF, Holmes KC (1990) Atomic model of the actin:DNase I complex. *Nature* 347:37–44
- Kimanius D, Forsberg BO, Scheres SH, Lindahl E (2016) Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *elife* 15:e18722
- Kimura Y, Vassilyev DG, Miyazawa A, Kidera A, Matsushima M, Mitsuoka K, Murata K, Hirai T, Fujiyoshi Y (1997) Surface of bacteriorhodopsin revealed by high-resolution electron crystallography. *Nature* 389:206–211
- Kitamura K, Tokunaga M, Iwane AH, Yanagida T (1999) A single myosin head moves along an actin filament with regular steps of 5.3 nanometres. *Nature* 397:129–134
- Li X, Mooney P, Zheng S, Booth CR, Braunfeld MB, Gubbens S, Agard DA, Cheng Y (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* 10:584–590
- Liao M, Cao E, Julius D, Cheng Y (2013) Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* 504:107–112
- Llinas P, Isabet T, Song L, Ropars V, Zong B, Benisty H, Sirigu S, Morris C, Kikuti C, Safer D, Sweeney HL, Houdusse A (2015) How actin initiates the motor activity of myosin. *Develop Cell* 33:401–412
- Lynn RW, Taylor EW (1971) Mechanism of adenosine triphosphate hydrolysis by actomyosin. *Biochemist* 10:4617–4624
- Ménétry J, Bahloul A, Wells AL, Yengo CM, Morris CA, Sweeney HL, Houdusse A (2005) The structure of the myosin VI motor reveals the mechanism of directionality reversal. *Nature* 435:779–785
- Ménétry J, Llinas P, Cicolari J, Squires G, Liu X, Li A, Sweeney HL, Houdusse A (2008) The post-rigor structure of the myosin VI and implications for the recovery stroke. *EMBO J* 27:244–252
- Mentes A, Huehn A, Liu X, Zwolak A, Dominguez R, Shuman H, Ostap EM, Sindelar CV (2018) High-resolution cryo-EM structures of actin-bound myosin states reveal the mechanism of myosin force sensing. *Proc Natl Acad Sci U S A* 115:1292–1297
- Mimori Y, Yamashita I, Murata K, Fujiyoshi Y, Yonekura K, Toyoshima C, Namba K (1995) The structure of the R-type straight flagellar filament of Salmonella at 9 Å resolution by electron cryomicroscopy. *J Mol Biol* 249:69–87
- Mitsuoka K, Hirai T, Murata K, Miyazawa A, Kidera A, Kimura Y, Fujiyoshi Y (1999) The structure of bacteriorhodopsin at 3.0 Å resolution based on electron crystallography: implication of the charge distribution. *J Mol Biol* 286:861–882
- Miyazawa A, Fujiyoshi Y, Unwin N (2003) Structure and gating mechanism of the acetylcholine receptor pore. *Nature* 423:949–955
- Murata K, Mitsuoka K, Hirai T, Walz T, Agre P, Heymann JB, Engel A, Fujiyoshi Y (2000) Structural determinants of water permeation through aquaporin-1. *Nature* 407:599–605
- Nagy B, Jencks WP (1965) Depolymerization of F-actin by concentrated solutions of salts and denaturing agents. *J Am Chem Soc* 87:2480–2488
- Namba K, Stubbs G (1985) Solving the phase problem in fiber diffraction. Application to tobacco mosaic virus at 3.6 Å resolution. *Acta Crystallogr A* 41:252–262
- Namba K, Stubbs G. (1986) Structure of tobacco mosaic virus at 3.6 Å resolution: implications for assembly. *Science* 231:1401–1406
- Oda T, Iwasa M, Aihara T, Maeda Y, Narita A (2009) The nature of the globular- to fibrous-actin transition. *Nature* 457:441–445
- Otterbein LR, Graceffa P, Dominguez R (2001) The crystal structure of uncomplexed actin in the ADP state. *Science* 293:708–711
- Pollard TD, Borisy GG (2003) Cellular motility driven by assembly and disassembly of actin filaments. *Cell* 112:453–465
- Rayment, I., Rypniewski, W. R., Schmidt-Bäse, K., Smith, R., Tomchick, D. R., Benning, M. M., Winkelmann D. A., Wesenberg, G. & Holden HM. (1993) Three-dimensional structure of myosin subfragment-1: a molecular motor. *Science* 261, 50–58
- Reubold TF, Eschenburg S, Becker A, Kull FJ, Manstein DJ (2003) A structural model for actin-induced nucleotide release in myosin. *Nature Struct Biol* 10:826–830

- Reubold, T. F., Eschenburg, S., Becker, Loonard, M, Schmid, S. L., Vallee, R. B., Kull, F. J. & Manstein, D. J. (2005) Crystal structure of the GTPase domain of rat dynamin 1. *Proc Natl Acad Sci U S A* 102, 13093–13098
- Sachse C, Chen JZ, Coureux P, Stroupe ME, Fandrich M, Grigorieff N (2007) High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus. *J Mol Biol* 371:812–835
- Samatey FA, Imada K, Nagashima S, Kumasaka T, Yamamoto M, Vonderviszt F, Namba K (2001) Structure of the bacterial flagellar protofilament and implication for a switch for supercoiling. *Nature* 410:331–337
- Samatey FA, Matsunami H, Imada K, Nagashima S, Shaikh TR, Thomas DR, Chen JZ, Derosier DJ, Namba K (2004) Structure of the bacterial flagellar hook and implication for the molecular universal joint mechanism. *Nature* 431:1062–1068
- Scheres SH (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* 180:519–530
- Schroder GF, Brunger AT, Levitt M (2007) Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Struct* 15:1630–1641
- Sweeney HL, Houdusse A (2004) The motor mechanism of myosin V: insights for muscle contraction. *Philos Trans R Soc B* 359:1829–1841
- Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A (2008) Protein structure fitting and refinement guided by cryo-EM density. *Struct*. 16:295–307
- von der Ecken J, Heissler SM, Pathan-Chhatbar S, Manstein DJ, Raunser S (2016) Cryo-EM structure of a human cytoplasmic actomyosin complex at near-atomic resolution. *Nature* 534:724–728
- Wulf SF, Roparsb V, Fujita-Beckera S, Ostera M, Hofhaus G, Trabucoc LG, Pylypenkob O, Sweeney HL, Houdusseb AM, Schröder R (2016) Force-producing ADP state of myosin bound to actin. *Proc Natl Acad Sci U S A* 113:E1844–E1852
- Yanagida T, Arata T, Oosawa F (1985) Sliding distance of actin filament induced by a myosin crossbridge during one ATP hydrolysis cycle. *Nature* 316:366–369
- Yang Y, Gourinath S, Kovács M, Mitray L, Reutzel R, Himmel DM, O’Neill-Hennessey E, Reshetnikova L, Szent-Györgyi AG, Brown JH, Cohen C (2007) Rigor-like structures from muscle myosins reveal key mechanical elements in the transduction pathways of this allosteric motor. *Structure* 15:553–564
- Yonekura K, Maki-Yonekura S, Namba K (2003) Complete atomic model of the bacterial flagellar filament by electron cryomicroscopy. *Nature* 424:643–650
- Yount RG, Lawson D, Rayment I (1995) Is myosin a “Back Door” Enzyme? *Biophys J* 68:44s–49s

# Chapter 5

## Current Solution NMR Techniques for Structure-Function Studies of Proteins and RNA Molecules



John L. Markley

**Abstract** We briefly review current technology for structure-function investigations of biological macromolecules in solution by nuclear magnetic resonance spectroscopy, which enable hybrid methods. An advantage of NMR is that biomolecules can be studied at atomic resolution under near physiological conditions where they are dynamically active. We outline stable isotope labeling strategies, NMR data collection methodology, and procedures for data analysis leading to structure-function information. We discuss issues related to NMR software and data deposition.

**Keywords** Dynamics · Stable isotope labeling · NMR data collection strategies · NMR observables · Spectral assignment · Structural restraints · NMR software packages · Validation of NMR results · Functional studies · Data deposition

### 5.1 Introduction

This review focuses on recent developments in solution NMR. The growing field of solid-state NMR, which has particular applicability to studies of membrane proteins, fibrous proteins, and viruses, is not covered here: for reviews see: (Linser 2017; Molugu et al. 2017; Zhao et al. 2017). The advantages of solution NMR spectroscopy for investigations of biological macromolecules are that it enables atomic-level studies of their structure and dynamics in solution under a variety of conditions (pH, temperature, pressure, and added ligands). NMR signals can be resolved from residues in both ordered and disordered regions, and their observable parameters (chemical shift, spin-spin couplings, dipolar couplings, relaxation and cross-relaxation rates, etc.) provide structural and functional information. Solution NMR spectroscopy gives a very different, but complementary, view of proteins and nucleic acids than the static picture depicted by X-ray crystallography. Crystal

---

J. L. Markley (✉)

Biochemistry Department, University of Wisconsin-Madison, Madison, WI, USA

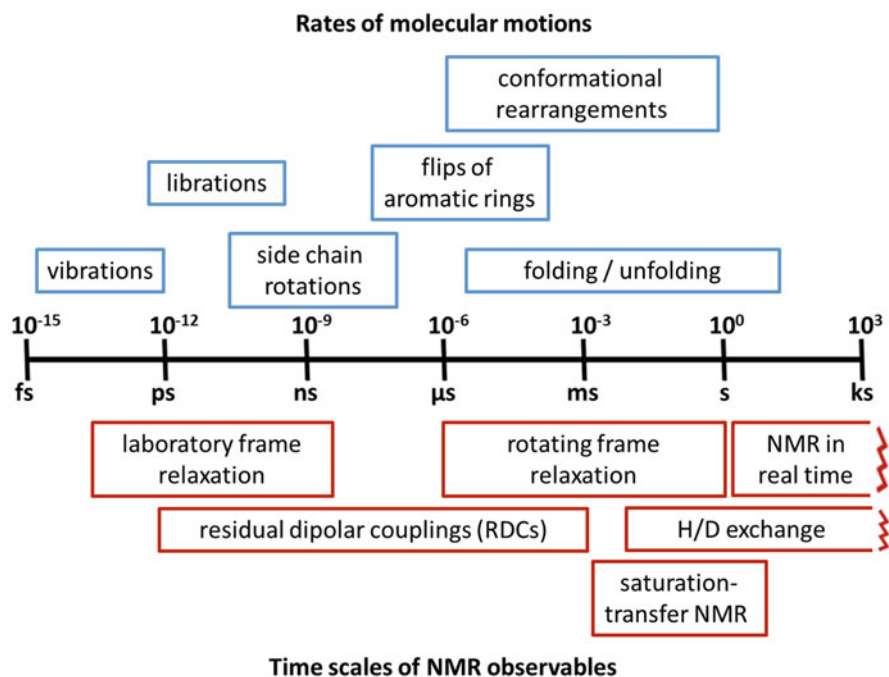
e-mail: [jmarkley@wisc.edu](mailto:jmarkley@wisc.edu)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,

Advances in Experimental Medicine and Biology 1105,

[https://doi.org/10.1007/978-981-13-2200-6\\_5](https://doi.org/10.1007/978-981-13-2200-6_5)



**Fig. 5.1** Time scales covered by (blue boxes) different NMR approaches in (red boxes) comparison with the rates of dynamic processes in proteins

packing forces tend to stabilize a single conformation, and the collection of X-ray data at low temperatures damps out motions leading to higher structural resolution. NMR experiments detect structural fluctuations over a time scale from  $10^{-15}$  s to minutes (Fig. 5.1) (Palmer et al. 2001). As a consequence, we know that proteins and nucleic acids are dynamic and undergo structural transitions. Solvent-exposed side chains are mobile, and the interiors of proteins undergo breathing motions that enable rotations of the aromatic side chains of Tyr and Phe. Some parts of a molecule or complex may be dynamically disordered. NMR is uniquely capable of detecting conformational states with low populations and/or following transitions between states. These minor states may be functionally important in catalysis or other functional properties. NMR can detect differences in chemical properties of states, such as a different protonation or redox state. As many as 40% of proteins in the human proteome are predicted to be intrinsically disordered, and many of these are known to become ordered when they interact with binding partners. NMR spectroscopy offers the most comprehensive way of investigating the properties of disordered states and how regions become ordered as a consequence of intermolecular interactions.

Solution NMR does have definite limitations: the size of molecules and complexes limits the resolution of solution NMR signals as do dynamic processes that



occur on an unfavorable time scale. In addition, as detailed below, macromolecules need to be labeled with stable isotopes and prepared in sufficient quantity (generally >1 mg).

The usual workflow in NMR-based structure determination involves the preparation of suitably labeled samples, the collection of several types of NMR data, analysis of these data to assign NMR observables to particular groups in the covalent structure of the molecule and to derive secondary, tertiary, and possibly quaternary structure. Finally, structures are validated for consistency with the experimental data, and the structures and associated data are deposited in the Protein Data Bank (PDB) (Berman et al. 2009) and BioMagResBank (BMRB) (Ulrich et al. 2008). Structures determined by NMR represent a statistical ensemble of the dynamic states in solution.

Many biomolecular NMR investigations do not have the generation of 3D coordinates as their goal. They may go beyond structure to investigate thermodynamic or kinetic properties of the molecule or complex, rates of conformational transitions, or effects of ligand binding. Experimental data from such studies are archived at BMRB.

## 5.2 Sample Preparation and Isotope Labeling

Genes coding for proteins are cloned or synthesized. *Escherichia coli* is usually the first choice for protein production because of the large number of available cloning vectors and specialized strains including auxotrophs (Hewitt and McDonnell 2004; Markley et al. 2009). *E. coli* can be grown on inexpensive labeled precursors ( $^{13}\text{C}$ -labeled glucose or  $^{15}\text{N}$ -labeled ammonia). In addition, methods with *E. coli* support perdeuteration and residue-selective labeling (Matthews 2004; Rajesh et al. 2003). For proteins that cannot be produced from *E. coli*, *Pichia pastoris* (Pickford and O'Leary 2004), baculovirus grown on insect cells (Kost et al. 2005), and cell-free methods (Makino et al. 2014; Takeda and Kainosho 2012; Kigawa et al. 2007) offer alternatives. The latter two methods require labeled amino acids rather than inexpensive precursors.

Many types of labeled precursors are commercially available. For proteins up to 25 kDa, it is common to label uniformly with both  $^{13}\text{C}$  and  $^{15}\text{N}$ . This can be achieved by growing *E. coli* on [ $^{13}\text{C}_6$ ]-glucose as the sole carbon source and  $^{15}\text{NH}_4\text{Cl}$  as the sole nitrogen source. With larger proteins (65 kDa or higher), deuterated  $^{13}\text{C}$  glucose [ $^{13}\text{C}_6$  1,2,3,4,5,6-d7] is used as the sole carbon source and ammonium- $^{15}\text{N}$ ,d4 chloride as the nitrogen source for *E. coli* cells grown in  $\text{D}_2\text{O}$ . The cells need to become adapted to growth in heavy water. An alternative approach is to grow the cells on an algal hydrolysate with the desired isotopic composition. Following perdeuteration, it is customary to back-exchange the protein in  $\text{H}_2\text{O}$  to replace  $^2\text{H}$  on labile backbone and sidechain amides with  $^1\text{H}$ .

Feeding *E. coli* a mixture of glycerol-1-3- $^{13}\text{C}$  and glycerol-2- $^{13}\text{C}$  leads to rough labeling of every other carbon in an amino acid with  $^{13}\text{C}$ . This labeling pattern

along with direct  $^{13}\text{C}$  detection has advantages with larger proteins and protein complexes (Takeuchi et al. 2008).

Kainosho and co-workers have designed amino acids with optimal patterns of  $^2\text{H}$ ,  $^{13}\text{C}$ , and  $^{15}\text{N}$  for protein NMR studies (Kainosho et al. 2006). These stereo-array isotope labeled (SAIL) amino acids yield sharper and simpler spectra through reduction in the number of  $^1\text{H}$  spins, spin-spin couplings, and spin diffusion pathways.

Methyl-labeling (Tugarinov and Kay 2005) or incorporation of fluorine-labeled amino acids (Sharaf and Gronenborn 2015) offer probes for NMR investigations of proteins and complexes of 100 kDa or larger. However, they do not provide an easy pathway for structure determination.

NMR structures of integral membrane proteins are challenging because they need to be stabilized in a membrane-like environment (Rajesh et al. 2016). Detergent micelles provide a solubilization mechanism a minimum increase in tumbling time and thus line broadening, but may not support a native active conformation. Detergent bicelles can be better at protein stabilization but lead to decreased tumbling rates. A promising approach is to incorporate integral membrane proteins into nanodiscs, discrete phospholipid mimetics modeled on high-density lipoprotein particles (Denisov et al. 2004). Wagner and co-workers have developed covalently circularized nanodiscs whose size can be tailored to a specific integral membrane protein (Nasr et al. 2017). Moreover, because they are covalent circles, they enable NMR data collection at higher temperatures where NMR signals are sharper.

### 5.3 NMR Data Collection

In order to resolve peak overlaps, macromolecular NMR data are collected as  $n$ -dimensional spectra. Early 2D  $^1\text{H}$ - $^{13}\text{C}$  (Chan and Markley 1982) and  $^1\text{H}$ - $^{15}\text{N}$  (Ortiz-Polo et al. 1986) studies of proteins utilized  $^{13}\text{C}$ - and  $^{15}\text{N}$  detection, respectively. With advances in instrumentation, indirect  $^1\text{H}$  detection became the norm for multinuclear NMR studies, because of the higher sensitivity of  $^1\text{H}$  sensitivity. More recently, the direct detection  $^{13}\text{C}$  and  $^{15}\text{N}$  has been re-investigated and shown to be advantageous for studies of larger proteins and nucleic acids. A suite of “protonless” direct  $^{13}\text{C}$ -detected experiments has been devised for complete protein assignments (Bermel et al. 2006). And direct  $^{15}\text{N}$ -detected experiments have been developed for proteins (Takeuchi et al. 2010; Gal et al. 2011). A  $^{15}\text{N}$ -detected TROSY-HSQC experiment shows promise as a way to study larger proteins without the need for perdeuteration and back-exchange (Takeuchi et al. 2016; Takeuchi et al. 2015).  $^{15}\text{N}$ -detected  $^1\text{H}$ - $^{15}\text{N}$  correlation experiments of larger RNA molecules have shown recent promise (Schnieders et al. 2017).

Advances in NMR instrumentation and data collection have led to increased spectral sensitivity. Cryogenic probes achieve increased sensitivity by cooling transmitter/receiver coils to liquid nitrogen or lower temperatures to reduce thermal

noise (Kovacs et al. 2005). Higher field magnets increase sensitivity by increasing equilibrium spin polarization. Data collection by Transverse Relaxation Optimized Spectroscopy (TROSY) methods (Pervushin et al. 1998; Salzmänn et al. 1998) leads to increased sensitivity, particularly for larger macromolecules.

Despite these advances, NMR data collection requires time averaging, with the amount of time required increasing geometrically with spectral dimensionality. Recently, two approaches to higher sensitivity in less time have been developed that take advantage of the sparsity of multidimensional NMR data: reduced dimensionality (Eghbalnia and Markley 2017) and non-uniform sampling (Hyberts et al. 2011). The general idea behind reduced dimensionality is illustrated by the collection of a 3D ( $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ ) spectrum as a series of tilted 2D planes, where one dimension is  $^1\text{H}$  and the other is a mixture of  $^{13}\text{C}$  and  $^{15}\text{N}$  frequencies which define the tilt angle (Kupce and Freeman 2003). Because the peaks in 3D space are sparse, all peaks can be sampled by a fewer number of tilted planes than sampled by a stack of non-tilted planes. The reduced-dimensionality paradigm has been implemented in different ways. For example, Automated Projection Spectroscopy (APSY) (Hiller et al. 2008; Krahenbuhl et al. 2014) utilizes a pre-specified projection regime for nD spectroscopy, whereas High-resolution Iterative Frequency Identification for NMR (HIFI-NMR) uses a Bayesian updating procedure that tightly integrates data acquisition with data processing and analysis to yield spectral assignment in real-time (Eghbalnia et al. 2005a; Lee et al. 2013a, b). Non-uniform sampling and spectral reconstruction are now standard on commercial NMR spectrometers, and are widely used to accelerate data collection and reduce the data size of nD spectra. Non-uniform sampling strategies, along with tools for spectral processing and signal reconstruction, are still evolving and becoming more powerful; see (Billeter 2017) and references therein.

## 5.4 NMR Observables Used in Structure Determination

Chemical shifts are the primary observables protein NMR spectroscopy. Once assigned, they can be used in determining secondary structure (Eghbalnia et al. 2005b; Shen and Bax 2013), likely flexible or disordered regions (Berjanskii and Wishart 2008), side chain mobility (Berjanskii and Wishart 2013), and possible  $^{13}\text{C}$  chemical shift referencing errors (Wang et al. 2005). Chemical shifts can be used in homology modeling (Shen and Bax 2015). In addition, assigned chemical shifts can be used in conjunction with Rosetta software (CS-Rosetta) to determine three-dimensional structures of small proteins (Shen et al. 2008).

The nuclear Overhauser effect (NOE), which is used to obtain structural restraints, is the consequence of  $^1\text{H}$ - $^1\text{H}$  cross relaxation. Normally the effect can be observed for pairs of protons that are within 5 Å of one another (Wüthrich 1986). The mixing time (time during which cross-relaxation is allowed to build up) must be kept short to minimize spin diffusion effects that degrade the accuracy of distance measurements. The data can be collected as a 2D NOESY experiment, in which the

1D  $^1\text{H}$  spectrum lies along the diagonal, and cross peaks occur at the intersection of the chemical shifts of protons that are close to one another. With proteins labeled with  $^{13}\text{C}$  and  $^{15}\text{N}$ , 3D NOESY-HSQC experiments allow the editing of the NOE peaks by the chemical shifts of the  $^{13}\text{C}$  or  $^{15}\text{N}$  nuclei (X) to which the protons are attached. These 3D spectra have two  $^1\text{H}$  dimension and one X dimension. The 4D C, N-edited NOESY experiment leads to separation of NOE cross peaks by the chemical shifts of both  $^{13}\text{C}$  and  $^{15}\text{N}$ . The 4D spectrum has two  $^1\text{H}$  dimensions, a  $^{13}\text{C}$  dimension, and a  $^{15}\text{N}$  dimension. Sparse sampling is generally carried out to reduce data collection to a reasonable time (Stanek et al. 2012).

Residual dipolar couplings (RDCs) are another important observable NMR parameter. RDCs are determined from the difference in couplings observed in a partially orienting (J + D) and non-orienting (isotropic) environment (J). A variety of orientation media have been described including lipid bicelles (Metz et al. 1995), liquid crystalline bicelles (Tjandra and Bax 1997a, b), rod-shaped virus such as filamentous bacteriophage (Hansen et al. 1998; Clore et al. 1998), and DNA nanotubes, which are compatible with detergents used to solubilize membrane proteins (Douglas et al. 2007). Even small molecules, such as natural products, can be oriented for RDC measurements (Gayathri et al. 2010). Recent approaches for measuring RDCs include intensity modulation (McFeeters et al. 2005), direct  $^{13}\text{C}$  detection (Balayssac et al. 2006), and ARTSY (amide RDCs by TROSY spectroscopy) (Fitzkee and Bax 2010). Software packages are available for analyzing RDC data (Valafar and Prestegard 2004; Lorieau 2017; Schwieters et al. 2017).

Spin-spin couplings can be used as dihedral constraints, but their use has been largely supplanted by chemical shift analysis. J-couplings that traverse hydrogen bonds can be useful for detecting and quantifying hydrogen bonds (Cordier and Grzesiek 1999; Cornilescu et al. 1999).

Larger proteins, membrane proteins, and partially disordered proteins are challenging as structural targets. The sparse NMR data obtainable for such systems can be supplemented by the introduction of paramagnetic labels and by collecting a data from a variety of NMR experiments. These approaches and associated computational algorithms for determining structures from pseudocontact shifts have been reviewed recently (Pilla et al. 2017a, b). A recent study used paramagnetic-induced  $^{19}\text{F}$  relaxation enhancement (PRE) in conjunction with  $^{19}\text{F}$  labeling to obtain structural constraints in a large protein (Matei and Gronenborn 2015).

## 5.5 Software for Data Analysis and Assignment

A large variety of software tools have been developed for biomolecular NMR applications, and many of these have evolved through a progression of releases. The NMRbox project (Maciejewski et al. 2017) has the goal of archiving these software packages and of making them available for use from a virtual machine platform

to enable the replication of experiments. A further goal is to enable the pipelining of data from one software package to another while capturing relevant information about the workflow. This ambitious project promises important benefits to the field.

NMR data are collected as a function of time (time domain) and need to be transformed to the frequency domain to yield NMR spectra. Spectrometer manufacturers provide software for NMR data processing. An alternative is NMRpipe (Delaglio et al. 1995), a freely available software package with many processing features including the reconstruction of spectra from sparsely sampled NMR data.

Popular software packages for viewing, annotating, and analyzing spectra are NMRView (Johnson 2004) and Sparky (Kneller and Kuntz 1993). The National Magnetic Resonance Facility at Madison (NMRFAM) which has incorporated Sparky into its software packages, took over the development of this package and released an enhanced version named NMRFAM-SPARKY (Lee et al. 2015). These programs have built-in peak picking capability, but external peak picking software packages are available (Koradi et al. 1998; Shin et al. 2008).

Chemical shift assignments of small proteins are derived from combinations of two-, three-, and possibly higher-dimensional NMR data sets. Peak assignments can be carried out manually with assistance from spectral visualization packages, or from assignment software. The Integrative NMR package (Lee et al. 2016), combines this process by displaying assignments predicted from the probabilistic PINE package (Bahrami et al. 2009) on spectra. The user can accept and refine the position of the proposed assignment or negate the assignment with mouse clicks.

For proteins with known 3D structure, for example from X-ray crystallography, software packages have been developed to assist the assignment of methyl signals from NOE data: FLAMEnGO (Chao et al. 2011); MAGMA (Pritisanic et al. 2017).

## 5.6 Structure Determination

Structure derived from NMR data are always underdetermined; thus they are simply models that are consistent with the available experimental data (Mackay et al. 2017). Widely used structure determination programs include CYANA (Güntert 2004) and Xplor-NIH (Schwieters et al. 2017). Recent software packages, such as FLYA (Schmidt and Güntert 2012) and Integrative NMR (Lee et al. 2016) combine peak assignments with protein structure determination. The latter package, which is available as a virtual machine, incorporates NMRFAM-SPARKY (Lee et al. 2015) for spectral visualization and annotation, with APES for peak picking (Shin et al. 2008), PINE for automated assignment (Bahrami et al. 2009), ARECA (Dashti et al. 2016) for validation of peak assignments, TALOS-N for shift based torsion angle restraints (Shen and Bax 2013), CS-Rosetta (Shen et al. 2008), for structure determination from chemical shifts, AUDANA (Lee et al. 2016) and PONDEROSA-C/S (Lee et al. 2014) for automated structure determination from NOE spectra, and data visualization by NDP-PLOT and an enhanced mode of the PyMOL software package (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger,

LLC.). A new software package, PINE-SPARKY.2 (Lee and Markley 2018, which comes as a plug-in to NMRFAM-SPARKY, further integrates several of these tasks and provides, in addition, easy-to-use visual analysis tools based on probability theory (Lee and Markley 2018).

## 5.7 Hybrid Approaches with NMR

NMR structures have been used to solve the phase problem with X-ray diffraction maps. One study provided evidence that Rosetta refinement of NMR structures aided this process (Ramelot et al. 2009).

Assigned backbone NMR chemical shifts constitute a minimal data set that can be combined with Rosetta to determine a 3D structure (Mao et al. 2014; Rosato et al. 2012; Lange et al. 2012). Protein structures can be determined by co-evolutionary restraints alone (Ovchinnikov et al. 2017); however, the availability of sparse NMR data yields a hybrid method for improving such structures (Tang et al. 2015). This hybrid approach is described in detail in a separate Chap. 10 in this volume.

Small angle scattering (SAS) and NMR spectroscopy are useful combinations as reviewed recently (Mertens and Svergun 2017). NMR RDC measurements can be combined with SAXS data to characterize conformational ensembles (Venditti et al. 2016). One approach is to build NMR structures into SAXS envelopes as shown recently with gammaD-crystallin (Whitley et al. 2017) and NFU1 (Cai et al. 2016). SAS restraints have proven useful in refining NMR structures of RNA molecules (Cornilescu et al. 2016; Cantero-Camacho et al. 2017). The combination of Cryo-EM and NMR data has been reviewed recently (Cuniasse et al. 2017). One example is the refinement of the Cryo-EM structure of HIV-1 capsid protein with NMR data and MD simulations (Perilla et al. 2017). The integration of data from a variety of techniques, including NMR, is challenging. A promising approach involves Bayesian inferential structure determination (Habeck 2017).

## 5.8 Validation of NMR Data

The Worldwide PDB sponsored an NMR Validation Task Force charged with recommending methods for validating NMR data deposited in the PDB archive. The initial report of this Task Force (Montelione et al. 2013) identified three phases for validation: (Phase 1) validation by methods that are available by existing software that has been well documented, (Phase 2) validation by available methods that require further review, and (Phase 3) validation by methods that require development. The panel recommended immediate implementation of Phase 1 methods as part of the PDB validation report. These Reports should include four components: (1) a report validating the completeness and global referencing of chemical shift data, independent of 3D structure; (2) analysis of “well-defined”

versus “ill-defined” regions; (3) a knowledgebased model validation report; and (4) a restraint-based model-versus-data validation report, comparing each member of the ensemble of NMR models to the available NMR restraints. To date items 1 and 3 have been implemented as part of the OneDep system. Item 2 should be implemented soon, and software for implementing item 4 is under development at BMRB.

## 5.9 Use of NMR for Dynamics and Functional Studies

Although 3D structures of proteins can be determined by NMR spectroscopy, a major strength of NMR is its ability to investigate a variety of functional properties in solution (Barrett et al. 2013). NMR is ideal for detecting protein dynamics (Vallurupalli et al. 2017), functionally dynamic states (Kay 2016; Rosenzweig and Kay 2016), and excited states that have a low population (Sekhar and Kay 2013). NMR can be used to determine complex protein energy landscapes (Khirich and Loria 2015) and functional properties such as pKa values of individual sites in proteins, allostery in enzyme catalysis (Lisi and Loria 2017), protein-ligand interactions, and protein-protein interactions (Lipchock and Loria 2009). A recent review discusses these applications with regard to membrane proteins (Liang and Tamm 2016). The monitoring of hydrogen exchange by fast pressure jump NMR is opening new approaches to studying conformational changes in proteins including protein folding (Alderson et al. 2017).

## 5.10 Data Handling and Deposition

In 1996, BMRB converted its archive from a restrictive format akin to the old PDB format, to NMR-STAR (Ulrich et al. 1996). STAR is related to the CIF format adopted earlier by small-molecule crystallographers (Hall et al. 1991). STAR (Hall 1991; Hall and Cook 1995; Hall and Spadaccini 1994) differs from CIF by supporting a “save frame” architecture that enables a tabular format. This feature enables NMR-STAR to capture information pertaining to unique entities (molecules, samples, experimental procedures, sets of results, etc.) and to link these entities in a relatively efficient manner. This greatly reduces the number of redundant data tags needed within a single file. Because of its relation to the flat format CIF, NMR-STAR is easily converted to the mmCIF (PDBx) format used by the Protein Data Bank (Fitzgerald et al. 2005).

NMR-STAR is defined by a dictionary that evolves as new experimental methods are developed. BMRB has been working with the biomolecular NMR community to expand NMR-STAR to handle a wide range of NMR experiments and associated hybrid methods. Recent developments include ways of dealing with sparse

sampling and reduced dimensionality NMR data as well as data from NMR-based metabolomics studies.

In 2015, a group of NMR software developers, in cooperation with the Worldwide Protein Data Bank, proposed an NMR Exchange Format (NEF) as a streamlined representation of NMR data in STAR format (Gutmanas et al. 2015). The idea was that by adopting NEF, different software packages could more readily exchange data. BMRB in its latest version of the NMR-STAR dictionary adopted some of the features of NEF and produced NMR-STAR tags for each of the NEF STAR tags. This enabled BMRB to develop software to convert NEF to NMR-STAR. The wwPDB is accepting the deposition of NMR restraint data in NEF as well as in NMR-STAR format. The OneDep system (Young et al. 2017) will convert NEF to NMR-STAR prior to archiving the data.

## References

- Alderson TR, Charlier C, Torchia DA, Anfinrud P, Bax A (2017) Monitoring hydrogen exchange during protein folding by fast pressure jump NMR spectroscopy. *J Am Chem Soc* 139(32):11036–11039. <https://doi.org/10.1021/jacs.7b06676>
- Bahrami A, Assadi AH, Markley JL, Eghbalnia HR (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Comput Biol* 5(3):e1000307. <https://doi.org/10.1371/journal.pcbi.1000307>
- Balayssac S, Bertini I, Luchinat C, Parigi G, Piccioli M (2006)  $^{13}\text{C}$  direct detected NMR increases the detectability of residual dipolar couplings. *J Am Chem Soc* 128(47):15042–15043
- Barrett PJ, Chen J, Cho MK, Kim JH, Lu Z, Mathew S, Peng D, Song Y, Van Horn WD, Zhuang T, Sonnichsen FD, Sanders CR (2013) The quiet renaissance of protein nuclear magnetic resonance. *Biochemistry* 52(8):1303–1320. <https://doi.org/10.1021/bi4000436>
- Berjanskii MV, Wishart DS (2008) Application of the random coil index to studying protein flexibility. *J Biomol NMR* 40(1):31–48
- Berjanskii MV, Wishart DS (2013) A simple method to measure protein side-chain mobility using NMR chemical shifts. *J Am Chem Soc* 135(39):14536–14539. <https://doi.org/10.1021/ja407509z>
- Berman HM, Henrick K, Nakamura H, Markley JL (2009) The worldwide protein data bank. In: Gu J, Bourne P (eds) *Structural bioinformatics*, 2nd edn. Wiley, Chichester, pp 293–303
- Bermel W, Bertini I, Felli IC, Kummerle R, Pierattelli R (2006) Novel  $^{13}\text{C}$  direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins. *J Magn Reson* 178(1):56–64
- Billeter M (2017) Non-uniform sampling in biomolecular NMR. *J Biomol NMR* 68(2):65–66. <https://doi.org/10.1007/s10858-017-0116-7>
- Cai K, Liu G, Frederick RO, Xiao R, Montelione GT, Markley JL (2016) Structural/functional properties of human NFU1, an intermediate [4Fe-4S] carrier in human mitochondrial iron-sulfur cluster biogenesis. *Structure* 24(12):2080–2091. <https://doi.org/10.1016/j.str.2016.08.020>
- Cantero-Camacho A, Fan L, Wang YX, Gallego J (2017) Three-dimensional structure of the 3'X-tail of hepatitis C virus RNA in monomeric and dimeric states. *RNA* 23:1465–1476. <https://doi.org/10.1261/rna.060632.117>
- Chan TM, Markley JL (1982) Heteronuclear ( $^1\text{H}$ ,  $^{13}\text{C}$ ) two-dimensional chemical shift correlation NMR spectroscopy of a protein. Ferredoxin from *Anabaena variabilis*. *J Am Chem Soc* 104:4010–4011



- Chao FA, Shi L, Masterson LR, Veglia G (2011) FLAMEnGO: a fuzzy logic approach for methyl group assignment using NOESY and paramagnetic relaxation enhancement data. *J Magn Reson* 214:103–110. <https://doi.org/10.1016/j.jmr.2011.10.008>
- Clore GM, Gronenborn AM, Tjandra N (1998) Direct structure refinement against residual dipolar couplings in the presence of Rhombicity of unknown magnitude. *J Magn Reson* 131(1):159–162
- Cordier F, Grzesiek S (1999) Direct observation of hydrogen bonds in proteins by Interresidue  $^3\text{H}_{\text{NC}}$  scalar couplings. *J Am Chem Soc* 121(7):1601–1602
- Cornilescu G, Hu JS, Bax A (1999) Identification of the hydrogen bonding network in a protein by scalar couplings. *J Am Chem Soc* 121(12):2949–2950
- Cornilescu G, Didychuk AL, Rodgers ML, Michael LA, Burke JE, Montemayor EJ, Hoskins AA, Butcher SE (2016) Structural analysis of multi-helical RNAs by NMR-SAXS/WAXS: application to the U4/U6 di-snRNA. *J Mol Biol* 428(5 Pt A):777–789. <https://doi.org/10.1016/j.jmb.2015.11.026>
- Cuniasse P, Tavares P, Orlova EV, Zinn-Justin S (2017) Structures of biomolecular complexes by combination of NMR and cryoEM methods. *Curr Opin Struct Biol* 43:104–113. <https://doi.org/10.1016/j.sbi.2016.12.008>
- Dashti H, Tonelli M, Lee W, Westler WM, Cornilescu G, Ulrich EL, Markley JL (2016) Probabilistic validation of protein NMR chemical shift assignments. *J Biomol NMR* 64(1):17–25. doi:10.1007/s10858-015-0007-8
- Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPIPE – a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6(3):277–293
- Denisov IG, Grinkova YV, Lazarides AA, Sligar SG (2004) Directed self-assembly of monodisperse phospholipid bilayer nanodiscs with controlled size. *J Am Chem Soc* 126(11):3477–3487
- Douglas SM, Chou JJ, Shih WM (2007) DNA-nanotube-induced alignment of membrane proteins for NMR structure determination. *Proc Natl Acad Sci U S A* 104(16):6644–6648. <https://doi.org/10.1073/pnas.0700930104>
- Eghbalnia HR, Markley JL (2017) Chapter 5 acquisition and post-processing of reduced dimensionality NMR experiments. In: *Fast NMR data acquisition: beyond the Fourier transform*. The Royal Society of Chemistry, pp 96–118. <https://doi.org/10.1039/9781782628361-00096>
- Eghbalnia HR, Bahrami A, Tonelli M, Hallenga K, Markley JL (2005a) High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. *J Am Chem Soc* 127(36):12528–12536
- Eghbalnia HR, Wang L, Bahrami A, Assadi A, Markley JL (2005b) Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. *J Biomol NMR* 32(1):71–81
- Fitzgerald PMD, Westbrook JD, Bourne PE, McMahon B, Watenpaugh KD, Berman HM (2005) 4.5 macromolecular dictionary (mmCIF). In: Hall SR, McMahon B (eds) *International tables for crystallography G. Definition and exchange of crystallographic data*. Springer, Dordrecht, pp 295–443
- Fitzkee NC, Bax A (2010) Facile measurement of (1)H-(1)5N residual dipolar couplings in larger perdeuterated proteins. *J Biomol NMR* 48(2):65–70. <https://doi.org/10.1007/s10858-010-9441-9>
- Gal M, Edmonds KA, Milbradt AG, Takeuchi K, Wagner G (2011) Speeding up direct (15)N detection: hCaN 2D NMR experiment. *J Biomol NMR* 51(4):497–504. <https://doi.org/10.1007/s10858-011-9580-7>
- Gayathri C, Tsarevsky NV, Gil RR (2010) Residual dipolar couplings (RDCs) analysis of small molecules made easy: fast and tuneable alignment by reversible compression/relaxation of reusable PMMA Gels. *Chemistry (Easton)* 16(12):3622–3626. <https://doi.org/10.1002/chem.200903378>
- Güntert P (2004) Automated NMR structure calculation with CYANA. *Methods Mol Biol* 278:353–378
- Gutmanas A, Adams PD, Bardiaux B, Berman HM, Case DA, Fogh RH, Guntert P, Hendrickx PM, Herrmann T, Kleywegt GJ, Kobayashi N, Lange OF, Markley JL, Montelione GT,

- Nilges M, Ragan TJ, Schwieters CD, Tejero R, Ulrich EL, Velankar S, Vranken WF, Wedell JR, Westbrook J, Wishart DS, Vuister GW (2015) NMR exchange format: a unified and open standard for representation of NMR restraint data. *Nat Struct Mol Biol* 22(6):433–434. <https://doi.org/10.1038/nsmb.3041>
- Habeck M (2017) Bayesian modeling of biomolecular assemblies with Cryo-EM maps. *Front Mol Biosci* 4:15. <https://doi.org/10.3389/fmolb.2017.00015>
- Hall SR (1991) The STAR file: a new format for electronic data transfer and archiving. *J Chem Inform Comput Sci* 31:326–333
- Hall SR, Cook APF (1995) STAR dictionary definition language: initial specification. *J Chem Inform Comput Sci* 35:819–825
- Hall SR, Spadaccini N (1994) The STAR file: detailed specifications. *J Chem Inform Comput Sci* 34:505–508
- Hall SR, Allen FH, Brown ID (1991) The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Cryst A* 47:655–685
- Hansen MR, Mueller L, Pardi A (1998) Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat Struct Biol* 5(12):1065–1074
- Hewitt L, McDonnell JM (2004) Screening and optimizing protein production in *E. coli*. *Methods Mol Biol* 278:1–16. <https://doi.org/10.1385/1-59259-809-9:001>
- Hiller S, Wider G, Wüthrich K (2008) APSY-NMR with proteins: practical aspects and backbone assignment. *J Biomol NMR* 42(3):179–195
- Hyberts SG, Arthanari H, Wagner G (2011) Applications of non-uniform sampling and processing. *Top Curr Chem*. [https://doi.org/10.1007/128\\_2011\\_187](https://doi.org/10.1007/128_2011_187)
- Johnson BA (2004) Using NMRView to visualize and analyze the NMR spectra of macromolecules. *Methods Mol Biol* 278:313–352
- Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Mei Ono A, Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature* 440(7080):52–57
- Kay LE (2016) New views of functionally dynamic proteins by solution NMR spectroscopy. *J Mol Biol* 428(2 Pt A):323–331. <https://doi.org/10.1016/j.jmb.2015.11.028>
- Khirich G, Loria JP (2015) Complexity of protein energy landscapes studied by solution NMR relaxation dispersion experiments. *J Phys Chem B* 119(9):3743–3754. <https://doi.org/10.1021/acs.jpcc.5b00212>
- Kigawa T, Matsuda T, Yabuki T, Yokoyama S (eds) (2007) Bacterial cell-free system for highly efficient protein synthesis. *Cell-free protein synthesis: methods and protocols*, October, 2007 edn. Wiley-VCH, Weinheim
- Kneller DG, Kuntz ID (1993) UCSF SPARKY – an NMR display, annotation and assignment tool. *J Cell Biochem Suppl* 17C:254–254
- Koradi R, Billeter M, Engeli M, Güntert P, Wüthrich K (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J Magn Reson* 135(2):288–297
- Kost TA, Condeary JP, Jarvis DL (2005) Baculovirus as versatile vectors for protein expression in insect and mammalian cells. *Nat Biotechnol* 23(5):567–575. <https://doi.org/10.1038/nbt1095>
- Kovacs H, Moskau D, Spraul M (2005) Cryogenically cooled probes – a leap in NMR technology. *Prog Nucl Magn Reson Spectrosc* 46(2–3):131–155. <https://doi.org/10.1016/j.pnmrs.2005.03.001>
- Krahenbuhl B, El Bakkali I, Schmidt E, Güntert P, Wider G (2014) Automated NMR resonance assignment strategy for RNA via the phosphodiester backbone based on high-dimensional through-bond APSY experiments. *J Biomol NMR* 59(2):87–93. <https://doi.org/10.1007/s10858-014-9829-z>
- Kupce E, Freeman R (2003) Projection-reconstruction of three-dimensional NMR spectra. *J Am Chem Soc* 125(46):13958–13959. <https://doi.org/10.1021/ja038297z>
- Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc Natl Acad Sci U S A* 109(27):10873–10878. <https://doi.org/10.1073/pnas.1203013109>

- Lee W, Markley JL (2018) PINE-SPARKY.2 for automated NMR-based protein structure research. *Bioinformatics* 34(9):1586–1588
- Lee W, Bahrami A, Markley JL (2013a) ADAPT-NMR enhancer: complete package for reduced dimensionality in protein NMR spectroscopy. *Bioinformatics* 29(4):515–517. <https://doi.org/10.1093/bioinformatics/bts692>
- Lee W, Hu K, Tonelli M, Bahrami A, Neuhardt E, Glass KC, Markley JL (2013b) Fast automated protein NMR data collection and assignment by ADAPT-NMR on Bruker spectrometers. *J Magn Reson* 236:83–88. <https://doi.org/10.1016/j.jmr.2013.08.010>
- Lee W, Stark JL, Markley JL (2014) PONDEROSA-C/S: client-server based software package for automated protein 3D structure determination. *J Biomol NMR* 60(2–3):73–75. doi:10.1007/s10858-014-9855-x
- Lee W, Tonelli M, Markley JL (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* 31(8):1325–1327. <https://doi.org/10.1093/bioinformatics/btu830>
- Lee W, Cornilescu G, Dashti H, Eghbalnia HR, Tonelli M, Westler WM, Butcher SE, Henzler-Wildman KA, Markley JL (2016) Integrative NMR for biomolecular research. *J Biomol NMR* 64(4):307–332. <https://doi.org/10.1007/s10858-016-0029-x>
- Liang B, Tamm LK (2016) NMR as a tool to investigate the structure, dynamics and function of membrane proteins. *Nat Struct Mol Biol* 23(6):468–474. <https://doi.org/10.1038/nsmb.3226>
- Linsler R (2017) Solid-state NMR spectroscopic trends for supramolecular assemblies and protein aggregates. *Solid State Nucl Magn Reson* 87:45–53. <https://doi.org/10.1016/j.ssnmr.2017.08.003>
- Lipchick JM, Loria JP (2009) Monitoring molecular interactions by NMR. *Methods Mol Biol* 490:115–134. [https://doi.org/10.1007/978-1-59745-367-7\\_5](https://doi.org/10.1007/978-1-59745-367-7_5)
- Lisi GP, Loria JP (2017) Allostery in enzyme catalysis. *Curr Opin Struct Biol* 47:123–130. <https://doi.org/10.1016/j.sbi.2017.08.002>
- Lorieau JL (2017) Mollib: a molecular and NMR data analysis software. *J Biomol NMR* 69(2):69–80. <https://doi.org/10.1007/s10858-017-0142-5>
- Maciejewski MW, Schuyler AD, Gryk MR, Moraru II, Romero PR, Ulrich EL, Eghbalnia HR, Livny M, Delaglio F, Hoch JC (2017) NMRbox: a resource for biomolecular NMR computation. *Biophys J* 112(8):1529–1534. <https://doi.org/10.1016/j.bpj.2017.03.011>
- Mackay JP, Landsberg MJ, Whitten AE, Bond CS (2017) Whaddaya know: a guide to uncertainty and subjectivity in structural biology. *Trends Biochem Sci* 42(2):155–167. <https://doi.org/10.1016/j.tibs.2016.11.002>
- Makino S, Beebe ET, Markley JL, Fox BG (2014) Cell-free protein synthesis for functional and structural studies. *Methods Mol Biol* 1091:161–178. [https://doi.org/10.1007/978-1-62703-691-7\\_11](https://doi.org/10.1007/978-1-62703-691-7_11)
- Mao B, Tejero R, Baker D, Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *J Am Chem Soc* 136(5):1893–1906. <https://doi.org/10.1021/ja409845w>
- Markley JL, Aceti DJ, Bingman CA, Fox BG, Frederick RO, Makino S, Nichols KW, Phillips GN Jr, Primm JG, Sahu SC, Vojtki FC, Volkman BF, Wrobel RL, Zolnai Z (2009) The center for eukaryotic structural genomics. *J Struct Funct Genom* 10(2):165–179. <https://doi.org/10.1007/s10969-008-9057-4>
- Matei E, Gronenborn AM (2015) F paramagnetic relaxation enhancement: a valuable tool for distance measurements in proteins. *Angew Chem Int Ed Engl* 55:150–154. <https://doi.org/10.1002/anie.201508464>
- Matthews S (2004) Perdeuteration/site-specific protonation approaches for high-molecular-weight proteins. *Methods Mol Biol* 278:35–45
- McFeeters RL, Fowler CA, Gaponenko VV, Byrd RA (2005) Efficient and precise measurement of H(alpha)-C(alpha), C(alpha)-C', C(alpha)-C(beta) and H(N)-N residual dipolar couplings from 2D H(N)-N correlation spectra. *J Biomol NMR* 31(1):35–47. <https://doi.org/10.1007/s10858-004-6057-y>

- Mertens HDT, Svergun DI (2017) Combining NMR and small angle X-ray scattering for the study of biomolecular structure and dynamics. *Arch Biochem Biophys* 628:33–41. <https://doi.org/10.1016/j.abb.2017.05.005>
- Metz G, Howard KP, Vanliemt WBS, Prestegard JH, Lugtenburg J, Smith SO (1995) Nmr-studies of ubiquinone location in oriented model membranes – evidence for a single Motionally-averaged population. *J Am Chem Soc* 117 (1):564-565. DOI:DOI. <https://doi.org/10.1021/ja00106a078>
- Molugu TR, Lee S, Brown MF (2017) Concepts and methods of solid-state NMR spectroscopy applied to biomembranes. *Chem Rev* 117(19):12087–12132. <https://doi.org/10.1021/acs.chemrev.6b00619>
- Montelione GT, Nilges M, Bax A, Guntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS, Berman HM, Kleywegt GJ, Markley JL (2013) Recommendations of the wwPDB NMR validation task force. *Structure* 21(9):1563–1570. <https://doi.org/10.1016/j.str.2013.07.021>
- Nasr ML, Baptista D, Strauss M, Sun ZJ, Grigoriu S, Huser S, Pluckthun A, Hagn F, Walz T, Hogle JM, Wagner G (2017) Covalently circularized nanodiscs for studying membrane proteins and viral entry. *Nat Methods* 14(1):49–52. <https://doi.org/10.1038/nmeth.4079>
- Ortiz-Polo G, Krishnamoorthi R, Markley JL, Live DH, Davis DG, Cowburn D (1986) Natural-abundance  $^{15}\text{N}$  NMR studies of Turkey Ovomucoid third domain. Assignment of peptide  $^{15}\text{N}$  resonances to the residues at the reactive site region via proton-detected multiple-quantum coherence. *J Magn Reson* 68:303–310
- Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpidis NC, Baker D (2017) Protein structure determination using metagenome sequence data. *Science* 355(6322):294–298. <https://doi.org/10.1126/science.aah4043>
- Palmer AG III, Kroenke CD, Loria JP (2001) Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods Enzymol* 339:204–238
- Perilla JR, Zhao G, Lu M, Ning J, Hou G, Byeon IL, Gronenborn AM, Polenova T, Zhang P (2017) CryoEM structure refinement by integrating NMR chemical shifts with molecular dynamics simulations. *J Phys Chem B* 121(15):3853–3863. <https://doi.org/10.1021/acs.jpcc.6b13105>
- Pervushin K, Riek R, Wider G, Wüthrich K (1998) Transverse relaxation-optimized spectroscopy (TROSY) for NMR studies of aromatic spin systems in  $^{13}\text{C}$ -labeled proteins. *J Am Chem Soc* 120:6394–6400
- Pickford AR, O’Leary JM (2004) Isotopic labeling of recombinant proteins from the methylotrophic yeast *Pichia pastoris*. *Methods Mol Biol* 278:17–33
- Pilla KB, Gaalswyk K, MacCallum JL (2017a) Molecular modelling of biomolecules by paramagnetic NMR and computational hybrid methods. *Biochim Biophys Acta* 1865:1654–1663. <https://doi.org/10.1016/j.bbapap.2017.06.016>
- Pilla KB, Otting G, Huber T (2017b) 3D computational modeling of proteins using sparse paramagnetic NMR data. *Methods Mol Biol* 1526:3–21. [https://doi.org/10.1007/978-1-4939-6613-4\\_1](https://doi.org/10.1007/978-1-4939-6613-4_1)
- Pritisanac I, Degiacomi MT, Alderson TR, Carneiro MG, Ab E, Siegal G, Baldwin AJ (2017) Automatic assignment of methyl-NMR spectra of supramolecular machines using graph theory. *J Am Chem Soc* 139(28):9523–9533. <https://doi.org/10.1021/jacs.6b11358>
- Rajesh S, Nietlispach D, Nakayama H, Takio K, Laue ED, Shibata T, Ito Y (2003) A novel method for the biosynthesis of deuterated proteins with selective protonation at the aromatic rings of Phe, Tyr and Trp. *J Biomol NMR* 27(1):81–86
- Rajesh S, Overduin M, Bonev BB (2016) NMR of membrane proteins: beyond crystals. *Adv Exp Med Biol* 922:29–42. [https://doi.org/10.1007/978-3-319-35072-1\\_3](https://doi.org/10.1007/978-3-319-35072-1_3)
- Ramelot TA, Raman S, Kuzin AP, Xiao R, Ma LC, Acton TB, Hunt JF, Montelione GT, Baker D, Kennedy MA (2009) Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. *Proteins* 75(1):147–167
- Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P, Gutmanas A, Guntert P, He Y, Herrmann T, Huang YJ, Jaravine V, Jonker HR, Kennedy MA, Lange OF, Liu G, Malliavin TE, Mani R, Mao B, Montelione GT,

- Nilges M, Rossi P, van der Schot G, Schwalbe H, Szyperski TA, Vendruscolo M, Vernon R, Vranken WF, Vries S, Vuister GW, Wu B, Yang Y, Bonvin AM (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. *Structure* 20(2):227–236. <https://doi.org/10.1016/j.str.2012.01.002>
- Rosenzweig R, Kay LE (2016) Solution NMR spectroscopy provides an avenue for the study of functionally dynamic molecular machines: the example of protein disaggregation. *J Am Chem Soc* 138(5):1466–1477. <https://doi.org/10.1021/jacs.5b11346>
- Salzmann M, Pervushin KV, Wider G, Senn H, Wüthrich K (1998) TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. *Proc Natl Acad Sci USA* 95(23):13585–13590
- Schmidt E, Güntert P (2012) A new algorithm for reliable and general NMR resonance assignment. *J Am Chem Soc* 134(30):12817–12829. <https://doi.org/10.1021/ja305091n>
- Schnieders R, Richter C, Warhaut S, de Jesus V, Keyhani S, Duchardt-Ferner E, Keller H, Wohner J, Kuhn LT, Breeze AL, Bermel W, Schwalbe H, Furtig B (2017) Evaluation of <sup>15</sup>N-detected H-N correlation experiments on increasingly large RNAs. *J Biomol NMR* 69:31–44. <https://doi.org/10.1007/s10858-017-0132-7>
- Schwieters CD, Bermejo GA, Clore GM (2017) Xplor-NIH for molecular structure determination from NMR and other data sources. *Protein Sci* 27:26–40. <https://doi.org/10.1002/pro.3248>
- Sekhar A, Kay LE (2013) NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc Natl Acad Sci U S A* 110(32):12867–12874. <https://doi.org/10.1073/pnas.1305688110>
- Sharaf NG, Gronenborn AM (2015) (<sup>19</sup>F)-modified proteins and (<sup>19</sup>F)-containing ligands as tools in solution NMR studies of protein interactions. *Methods Enzymol* 565:67–95. <https://doi.org/10.1016/bs.mie.2015.05.014>
- Shen Y, Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J Biomol NMR* 56(3):227–241. <https://doi.org/10.1007/s10858-013-9741-y>
- Shen Y, Bax A (2015) Homology modeling of larger proteins guided by chemical shifts. *Nat Methods* 12(8):747–750. <https://doi.org/10.1038/nmeth.3437>
- Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci U S A* 105(12):4685–4690. <https://doi.org/10.1073/pnas.0800256105>
- Shin J, Lee W, Lee W (2008) Structural proteomics by NMR spectroscopy. *Expert Rev Proteomics* 5(4):589–601. <https://doi.org/10.1586/14789450.5.4.589>
- Stanek J, Augustyniak R, Kozminski W (2012) Suppression of sampling artefacts in high-resolution four-dimensional NMR spectra using signal separation algorithm. *J Magn Reson* 214(1):91–102. <https://doi.org/10.1016/j.jmr.2011.10.009>
- Takeda M, Kainosho M (2012) Cell-free protein production for NMR studies. *Methods Mol Biol* 831:71–84. [https://doi.org/10.1007/978-1-61779-480-3\\_5](https://doi.org/10.1007/978-1-61779-480-3_5)
- Takeuchi K, Sun ZY, Wagner G (2008) Alternate <sup>13</sup>C-<sup>12</sup>C labeling for complete mainchain resonance assignments using C alpha direct-detection with applicability toward fast relaxing protein systems. *J Am Chem Soc* 130(51):17210–17211. <https://doi.org/10.1021/ja806956p>
- Takeuchi K, Heffron G, Sun ZY, Frueh DP, Wagner G (2010) Nitrogen-detected CAN and CON experiments as alternative experiments for main chain NMR resonance assignments. *J Biomol NMR* 47(4):271–282. <https://doi.org/10.1007/s10858-010-9430-z>
- Takeuchi K, Arthanari H, Shimada I, Wagner G (2015) Nitrogen detected TROSY at high field yields high resolution and sensitivity for protein NMR. *J Biomol NMR* 63(4):323–331. <https://doi.org/10.1007/s10858-015-9991-y>
- Takeuchi K, Arthanari H, Imai M, Wagner G, Shimada I (2016) Nitrogen-detected TROSY yields comparable sensitivity to proton-detected TROSY for non-deuterated, large proteins under physiological salt conditions. *J Biomol NMR* 64(2):143–151. <https://doi.org/10.1007/s10858-016-0015-3>

- Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat Methods* 12(8):751–754. <https://doi.org/10.1038/nmeth.3455>
- Tjandra N, Bax A (1997a) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* 278(5340):1111–1114
- Tjandra N, Bax A (1997b) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Errat Sci* 278(5344):1697–1697
- Tugarinov V, Kay LE (2005) Methyl groups as probes of structure and dynamics in NMR studies of high-molecular-weight proteins. *Chembiochem* 6(9):1567–1577
- Ulrich EL, Argentar D, Klimowicz A, Markley JL (1996) STAR/CIF macromolecular NMR data dictionaries and data file formats. *Acta Crystallogr A* 52(a1):C577–C577
- Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. *Nucleic Acids Res* 36(Database issue):D402–D408
- Valafar H, Prestegard JH (2004) REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* 167(2):228–241. <https://doi.org/10.1016/j.jmr.2003.12.012>
- Vallurupalli P, Sekhar A, Yuwen T, Kay LE (2017) Probing conformational dynamics in biomolecules via chemical exchange saturation transfer: a primer. *J Biomol NMR* 67(4):243–271. <https://doi.org/10.1007/s10858-017-0099-4>
- Venditti V, Egner TK, Clore GM (2016) Hybrid approaches to structural characterization of conformational ensembles of complex macromolecular systems combining NMR residual dipolar couplings and solution X-ray scattering. *Chem Rev* 116:6305–6322. <https://doi.org/10.1021/acs.chemrev.5b00592>
- Wang L, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. *J Biomol NMR* 32(1):13–22
- Whitley MJ, Xi Z, Bartko JC, Jensen MR, Blackledge M, Gronenborn AM (2017) A combined NMR and SAXS analysis of the partially folded cataract-associated V75D gammaD-Crystallin. *Biophys J* 112(6):1135–1146. <https://doi.org/10.1016/j.bpj.2017.02.010>
- Wüthrich K (1986) *NMR of proteins and nucleic acids*. Wiley Interscience, New York
- Young JY, Westbrook JD, Feng Z, Sala R, Peisach E, Oldfield TJ, Sen S, Gutmanas A, Armstrong DR, Berrisford JM, Chen L, Chen M, Di Costanzo L, Dimitropoulos D, Gao G, Ghosh S, Gore S, Guranovic V, Hendrickx PM, Hudson BP, Igarashi R, Ikegawa Y, Kobayashi N, Lawson CL, Liang Y, Mading S, Mak L, Mir MS, Mukhopadhyay A, Patwardhan A, Persikova I, Rinaldi L, Sanz-Garcia E, Sekharan MR, Shao C, Swaminathan GJ, Tan L, Ulrich EL, van Ginkel G, Yamashita R, Yang H, Zhuravleva MA, Quesada M, Kleywegt GJ, Berman HM, Markley JL, Nakamura H, Velankar S, Burley SK (2017) OneDep: unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure* 25(3):536–545. <https://doi.org/10.1016/j.str.2017.01.004>
- Zhao L, Pinon AC, Emsley L, Rossini AJ (2017) DNP-enhanced solid-state NMR spectroscopy of active pharmaceutical ingredients. *Magn Reson Chem* 56:583–609. <https://doi.org/10.1002/mrc.4688>

# Chapter 6

## The PA Tag: A Versatile Peptide Tagging System in the Era of Integrative Structural Biology



Zuben P. Brown and Junichi Takagi

**Abstract** We have recently developed a novel protein tagging system based on the high affinity interaction between an antibody NZ-1 and its antigen PA peptide, a dodecapeptide that forms a  $\beta$ -turn in the binding pocket of NZ-1. This unique conformation allows for the PA peptide to be inserted into turn-forming loops within a folded protein domain and the system has been variously used in general applications including protein purification, Western blotting and flow cytometry, or in more specialized applications such as reporting protein conformational change, and identifying subunits of macromolecular complexes with electron microscopy. Thus the small and “portable” nature of the PA tag system offers a versatile and powerful tool that can be implemented in various aspects of integrative structural biology.

**Keywords** Protein tagging · Affinity purification · Monoclonal antibody · Peptide insertion · EM label

### 6.1 Introduction

There is a growing demand for the structural and functional characterization of biological phenomena at the molecular level. These phenomena may involve large networks of complex biomolecules interacting at varying spatial and temporal frames, and so it is becoming increasingly important to approach these biological questions with multiple methods and techniques to successfully elucidate their structural basis at the atomic level. Since most structural methods require purified proteins reconstituted in an artificial system, obtaining pure and high-quality protein samples is a key determinant for the success of structural biology projects. However,

---

Z. P. Brown · J. Takagi (✉)

Laboratory of Protein Synthesis and Expression, Institute for Protein Research, Osaka University, Suita, Osaka, Japan

e-mail: [takagi@protein.osaka-u.ac.jp](mailto:takagi@protein.osaka-u.ac.jp)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,

Advances in Experimental Medicine and Biology 1105,

[https://doi.org/10.1007/978-981-13-2200-6\\_6](https://doi.org/10.1007/978-981-13-2200-6_6)

large macromolecular complexes are generally unstable and/or difficult to produce in a recombinant manner, therefore, it is crucial to employ highly efficient systems for the production and purification of target proteins. With this in mind, we have developed multiple affinity tagging systems of our own (Nogi et al. 2008; Sangawa et al. 2013; Tabata et al. 2010) and applied them to structural biology projects that involve purification of high-value target proteins (Kato et al. 2012; Kitago et al. 2015; Morita et al. 2016; Nagae et al. 2008; Nishimasu et al. 2011; Nogi et al. 2010). In particular, the recently-developed PA tag system proves to outperform many existing peptide-based immunoaffinity purification systems because of its universal applicability, speed, and cost-efficiency (Fujii et al. 2014). More importantly, a unique character of the PA tag system revealed by the structural analysis of the peptide-antibody complex was exploited to allow its use in various labeling applications that had not been possible with conventional peptide-based tag systems (Fujii et al. 2016a). In this chapter, we will explain how this unique antibody-epitope system can greatly expand the repertoire of tools available for investigating the structure and function of proteins, and outline some areas that may see its utility in solving difficult questions in integrative structural biology.

## 6.2 Protein Purification and Biochemical Analyses

### 6.2.1 Overview of Tag-Based Affinity Purification Systems

As most target proteins subjected to structural analysis nowadays are produced recombinantly rather than purified from natural sources, it is a common practice to express the proteins as a fusion with certain unnatural polypeptides that function as a purification handle, collectively called affinity tags. The size of the tag moiety can range from less than 10 residues (e.g., poly-His tag) to more than 50 kDa (e.g., Fc tag), but they all must be capable of binding to a specific purification matrix to allow preferential capture of the target protein compared with other impurities (a detailed review of the various techniques in (Terpe 2003)).

In an effective affinity purification system, the interaction between the tag and the matrix needs to show a number of properties including: high specificity to reduce contamination from unwanted molecules, high affinity to achieve complete capture of the tagged protein from the dilute sample, slow dissociation kinetics to withstand extensive washing steps, and availability of elution conditions that can achieve complete removal of the bound proteins from the matrix while being chemically harmless and cost efficient. In addition, it is very important that the tag is attached in such a way to not impair the structural and functional integrity of both the tag itself and the target protein. The last property is usually ensured by the placement of the tag moiety at either the N- or C-terminal of the protein, in order to maximize the separation between the tag and unaltered portion of the polypeptide chain. Naturally, no 'perfect' tagging system suitable for all experiments exists, and



**Table 6.1** Selected list of epitope tag systems

| Name   | sequence     | Affinity (Kd)      | Elution condition <sup>b</sup> | Antibody | References           |
|--------|--------------|--------------------|--------------------------------|----------|----------------------|
| FLAG   | DYKDDDDK     | 28 nM <sup>a</sup> | Low pH, EDTA, peptide          | M2       | Hopp et al. (1988)   |
| Myc    | EQKLISEEDL   | 2.2 nM             | Low pH                         | 9E10     | Evan et al. (1985)   |
| HA     | YPYDVPDYA    | 1.6 nM             | Peptide                        | 12CA5    | Field et al. (1988)  |
| PA     | GVAMPGAEDDVV | 0.4 nM             | MgCl <sup>2+</sup> , peptide   | NZ-1     | Fujii et al. (2014)  |
| TARGET | 5x(YPGQ)V    | 10 nM              | Propylene glycol, peptide      | P20.1    | Tabata et al. (2010) |
| MAP    | GDGMVPPGIEDK | 3.7 nM             | Peptide                        | PMab-1   | Fujii et al. (2016b) |
| AGIA   | EAAAGIARP    | 4.9 nM             | Peptide                        | Ra48     | Yano et al. (2016)   |
| CP5    | GQHVT        | 7.5 nM             | Peptide                        | Ra62     | Takeda et al. (2017) |
| RAP    | DMVNPGLRDRIE | 9.7 nM             | Peptide                        | PMab-2   | Fujii et al. (2017)  |

<sup>a</sup>Reported by Fuji and coworkers (2014). All other values are from the respective reference

<sup>b</sup>“peptide” refers to the competitive elution with a solution containing free epitope peptide

the ideal combination of purification tag and the target protein will depend on the intended experimental purpose and must be empirically determined.

Many purification systems have been developed including those based on metal-chelate interaction between Ni-bearing resin and poly-histidine (Sassenfeld and Brewer 1984), maltose binding protein binding to amylose resin (Maina et al. 1988), glutathione S-transferase binding to glutathione-resin (Smith and Johnson 1988), calmodulin binding peptide binding to calmodulin (Stofkohahn et al. 1992), or Strep-tag binding to streptavidin (Schmidt and Skerra 2007). Anti-peptide antibodies bound to an inert matrix offer another attractive set of protein purification strategies given the high affinity and specificity of antibodies, and the relatively small size of their epitopes. Several popular epitope-based purification systems are in use that involve the fusion with peptides such as FLAG (Hopp et al. 1988), HA (Field et al. 1988), and Myc (Evan et al. 1985) that can be captured by their respective antibodies. Epitope tag systems have a range of affinities, epitope sizes, chemical properties, viable cell expression systems and elution conditions (Table 6.1) and so the appropriate tag and affinity matrix needs to be selected based on experimental constraints. Accordingly, a great deal of research to develop new and potentially better-performing purification systems is still being underway (Yano et al. 2016; Fujii et al. 2016b).

## 6.2.2 Development of the PA Tag System

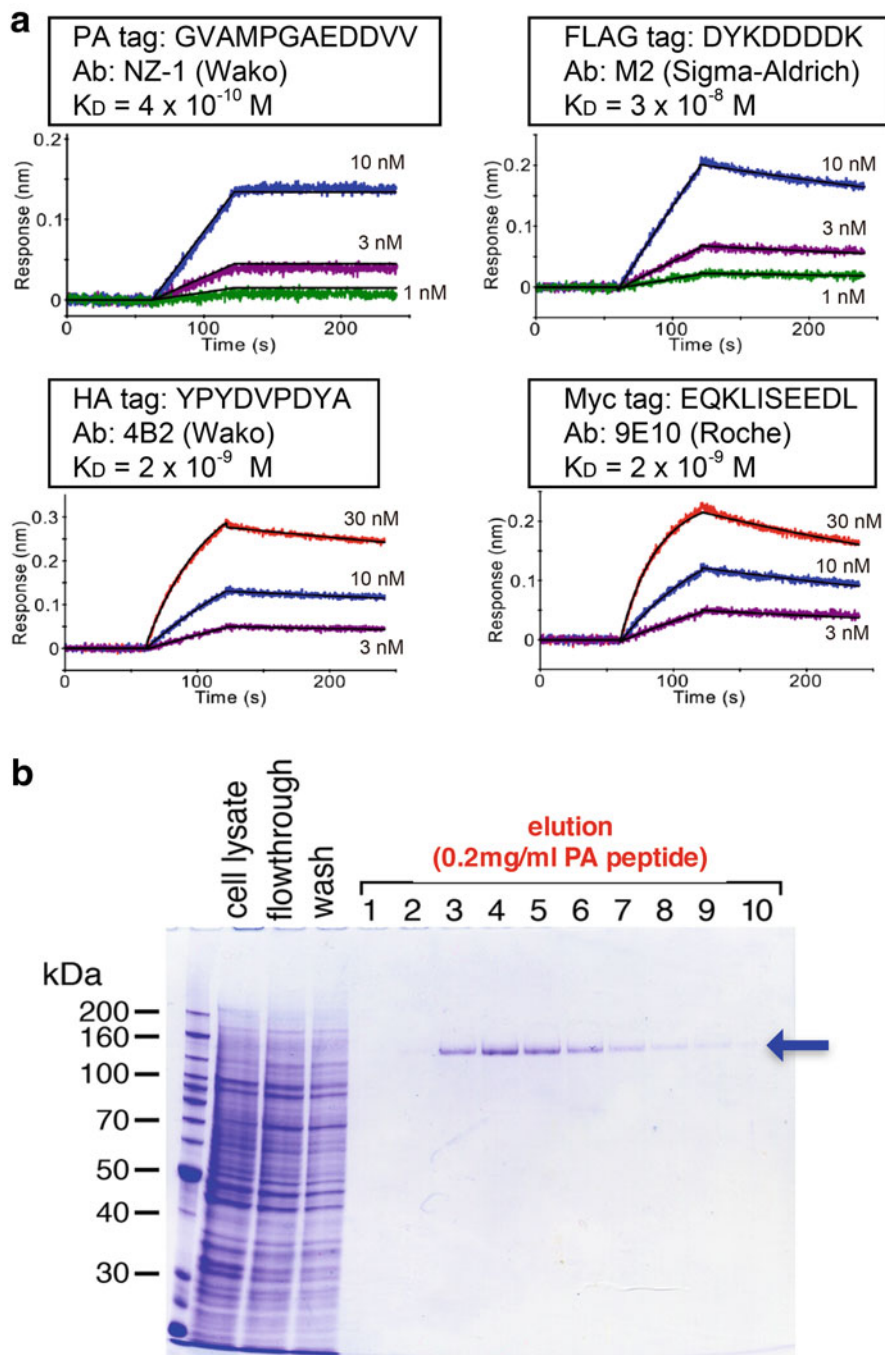
We recently reported the development of a novel epitope tag purification system based on the high affinity interaction between the NZ-1 antibody and a dodecapeptide (GVAMPGAEDDVV) called PA tag (Fujii et al. 2014). NZ-1 was established

during the search for anti-cancer antibodies as an inhibitor of platelet aggregation by its strong binding to the PLAG domain of podoplanin, a type I transmembrane protein that is over-expressed in cancer cells (Kato et al. 2006). As NZ-1 recognized not only the native podoplanin protein but also a synthetic peptide derived from the PLAG domain, we decided to see if it can be used as an anti-tag antibody.

During the initial characterization of the NZ-1 interaction with the epitope peptide, it showed a binding affinity that was orders of magnitude higher than popular and commercially available anti-tag antibodies including M2 (anti-FLAG), 9E10 (anti-Myc), or 4B2(anti-HA) when measured using Biolayer interferometry (Fig. 6.1a). More importantly, this high affinity was due to the very slow dissociation of antibody-peptide interaction, as evident from the near absence of the signal decline during the dissociation phase (i.e., time point after 120 sec in Fig. 6.1a). This property is highly desirable for an affinity tag system, because it allows for extensive washing steps to reduce the level of contamination from nonspecific binding. In fact, we successfully purified recombinant human epidermal growth factor receptor from the total cell lysate without contaminating proteins by fusing podoplanin derived dodecapeptide to the C-terminal and capturing the protein with NZ-1-immobilized Sepharose (Fig. 6.1b). Therefore, it became clear that NZ-1 can be implemented in a very efficient affinity purification system, and we designated the epitope dodecapeptide as PA tag. PA tag can be used in applications typical for any peptide-based tag systems, such as Western blotting, flow cytometry, and immunoprecipitation (Fujii et al. 2014). However, the greatest advantage of the PA tag over other existing systems is its ability to achieve complete affinity purification of the target protein in just one step, even from a very dilute and heavily contaminated crude material (Fig. 6.1b). Many structural biologists would agree that it is essential to use freshly prepared proteins to produce well-diffracting crystals or obtain high quality cryo-EM images. Since PA-tagged proteins purified by immobilized NZ-1 generally require less time during sample preparation compared to other technologies, we believe that the use of PA tag system will increase the success rate of challenging structural analyses, as our group has already demonstrated with numerous examples (Kitago et al. 2015; Arimori et al. 2017; Matoba et al. 2017; Matsunaga et al. 2016; Hirai et al. 2017). Another advantage of this system is that the NZ-1 resin can be regenerated by washing with non-denaturing and inexpensive buffer (3 M MgCl) and allowing for repeated uses without the loss in binding capacity, significantly reducing the running costs of experiments (Fujii et al. 2014).

### 6.2.3 Crystal Structure of PA Peptide Bound to NZ-1 Fab

The X-ray crystal structure of the NZ-1 fragment antigen binding (Fab) in both *apo* and PA peptide-bound forms was determined to better understand the high affinity interaction. High resolution crystal structures were obtained for NZ-1 Fab *apo* form at 1.65 Å and PA peptide-bound form at 1.70 Å (Fujii et al. 2016a). Upon comparison between the two structures, it became immediately clear that they are



**Fig. 6.1** High affinity and specificity of PA tag/NZ-1 system. **(a)** Binding affinities of various anti-tag antibodies against their epitope tags as measured by biolayer interferometry. NZ-1 (anti-PA),

essentially identical, indicating that there is very small conformational change of the antibody before and after the peptide binding. Typically, the complementary determining region (CDR) of an antibody undergoes significant conformational changes upon antigen binding, often showing the “induced-fit” type of ligand recognition mode. However, the total RMSD for the CDR region between *apo* and bound structures was only 0.466 Å. Furthermore, several water molecules that participate in the hydrogen bonding network to stabilize the bound PA peptide were already present in the *apo* form. The small conformational change between *apo* and bound states, as well as the presence of water molecules that mediate antigen binding in the absence of the peptide indicate that the binding pocket of NZ-1 is preformed or ‘primed’ for antigen recognition before the encounter, which could contribute to the high affinity of NZ-1 as there would be a very low entropic cost that NZ-1 needs to pay during a binding event.

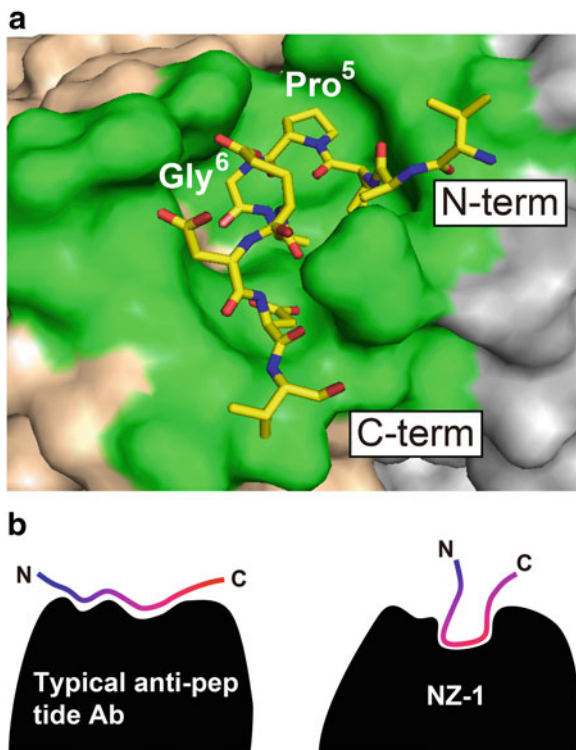
The overall structure of the binding pocket may also contribute to the high affinity of NZ-1 to PA peptide as the heavy and light chains of NZ-1 form a deep cleft that buries the PA peptide and covers over 1200 Å<sup>2</sup> of the total solvent-accessible surface area (ASA) (Fig. 6.2a). Although this value is not particularly high when compared to other known protein-peptide interaction surfaces (Chen et al. 2013), there are many hydrogen bonds and salt bridges formed across the interface together with numerous van der Waals contacts and a high shape complementarity all of which likely accounts for the large enthalpic gain upon complex formation.

The final component that may explain the high binding affinity of NZ-1 and PA peptide compared with other common epitope tag systems is the secondary structure of the PA peptide itself. Prior to the crystal structure being available, it was demonstrated using alanine scanning experiments that the central 7 residues of the PA peptide (shown in bold GVAM**PGAEDD**VV) were critical for recognition by NZ-1 (Fujii et al. 2014). This was confirmed by the X-ray structure as these amino acids were in direct contact with the antibody (Fig. 6.2a). Furthermore, the central “MPGA” motif formed a type II β-turn in the binding pocket, which is a commonly observed conformation for Pro-Gly sequence-containing peptides in solution (Guruprasad and Rajkumar 2000). This suggests that the PA peptide is also ‘primed’ for recognition by the NZ-1 CDR, giving another entropic advantage to the interaction.



**Fig. 6.1** (continued) M2 (anti-FLAG), 4B2 (anti-HA) or 9E10 (anti-Myc) antibodies were immobilized and serial dilutions of epitope tag attached to T4 lysozyme protein were tested. Equilibration (0–60s), association (60–120 s) and dissociation (120–240 s) stages are shown. **(b)** One-step purification of human EGFR C-terminally tagged with PA tag from total cell lysate using NZ-1 immobilized Sepharose. Purified EGF is marked with an arrow. (Reproduced after modifications from Fujii et al. 2014)

**Fig. 6.2** Unique mode of PA tag recognition by NZ-1. (a) X-ray crystal structure of PA peptide in the binding pocket of NZ-1 Fab (PDB ID: 4yo0). Peptide terminals, and both central proline and glycine residues that form the type II  $\beta$ -turn characteristic of the PA peptide are labeled. (b) Schematic comparison between the peptide-recognition modes of typical anti-peptide antibodies and NZ-1



#### 6.2.4 PA Tag as a “Mobile Epitope”

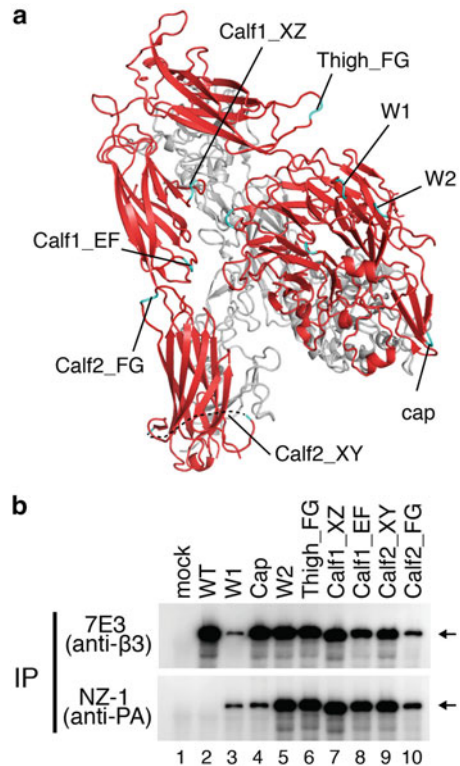
The structural analysis of the interaction between NZ-1 and PA peptide (described above) unraveled the structural causes for the extremely high affinity and showed that it involved multiple factors with favorable entropic and enthalpic energy terms. Although this alone was interesting information, we realized that the structure had a far more important implication regarding its utility as an epitope tag. In the NZ-1 binding pocket, the tip of the Pro-Gly  $\beta$ -turn of the PA peptide is inserted into the groove between the heavy and light chains (Fig. 6.2a). As a result, both the N-, and C-terminal portions of the peptide are not involved in binding recognition, and, importantly, point away from the antibody while being separated with a distance of  $\sim 10$  Å. This arrangement is rather unique, because most high affinity anti-peptide antibodies recognize relatively extended conformations of linear peptide within their antigen recognition groove to maximize the interacting surface (Fig. 6.2b). The recognition topology of NZ-1 suggests that the PA peptide could potentially remain

a viable epitope for NZ-1 even when constrained by neighboring residues; i.e., when inserted into the middle of a folded protein domain. The lack of direct interaction between the peripheral residues of the PA tag and NZ-1 also raise a possibility that the central segment of the PA tag may assume ideal conformation regardless of the flanking structures. This is fundamentally different from typical peptide tags, which are usually placed at either end of the target polypeptide because conformational flexibility and accessibility are generally the highest at these locations to ensure the full reactivity with a cognate antibody.

Many anti-peptide antibodies are generated by immunizing animals with synthetic peptides with a sequence that matches a certain segment of the original target protein. Such antibodies do not always recognize the native target antigen protein efficiently, because the *in situ* conformation of the peptide can be very different from that in solution (Dyson et al. 1988), leading to weak or no binding to the target epitope in the native protein (Hancock and O'Reilly 2005). For the very same reason, a peptide tag inserted into a topologically constrained protein domain may suffer from lower binding affinity with its anti-tag antibody due to unwanted conformational changes of the reactive epitope. While a systematic review of the reactivity of anti-tag antibodies toward peptide tags inserted into folded domains has not been done, our own investigation revealed that some common epitope tags (such as FLAG and Myc) lose reactivity to their antibody when inserted into these domains (see later section), presumably due to the inability of assuming the desired conformation in the context of the inserted topology. In order to be able to function in an “inserted” form, tag peptides need to be flanked by additional linker sequences (Facey and Kuhn 2003; Kendall and Senogles 2006) or strategically placed in preexisting long loop regions (Dinculescu et al. 2002; Morlacchi et al. 2012). Therefore, if the PA tag is universally “insertion-compatible” without the need for the linker optimization, it will have a high utility in a variety of research areas.

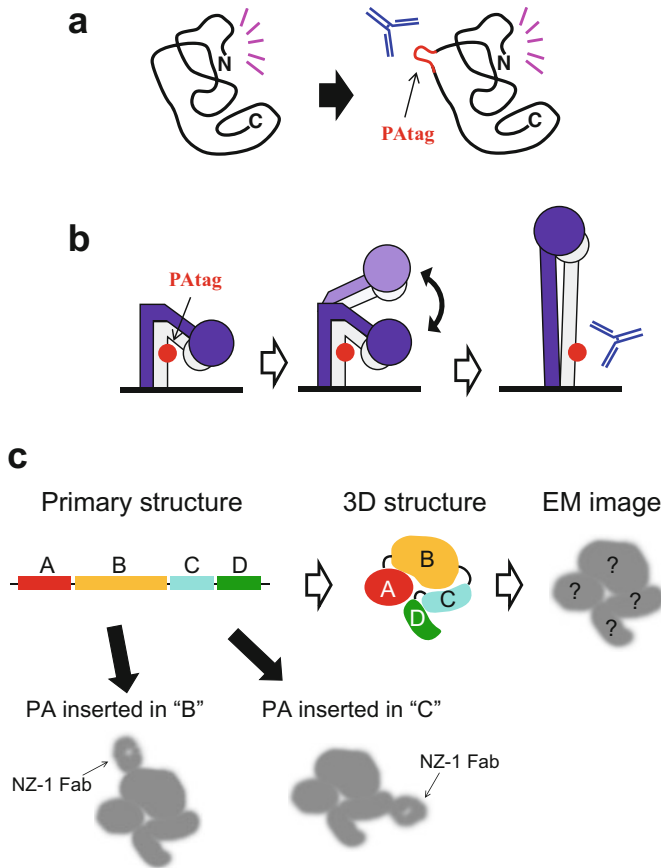
In order to test the insertion compatibility of the PA tag, we chose a platelet adhesion receptor  $\alpha$ IIB $\beta$ 3 integrin as the base protein. Integrins are a structurally and functionally diverse group of cell adhesion receptors made up of 18  $\alpha$ - and 8  $\beta$ -subunits to form 24 non-covalent heterodimers (Takagi and Springer 2002). They have a well characterized biology and undergo a distinct conformational change upon activation that involves an extension of the subunits from a bent to an extended conformation (Takagi and Springer 2002). Importantly, the extracellular portion of the  $\alpha\beta$ -heterodimer can be reconstituted as a soluble recombinant protein using an established design strategy (Takagi et al. 2002). The  $\alpha$ IIB subunit has a large extracellular region composed of four  $\beta$ -rich domains with multiple loops (Zhu et al. 2008), and is suitable for systematically investigating the insertion of the PA tag into loop regions (Fig. 6.3a). When the PA tag dodecapeptide was inserted in the middle of 8 selected loops of the  $\alpha$ IIB subunit and co-expressed with the  $\beta$ 3 subunit, most mutant integrins were efficiently expressed and secreted as in the case of the wild type version (Fig. 6.3b), indicating that the insertion did not cause serious structural disturbance. Furthermore, all these “PA-inserted” integrins were well recognized by NZ-1, suggesting that the native epitope structure was maintained

**Fig. 6.3** Insertion compatibility of PA tag. **(a)** Crystal structure of  $\alpha$ IIb $\beta$ 3 integrin in bent conformation (PDB ID: 3FCS) showing insertion sites for PA tag (cyan) in the  $\alpha$ -subunit (red). **(b)** Immunoprecipitation data showing the reactivity of NZ-1 to the PA tag when inserted into the indicated location. 7E3 antibody (anti- $\beta$ 3) was used to confirm the correct expression and formation of the  $\alpha$ IIb $\beta$ 3 integrin heterodimer. (Reproduced after modifications from (Fujii et al. 2016a)



(Fig. 6.3b). We confirmed that this insertion compatibility is highly unique to PA/NZ-1 interaction, because the identically constructed FLAG (DYKDDDDK) or Myc (EQKLISEEDL) tag-inserted integrin constructs completely lost reactivity with their antibodies (M2 and 9E10, respectively). It is surprising that PA tag can be successfully inserted into loops with varying base distances ranging from only 5 Å (between neighboring strands) to more than 15 Å (inter-sheet loops) (Fig. 6.3a). This supports our prediction that the NZ-1 recognition mode is insensitive to the flanking structures of PA tag, and the tag terminals are highly ‘adjustable’ and protect the epitope from a wide range of topological variability.

In addition to the  $\alpha$ IIb $\beta$ 3 integrin, we have also inserted PA tag into various loops of other membrane and soluble proteins and succeeded in purifying them (manuscripts in preparation). Although the insertion design has to be empirically determined for each case, we are confident that functionally active PA tag can be inserted into most folded domains. The question then becomes for what type of experimental applications will the insertion capability of PA tag becomes critically important? One obvious case is when the terminal regions of a target protein are not available for tagging, due to their inaccessibility in the native structure, or if they directly participate in a functionally important domain such as an active site. The N-terminal myristoylation motif and the C-terminal PDZ motif are examples of amino



**Fig. 6.4** Examples for utility of “portable” epitope system. (a) The PA peptide can be inserted into a central region of the polypeptide chain allowing antibody binding in cases where the terminal regions are unavailable due to its proximity with active site (depicted by magenta eyelash) or burial in the protein interior. (b) Antibody binding can be used as a conformational reporter in cases where the PA-tagged site alternates between hidden and exposed due to structural changes in the target protein. (c) In multi-module proteins, domain identity may be unclear during EM if there is no prior information about the domain architecture. Insertion of PA tag into target domains followed by labeling with NZ-1 Fab enables domain localization via differential EM imaging

acid sequence at the terminals that cannot be changed, and so alternative tagging strategies, such as insertions into central domain loops, are needed to preserve native-like structure and function. (Fig. 6.4a). We have already applied this strategy in the purification of neuroguidance factor semaphorin 3A which requires intact N- and C-termini to exhibit full biological activity (Fujii et al. 2016a), and for the adeno-associated virus capsid protein VP3 whose terminals are buried in the capsid (unpublished results).



### 6.2.5 *Monitoring and Controlling Conformational Change*

With careful design the PA tag can be inserted into the exposed loops of a target protein, acting as a ‘portable’ epitope, allowing the PA/NZ-1 pair to be used as a site-specific labeling system. One obvious use of such a system would be the monitoring of conformational change in flexible proteins. Proteins that undergo large conformational shifts resulting in the exposure of certain epitopes can be monitored by the change in binding of specific antibodies against them (Dennison et al. 2014; Humphries et al. 2003; Irannejad et al. 2013; Walker et al. 2004).

However, such special ‘conformation reporter’ antibodies are essentially only obtained by chance, and are not available for many proteins despite the obvious experimental applications that single molecule reporting can have. Several attempts have been made to fill this experimental niche by designing reporters based on small chromophore-bearing proteins such as GFP or cutinase as such conformation monitoring tags (Calleja et al. 2003; Bonasio et al. 2007), but they have not become widely used due to the potential structural and functional disturbances caused by their insertion. The PA peptide has two distinct advantages compared with other conformational reporting strategies. First, it is recognized by a single high affinity antibody (NZ-1) and so does not require any search of epitope-paratope space for antibodies that target a particular location, rather, the PA tag can be inserted into various locations enabling the identification of the tagging site with maximum reporting power (Fig. 6.4b). Second, the PA peptide is only 12 residues and so with rational design has minimal effect on the global architecture of the target protein after insertion. By embedding the epitope in a location that alternates between exposed and hidden depending on some structural and functional changes NZ-1 binding can be used as a conformational monitor (Fig. 6.4b). The integrins are known to undergo a major structural change on the cell surface during activation (Takagi et al. 2002), which makes it a perfect candidate for demonstrating the utility of PA tag and NZ-1 as conformational reporters. Among the 8 PA insertion positions tested in the  $\alpha$ IIB subunit, the Calf1\_EF site is located inside the subdomain interface and hence unavailable for NZ-1 antibody binding when integrin is inactive (Fig. 6.3a). However, during activation integrin takes on an extended conformation and so the Calf1\_EF insertion is predicted to be exposed. In line with our prediction, when we expressed Calf1\_EF integrin on the cell surface we saw an increase in NZ-1 binding upon cellular activation. Similar results were obtained when PA tag was inserted into different integrin subunits (mouse  $\beta$ 1) (Fujii et al. 2016a), indicating the broad applicability of this strategy.

In general, antibodies used for monitoring structural changes are also capable of affecting the equilibrium of functional states, because upon binding they may block a return to previous conformations and hence alter the structural equilibrium. In fact, binding of NZ-1 to the Calf1\_EF mutant integrin upregulates ligand binding by locking the receptor in an activated state (Fujii et al. 2016a). This is another area where the PA/NZ-1 system has many potential experimental applications. For example, the PA tag can be strategically placed in a surface-exposed loop region of

some protein with motile function (e.g., a motor protein) where the tag alone does not affect the function, but binding of  $\sim 50$  kDa NZ-1 Fab fragment onto the epitope physically inactivates the protein by enforcing a uniform conformation, which at the same time gives an ideal condition for the static structural analysis. Here, the important advantage of the PA tag/ NZ-1 system is the less invasive nature of the tag itself due to the small size, and the ability to achieve controlled ‘activation’ of the tag by labeling it with a large obstacle (i.e., NZ-1 Fab). We further exploited this property to expand the utility of the PA tag in another area of structural biology: electron microscopy.

## 6.3 Protein Labeling in EM Studies

### 6.3.1 Demands for EM Labeling Technologies

For large proteins, electron microscopy (EM) is becoming a highly popular method and EM-derived structures are routinely reaching atomic resolution in some cases within as little as 24 hours (Forsberg et al. 2017). On the other hand, EM analysis of small, flexible, or conformationally diverse proteins is more difficult and so reaching atomic resolution may not be possible. In these experiments, they may only yield intermediate resolution maps (typically  $>20$  Å). While these resolutions do not allow for the precise localization of amino acids, intermediate resolutions still give valuable global architectural and mechanistic information that is useful in understanding the function and structure of proteins, particularly when integrated with other sources of structural information (such as X-ray crystallography) (Matoba et al. 2017). Under these resolutions, however, the identity of the subunits or domains may be unclear, and so methods are required that can unambiguously identify them in the density map. One strategy is to use EM labeling techniques which utilize genetic manipulation of the target protein with insertions of extra polypeptides or deletions at a region of interest. The difference(s) between EM images (or 3D densities) of wild type and the mutant allows for the recognition of the altered density features as the site of modification and hence its identification.

The criteria for an ideal EM labeling method may include (1) the smallest possible genetic modification to the target complex to reduce the potential for unwanted structural alterations; (2) availability of the labeling agents with high specificity, high affinity, and an easily recognizable feature under EM; (3) a simple and efficient labeling step to ensure high occupancy without causing artificial conformational changes; and (4) temporal control over the ‘activation’ of the visualization label (c.f., genetically encoded constitutively visible tags). In the following section, we will outline some of the available techniques that will illustrate the basic EM labeling principles and discuss the use of PA tag and NZ-1 as a novel EM label method.

### 6.3.2 Presently Available EM Tags and Labeling Strategies

An early and demonstrative example of EM labeling is the identification of two toxin recognition sites in the acetylcholine receptor (AChR) by comparing averaged images of the unbound and bound proteins (Zingsheim et al. 1982). It was possible to identify the  $\alpha$ -subunits of the AChR by the additional densities that were present when the protein was in complex with the snake  $\alpha$ -neurotoxin, which binds nearly irreversibly to the  $\alpha$ -subunit. In order for this kind of analysis to be successful, two conditions have to be met; first, ligands such as an antibody or natural ligand (e.g., snake-derived neurotoxin) must be available, and second, the affinity and specificity of the ligand is high enough, with the location of the binding known to some extent. The second condition is particularly important, because partial (i.e., non-saturated) or non-specific binding would generate structural noise and results in many sub-populations during the EM image analysis, eventually leading to the unsuccessful identification of the binding locations.

In cases where ligands are not available or they bind with a low affinity, a deletion of subunits and/or domains followed by comparison between these mutants with the wild type complex can offer an alternative route for subunit/domain identification. In these cases, the missing density will show the location of the deleted subunit. The use of mutants that lack various components has been used quite successfully to determine the molecular architecture of complex cellular machineries such as cilia and flagella (Bui et al. 2008; Heuser et al. 2012; Heuser et al. 2009; Pigino et al. 2011). However, generalized application of this technique may be limited, as it is unreasonable to expect that such a range of mutants will be available for all target macromolecular complexes, and, if the mutation is in a structurally important location then its removal will prevent the correct structure being observed.

Another more direct strategy to mark and visualize particular sites within a target protein is to make recombinant proteins that incorporate an additional domain onto the site of interest. Various domain-incorporation labels have been developed, such as metallothionein tags that use heavy metal clusters to improve the EM contrast (Mercogliano and Derosier 2007; Nishino et al. 2007). However, many of these large tags are only tested in the ‘terminal fusion’ condition, and their applicability to a non-terminal marking (i.e., domain insertion) is not established. Internal placement of domain labels have been reported by utilizing some relatively small proteins such as GFP, taking advantage of the close spacing between the C-, and N-terminals that is compatible with insertion topology (Ciferri et al. 2012; Ciferri et al. 2015). As in the case of the domain deletion strategy, permanently attached labels may interfere with correct complex formation or execution of the function by the target protein, leaving uncertainty as to whether the obtained structure is authentic.

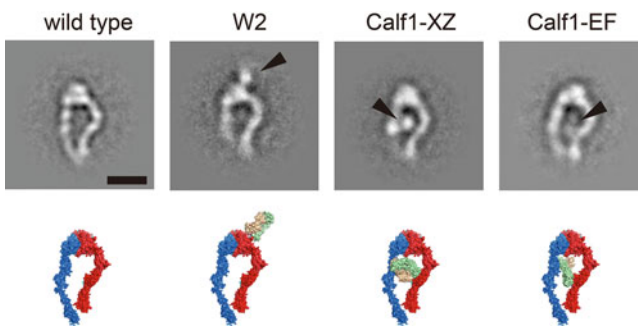
Labeling systems that are assembled at a later stage give temporal control over when to ‘activate’ the binding signal, and so can overcome some of the problems associated with genetically encoded tags or deletion-based strategies. The DID tag is based on yeast dynein light chain-interacting domain, and is assembled after protein expression upon addition of the appropriate binding partners (Flemming et al. 2010).

This tag is relatively small ( $\sim 80$  residues) and so is less likely to interfere with complex formation or folding, and can be visualized as a highly conspicuous feature after the label assembly. However, it must be placed at the terminal region of a protein, potentially limiting its utility. Antibody-based labeling is another way to realize positional mapping of proteins that can be ‘switched on’ at the desired time, and has been demonstrated with various monoclonal antibodies (Boisset et al. 1995; Boisset et al. 1993; Prasad et al. 1990). As production of good labeling antibodies against each target protein demands various resources and may not be possible in some cases (see sect. 2.4), the only reasonable option is to use epitope tags and their cognate antibodies. As mentioned earlier, however, such applications are frequently limited to the labeling of terminal regions of proteins (Buchel et al. 2001; Kelly et al. 2010).

### 6.3.3 PA Tag as an EM Label

As described in the earlier section, PA tag is small (12 residues) and can be placed in a variety of locations including in the middle of folded domains, constituting a highly unique ‘portable’ tagging system. Furthermore, the anti-PA tag antibody NZ-1 has very high affinity with extremely slow dissociation kinetics. All these features point to the possibility that the PA tag/NZ-1 may be the perfect tool to realize EM domain labeling (Fig. 6.4c).

To test this, we used the PA-inserted integrin constructs (Brown et al. 2017). In negative-stain EM, the soluble ectodomain fragment of  $\alpha$ IIb $\beta$ 3 integrin revealed particles with a flattened ring-like shapes made by two thin legs connected at both ends (Fig. 6.5, wild type), in agreement with the previously published integrin EM



**Fig. 6.5** Inserted PA tag can be visualized by NZ-1 Fab under negative stain EM. Representative 2D averages for wild type or PA-inserted  $\alpha$ IIb $\beta$ 3 integrin mutants are shown in the upper panels. Bound NZ-1 Fab is marked with a black arrowhead. Below each class average are shown predicted structures of  $\alpha$ IIb $\beta$ 3 integrin in the extended conformation, with the NZ-1 Fab binding simulated. The  $\alpha$ IIb subunit is shown in red,  $\beta$ 3 subunit in blue, and the heavy and light chains of NZ-1 Fab are in wheat and pale green, respectively. (Reproduced after modifications from Brown et al. 2017)

structures (Takagi et al. 2002). The major 2D class averages exhibited excellent agreement with the atomic model of the  $\alpha$ IIB $\beta$ 3 integrin created from the crystal structure (PDB ID: 3FCS) (Zhu et al. 2008). Particularly, the density profile of the  $\alpha$ IIB subunit was remarkably good in its detail, where all four domains ( $\beta$ -propeller, thigh, calf-1, and calf-2) could be resolved in most of the class averages. We chose three PA-insertion mutants (W2, Calf1\_XZ, and Calf1\_EF) and incubated them with an excess amount of NZ-1 Fab fragment. As expected, all mutants formed a very stable complex with NZ-1 Fab that could be isolated using size exclusion chromatography with no signs of dissociation. Upon negative staining and TEM observation, these mutant integrins exhibited structures identical to that of wild type integrin, except for the extra densities corresponding to the bound Fab (Fig. 6.5). The locations of the Fab densities were in perfect agreement with the that of PA tag insertion in each mutant, indicating that successful domain labeling was achieved. Importantly, a great majority (50–90 percent) of the integrin particles in the TEM images had clear density of bound Fab (Brown et al. 2017). This high prevalence is remarkable considering multiple factors including a chemical condition of negative staining that may facilitate Fab dissociation, potential invisibility of bound Fab due to projection overlap, and heterogeneous nature of the Fab-integrin orientation leading to the disappearance after image averaging. Since this method is applicable to a protein like integrin that is so small and have a highly polymorphic nature, we believe that PA tag insertion followed by NZ-1 Fab labeling should be considered as a useful and generally applicable method for EM domain mapping. In fact, this method was successfully applied to a recent cryo-EM analysis of yeast group II chaperonin TRiC/CCT complex made up with eight homologous but distinct subunits, allowing the unambiguous identification of each subunit which had been difficult due to the dynamic nature and the inherent pseudosymmetry (Wang et al. 2018).

## 6.4 Concluding Remarks

The discovery of the NZ-1 antibody and the subsequent structural characterization of its complex with the high affinity ligand peptide PA tag has allowed for the development of unique protein tagging system that has its utility in (1) protein purification, (2) sensitive immunodetection with Western blotting, flow cytometry, and immunoprecipitation, (3) analyzing and manipulating receptor conformation on cell surface, and (4) EM domain labeling. Given the useful properties of the PA-tag/NZ-1 system with its high affinity interaction and portable epitope functionality, we suspect that there are other applications within the structural biology field for this epitope-paratope system. For example, crystallization chaperoning is attracting much attention as a promising way to increase the likelihood of obtaining well-ordered crystals of biologically important and difficult targets for structural studies (Koide 2009). If NZ-1 Fab bound to the inserted PA tag can provide sufficient surface for the lattice contacts and promote crystallization, it will help crystallize, and hence solve structure of, proteins for which no antibodies or other binders/ligands

are available. PA-tag/NZ-1 may also have applications in cryo-EM where antibodies have been used to increase the mass of target proteins to improve image alignment, again PA-tag/NZ-1 offers a ready-made antibody system in cases where other high affinity antibodies are unavailable. The demonstrated utility of the PA tag system when used individually or in combination with other techniques can contribute greatly to the structural studies of difficult target proteins, and help solve important biological questions within the integrative structural biology field.

## References

- Arimori T, Kitago Y, Umitsu M et al (2017) Fv-clasp: an artificially designed small antibody fragment with improved production compatibility, stability, and crystallizability. *Structure* 25:1611–1622
- Boisset N, Radermacher M, Grassucci R et al (1993) Three-dimensional immunoelectron microscopy of scorpion hemocyanin labeled with a monoclonal fab fragment. *J Struct Biol* 111:234–244
- Boisset N, Penczek P, Taveau JC et al (1995) Three-dimensional reconstruction of androctonus australis hemocyanin labeled with a monoclonal fab fragment. *J Struct Biol* 115:16–29
- Bonasio R, Carman CV, Kim E et al (2007) Specific and covalent labeling of a membrane protein with organic fluorochromes and quantum dots. *Proc Natl Acad Sci U S A* 104:14753–14758
- Brown Z, Arimori T, Iwasaki K et al (2017) Development of a new protein labeling system to map subunits and domains of macromolecular complexes for electron microscopy. *J Struct Biol* 201:247–251
- Buchel C, Morris E, Orlova E et al (2001) Localisation of the PsbH subunit in photosystem II: a new approach using labelling of His-tags with a Ni(2+)-NTA gold cluster and single particle analysis. *J Mol Biol* 312:371–379
- Bui KH, Sakakibara H, Movassagh T et al (2008) Molecular architecture of inner dynein arms in situ in *Chlamydomonas reinhardtii* flagella. *J Cell Biol* 183:923–932
- Calleja V, Ameer-Beg SM, Vojnovic B et al (2003) Monitoring conformational changes of proteins in cells by fluorescence lifetime imaging microscopy. *Biochem J* 372:33–40
- Chen J, Sawyer N, Regan L (2013) Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci* 22:510–515
- Ciferri C, Lander GC, Maiolica A et al (2012) Molecular architecture of human polycomb repressive complex 2. *elife* 1:e00005
- Ciferri C, Lander GC, Nogales E (2015) Protein domain mapping by internal labeling and single particle electron microscopy. *J Struct Biol* 192:159–162
- Dennison SM, Anasti KM, Jaeger FH et al (2014) Vaccine-induced HIV-1 envelope gp120 constant region I-specific antibodies expose a CD4-inducible epitope and block the interaction of HIV-1 gp140 with galactosylceramide. *J Virol* 88:9406–9417
- Dinculescu A, McDowell JH, Amici SA et al (2002) Insertional mutagenesis and immunochemical analysis of visual arrestin interaction with rhodopsin. *J Biol Chem* 277:11703–11708
- Dyson HJ, Lerner RA, Wright PE (1988) The physical basis for induction of protein-reactive antipeptide antibodies. *Annu Rev Biophys Chem* 17:305–324
- Evan GI, Lewis GK, Ramsay GB, J. M. (1985) Isolation of monoclonal antibodies specific for human c-myc proto-oncogene product. *Mol Cell Biol* 5:3610–3616
- Facey SJ, Kuhn A (2003) The sensor protein KdpD inserts into the *Escherichia coli* membrane independent of the sec translocase and YidC. *Eur J Biochem* 270:1724–1734
- Field J, Nikawa J, Broek D et al (1988) Purification of a RAS-responsive adenylyl cyclase complex from *Saccharomyces cerevisiae* by use of an epitope addition method. *Mol Cell Biol* 8:2159–2165

- Flemming D, Thierbach K, Stelter P et al (2010) Precise mapping of subunits in multiprotein complexes by a versatile electron microscopy label. *Nat Struct Mol Biol* 17:775–778
- Forsberg BO, Aibara S, Kimanius D et al (2017) Cryo-EM reconstruction of the chlororibosome to 3.2 Å resolution within 24 h. *IUCr J* 4:723–727
- Fujii Y, Kaneko M, Neyazaki M et al (2014) PA tag: a versatile protein tagging system using a super high affinity antibody against a dodecapeptide derived from human podoplanin. *Protein Expres Purif* 95:240–247
- Fujii Y, Matsunaga Y, Arimori T, et al. (2016a) Tailored placement of a turn-forming PA tag into the structured domain of a protein to probe its conformational state. *J. Cell Sci.*, 1512–1522
- Fujii Y, Kaneko MK, Kato Y (2016b) MAP tag: a novel tagging system for protein purification and detection. *Monoclon Antib Immunodiagn Immunother* 35:293–299
- Fujii Y, Kaneko MK, Ogasawara S et al (2017) Development of RAP tag, a novel tagging system for protein detection and purification. *Monoclon Antib Immunodiagn Immunother* 36:68–71
- Guruprasad K, Rajkumar S (2000) Beta-and gamma-turns in proteins revisited: a new set of amino acid turn-type dependent positional preferences and potentials. *J Biosci* 25:143–156
- Hancock DC, O'Reilly NJ (2005) Synthetic peptides as antigens for antibody production. *Methods Mol Biol* 295:13–26
- Heuser T, Raytchev M, Krell J et al (2009) The dynein regulatory complex is the nexin link and a major regulatory node in cilia and flagella. *J Cell Biol* 187:921–933
- Heuser T, Barber CF, Lin J et al (2012) Cryoelectron tomography reveals doublet-specific structures and unique interactions in the II dynein. *Proc Natl Acad Sci U S A* 109:E2067–E2076
- Hirai H, Yasui N, Yamashita K et al (2017) Structural basis for ligand capture and release by the endocytic receptor ApoER2. *EMBO Rep* 18:982–999
- Hopp TP, Prickett KS, Price VL et al (1988) A short polypeptide marker sequence useful for recombinant protein identification and purification. *Nat Biotechnol* 6:1204–1210
- Humphries MJ, Symonds EJ, Mould AP (2003) Mapping functional residues onto integrin crystal structures. *Curr Opin Struct Biol* 13:236–243
- Irannejad R, Tomshine JC, Tomshine JR et al (2013) Conformational biosensors reveal GPCR signalling from endosomes. *Nature* 495:534–538
- Kato Y, Kaneko MK, Kuno A et al (2006) Inhibition of tumor cell-induced platelet aggregation using a novel anti-podoplanin antibody reacting with its platelet-aggregation-stimulating domain. *Biochem Biophys Res Commun* 349:1301–1307
- Kato K, Nishimasu H, Okudaira S et al (2012) Crystal structure of Enpp1, an extracellular glycoprotein involved in bone mineralization and insulin signaling. *Proc Natl Acad Sci U S A* 109:16876–16881
- Kelly DF, Lake RJ, Middelkoop TC et al (2010) Molecular structure and dimeric organization of the notch extracellular domain as revealed by electron microscopy. *PLoS One* 5:e10532
- Kendall RT, Senogles SE (2006) Investigation of the alternatively spliced insert region of the D2L dopamine receptor by epitope substitution. *Neurosci Lett* 393:155–159
- Kitago Y, Nagae M, Nakata Z et al (2015) Structural basis for amyloidogenic peptide recognition by sorLA. *Nat Struct Mol Biol* 22:199–206
- Koide S (2009) Engineering of recombinant crystallization chaperones. *Curr Opin Struct Biol* 19:449–457
- Maina CV, Riggs PD, Granda AG 3rd et al (1988) An Escherichia coli vector to express and purify foreign proteins by fusion to and separation from maltose-binding protein. *Gene* 74:365–373
- Matoba K, Mihara E, Tamura-Kawakami K et al (2017) Conformational freedom of the Irp6 ectodomain is regulated by n-glycosylation and the binding of the wnt antagonist dkk1. *Cell Rep* 18:32–40
- Matsunaga Y, Bashiruddin NK, Kitago Y et al (2016) Allosteric inhibition of a semaphorin 4d receptor plexin b1 by a high-affinity macrocyclic peptide. *Cell Chem Biol* 23:1341–1350
- Mercogliano CP, Derosier DJ (2007) Concatenated metallothionein as a clonable gold label for electron microscopy. *J Struct Biol* 160:70–82

- Morita J, Kano K, Kato K et al (2016) Structure and biological function of ENPP6, a choline-specific glycerophosphodiester-phosphodiesterase. *Sci Rep* 6:20995
- Morlacchi S, Sciandra F, Bigotti MG et al (2012) Insertion of a myc-tag within alpha-dystroglycan domains improves its biochemical and microscopic detection. *BMC Biochem* 13:14
- Nagae M, Nishikawa K, Yasui N et al (2008) Structure of the F-spondin reeler domain reveals a unique beta-sandwich fold with a deformable disulfide-bonded loop. *Acta Crystallogr D Biol Crystallogr* 64:1138–1145
- Nishimasu H, Okudaira S, Hama K et al (2011) Crystal structure of autotaxin and insight into GPCR activation by lipid mediators. *Nat Struct Mol Biol* 18:205–212
- Nishino Y, Yasunaga T, Miyazawa A (2007) A genetically encoded metallothionein tag enabling efficient protein detection by electron microscopy. *J Electron Microsc* 56:93–101
- Nogi T, Sangawa T, Tabata S et al (2008) Novel affinity tag system using structurally defined antibody-tag interaction: application to single-step protein purification. *Protein Sci* 17: 2120–2126
- Nogi T, Yasui N, Mihara E et al (2010) Structural basis for semaphorin signalling through the plexin receptor. *Nature* 467:1123–1127
- Pigino G, Bui KH, Maheshwari A et al (2011) Cryoelectron tomography of radial spokes in cilia and flagella. *J Cell Biol* 195:673–687
- Prasad BV, Burns JW, Marietta E et al (1990) Localization of VP4 neutralization sites in rotavirus by three-dimensional cryo-electron microscopy. *Nature* 343:476–479
- Sangawa T, Tabata S, Suzuki K et al (2013) A multipurpose fusion tag derived from an unstructured and hyperacidic region of the amyloid precursor protein. *Protein Sci* 22:840–850
- Sassenfeld HM, Brewer SJ (1984) A polypeptide fusion designed for the purification of recombinant proteins. *Bio-Technol* 2:76–81
- Schmidt TG, Skerra A (2007) The strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nat Protoc* 2:1528–1535
- Smith DB, Johnson KS (1988) Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione *s*-transferase. *Gene* 67:31–40
- Stofkohahn RE, Carr DW, Scott JD (1992) A single step purification for recombinant proteins – characterization of a microtubule associated protein (map-2) fragment which associates with the type-II camp-dependent protein-kinase. *FEBS Lett* 302:274–278
- Tabata S, Nampo M, Mihara E et al (2010) A rapid screening method for cell lines producing singly-tagged recombinant proteins using the “TARGET tag” system. *J Proteome* 73:1777–1785
- Takagi J, Springer TA (2002) Integrin activation and structural rearrangement. *Immunol Rev* 186:141–163
- Takagi J, Petre BM, Walz T et al (2002) Global conformational rearrangements in integrin extracellular domains in outside-in and inside-out signaling. *Cell* 110:599–611
- Takeda H, Zhou W, Kido K et al (2017) CP5 system, for simple and highly efficient protein purification with a C-terminal designed mini tag. *PLoS One* 12:e0178246
- Terpe K (2003) Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol* 60:523–533
- Walker F, Orchard SG, Jorissen RN et al (2004) CR1/CR2 interactions modulate the functions of the cell surface epidermal growth factor receptor. *J Biol Chem* 279:22387–22398
- Wang H, Han W, Takagi J et al (2018) Yeast inner-subunit PA-NZ-1 labeling strategy for accurate subunit identification in a macromolecular complex through cryo-EM analysis. *J Mol Biol* 430:1417–1425
- Yano T, Takeda H, Uematsu A et al (2016) AGIA tag system based on a high affinity rabbit monoclonal antibody against human dopamine receptor D1 for protein analysis. *PLoS One* 11:e0156716
- Zhu J, Luo BH, Xiao T et al (2008) Structure of a complete integrin ectodomain in a physiologic resting state and activation and deactivation by applied forces. *Mol Cell* 32:849–861
- Zingsheim HP, Barrantes FJ, Frank J et al (1982) Direct structural localization of two toxin-recognition sites on an ACh receptor protein. *Nature* 299:81–84



# Chapter 7

## Small Angle Scattering and Structural Biology: Data Quality and Model Validation



Jill Trehwella

**Abstract** This chapter provides a brief review of the current state-of-the-art in small-angle scattering (SAS) from biomolecules in solution in regard to: (1) sample preparation and instrumentation, (2) data reduction and analysis, and (3) three-dimensional structural modelling and validation. In this context, areas of ongoing research in regard to the interpretation of SAS data will be discussed with a particular focus on structural modelling using computational methods and data from different experimental techniques, including SAS (hybrid methods). Finally, progress made in establishing community accepted publication guidelines and a standard reporting framework that includes SAS data deposition in a public data bank will be described. Importantly, SAS data with associated meta-data can now be held in a format that supports exchange between data archives and seamless interoperability with the world-wide Protein Data Bank (wwPDB). Biomolecular SAS is thus well positioned to contribute to an envisioned federation of data archives in support of hybrid structural biology.

**Keywords** Small-angle scattering · SAXS · SANS · Biomolecular structure · Protein structure · Modelling · Data archive · Publication guidelines

### 7.1 Introduction

The potential for small-angle scattering (SAS) applications in structural biology was foreseen early in the development of the field. In their 1955 monograph Guinier and Fournet (1955) observed that, unlike synthetic polymers, biomolecules fold into well-defined structures that can meet the stringent requirements of purity and

---

J. Trehwella (✉)

School of Life and Environmental Sciences, The University of Sydney, NSW, Australia

Department of Chemistry, University of Utah, Salt Lake City, UT, USA

e-mail: [jill.trehwella@sydney.edu.au](mailto:jill.trehwella@sydney.edu.au)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,

Advances in Experimental Medicine and Biology 1105,

[https://doi.org/10.1007/978-981-13-2200-6\\_7](https://doi.org/10.1007/978-981-13-2200-6_7)

mono-dispersity necessary for accurate structural interpretation of solution SAS data. More than 60 years hence, it seems likely that the current level of activity in biomolecular SAS with sophisticated structural modelling for interpretation of data would exceed even the imagination of these pioneers.

The last decade has seen unprecedented advances in synchrotron and neutron sources with specialized beam-lines supporting biomolecular SAS, in commercial SAS instrumentation, in desk-top computing power, and in user-friendly SAS data analysis and modelling programs designed for the expert and non-expert alike. There also have been advances in the tools of molecular biology, biochemistry and sample characterization that have made possible solution SAS studies of increasingly challenging biomolecular complexes and assemblies that represent today's structural biology frontier. The result has been a steady rise in publications of biomolecular SAS studies, with a more than four-fold increase in annual totals over a dozen years to reach ~500 publications in 2016 (Franke et al. 2017).

The SAS intensity profile (generally expressed as  $I(q)$  vs  $q$ ; where  $q = \frac{4\pi \sin\theta}{\lambda}$ ,  $2\theta$  is the scattering angle and  $\lambda$  the wavelength of the radiation) contains information related to the shape of a scattering object and the distribution of scattering density within that shape. The intensity of the scattering signal is proportional to the square of the mean scattering density difference between the particle and its solvent (*i.e.* its “contrast”) and the square of its volume ( $V$ ). For biomolecules tumbling in solution, their random orientations result in rotational averaging of the scattering signal. As a result, all directional information is lost and the Fourier transform of  $I(q)$  vs  $q$  gives only the distribution of the pair-wise distances between scattering centers (atoms) within the biomolecule weighted by the product of their scattering powers relative to the solvent. Further, a solution SAS experiment measures the time and ensemble average of the scattering particles present. If the solution contains a mixture of different sized biomolecules, or there is an ensemble of conformers or flexibility, the measured profile represents the population weighted average of the structures present over the measurement period. For general biomolecular SAS reviews see (Jacques and Trewhella 2010; Koch et al. 2003; Rambo and Tainer 2010); for a comprehensive modern text on the subject see (Svergun et al. 2013).

An important goal for the structural biologist is an accurate and as precise as possible three-dimensional (3D) model of a biomolecule or biomolecular complex or assembly that informs our understanding of biological function. For a mono-disperse solution of essentially identical particles, the SAS profile yields accurate and precise parameters related to particle's size, shape, and internal structure; for example, radius of gyration ( $R_g$ ) to within a few 10th's of an Å, and volume ( $V$ ) or molecular mass ( $M$ ) to within 5–10%. The Fourier transform of  $I(q)$  yields the pair-wise atomic distance distribution,  $P(r)$  vs  $r$ , which is zero at  $r = 0$  and at the maximum dimension ( $d_{max}$ ) of the particle and provides further information on the scattering density distribution within the particle boundary.

Small-angle X-ray scattering (SAXS) is widely used for biomolecular analysis, with high intensity sources providing vast amounts of high precision data sets. Neutrons are more difficult to come by, but small-angle neutron scattering (SANS) with deuterium substitution and contrast variation enables structural analysis of individual components within complexes. In either case the SAS experiment is

conceptually simple, but technically demanding in terms of both sample preparation and instrumentation. The one-dimensional (1D) nature of the structural information encoded in the SAS profile and the averaging over the ensemble of structures present in the sample make it vulnerable to overfitting, over-interpretation, and even mis-interpretation. Nevertheless, with appropriate sample and data quality checks the SAS profile or SAS derived structural parameters can provide powerful restraints for 3D structural modelling, most especially when combined with complementary data (Trehwella 2016). The growth in biomolecular SAS, with an increasingly diverse community of users of the technique and increased focus on it as a contributor to hybrid/integrative structural modelling, made it imperative to establish a community agreed reporting framework for the field.

This chapter will present a brief outline of the current state of the art for SAS experiment and interpretation, significant issues regarding data interpretation that are the subject of ongoing research, and work that has been facilitated by the Commissions of the International Union of Crystallography (IUCr) and the world-wide Protein Data Bank (wwPDB) SAS validation task force (SASvtf) to establish a community agreed reporting framework for biomolecular SAS and tools for assessing data quality and model validation (Trehwella et al. 2017).

## 7.2 Current State of the Art

### 7.2.1 *Sample Preparation and Instrumentation*

To interpret solution SAS data accurately in terms of a 3D model, it is essential to demonstrate the SAS profile represents the form factor that encodes for the shape and scattering density distribution of the particle of interest. The samples must be highly pure and contain identical particles with respect to the resolution of the data (typically 10's of Å). Measurements of the sample plus an exact solvent blank are required in order to be able to accurately subtract the solvent contribution to the scattering. The subtracted SAS profile must represent the scattering from particles in the infinite dilution regime; that is free of aggregates (*i.e.* mono-disperse) and of inter-particle distance correlations. The dependence of the scattering signal on the square of the volume of the scattering particle means that small amounts of aggregation or oligomerization will measurably impact the SAS profile and the derived structural parameters will be too large. Distance correlations between particles that might arise from Columbic repulsion will give rise to a structure factor contribution to the scattering that suppresses the lowest-angle data and the derived structural parameters will be too small. Early reviews promoting the power of biomolecular SAS would often boast of the lack of need to crystallize or isotopically label the target of interest, as required for crystallography or NMR. In reality, crystallization can be a final purification step that rids a sample of impurities that would interfere with a SAS measurement and, unlike SAS, NMR is not sensitive to small amounts of large impurities or aggregates.

Thus, the requirements for purity and mono-dispersity for SAS are most stringent and have been a major limitation for accurate and precise measurement of the SAS profile for many, if not the majority, of high priority targets for structural biology research. As a result, success of the SAS experiment has always been highly dependent on the solubility of the target and the capacity to tune solvent conditions to find an optimal set where the measured SAS profile is in the infinite dilution regime. In some cases, measurement of a concentration series and point-by-point linear extrapolation of the SAS profiles to infinite dilution can remove concentration-dependent effects such as inter-particle distance correlations. Preparing for a biomolecular SAS experiment aimed at deriving 3D structural parameters thus involves first assessing samples for any concentration dependence to the SAS data that would be diagnostic of non-specific aggregation or inter-particle correlations. As needed, solution conditions might be adjusted (*e.g.* pH and/or ionic strength/species) or it may be determined that a concentration series and extrapolation to infinite dilution is required to obtain the desired form factor.

The past decade has seen a significant increase in the number of vendors offering laboratory-based SAXS systems that can be of high value for training, and can also provide high quality data locally and aid in evaluating samples in preparation for experiments at synchrotrons or neutron scattering facilities where access is limited and time restricted. In this same period, there has been a proliferation of SAXS beam-lines at synchrotron facilities world-wide, many dedicated solely to biological applications, with X-ray beam intensities and robotics that enable rapid measurement of samples (10's of milliseconds to seconds) using very small amounts of material (mg and smaller quantities) (*e.g.* Hura et al. 2009; Blanchet et al. 2015; Round et al. 2015). There have also been substantial developments of in-line purification and characterization capabilities at many synchrotron beamlines. Size exclusion chromatography (SEC) has proven especially powerful in combination with SAXS (Brennich et al. 2017; David and Perez 2009; Graewert et al. 2015; Mathew et al. 2004; Blanchet et al. 2015; Ryan et al. 2017).

The SEC-SAXS set up provides for separation of contaminants and/or aggregates in a sample or of species in polydisperse mixtures immediately prior to SAXS measurement. It is thus especially helpful for samples that are subject to time-dependent aggregation. SEC-SAXS also aids in obtaining precise solvent subtraction, as the solvent measurement is made on the sample free column flow through, and potentially also measures a useful range of sample concentrations as the sample elutes from the SEC column. The statistical quality of the data is limited by sample dilution on the column and the speed with which it elutes. The speed of the SEC-SAXS experiment overall is limited by the time for sample to traverse the column. Taking full advantage of the brightness of the synchrotron source and by judicious choice of columns, one can complete a SEC-SAXS experiment in less than 10 minutes and obtain good quality data with sample loadings of a few 10ths of mg (*e.g.* 100  $\mu\text{L}$  of 5  $\text{mg mL}^{-1}$  of a 20 kDa protein (Ryan et al. 2017)). Elimination or reduction of void volumes in the SEC-SAXS setup can reduce sample dilution and facilitates accurate correlation of UV measurements with SAXS data measurement for concentration determination of the biomolecular solute (Ryan

et al. 2017). This allows for calculation of its molecular mass  $M$  from  $I(0)$ , which is a primary validation parameter demonstrating that the scattering is from the particle of interest.

With SANS and selective deuteration the individual subunits of complexes or assemblies can be distinguished in contrast variation experiments (Gabel 2015; Jacques and Trewhella 2010; Whitten and Trewhella 2009; Whitten et al. 2008; Zaccai et al. 2016; Zaccai and Jacrot 1983). However, neutron sources are many orders of magnitude less bright than even laboratory X-ray sources, and thus sample sizes (typically 100's of  $\mu\text{L}$  at  $\text{mg/mL}$  concentrations) and exposure times (minutes to hours) historically have been a significant limitation. Also, neutron sources require a reactor or particle accelerator, and there are many fewer neutron scattering facilities compared to synchrotrons. Even so, the power of the contrast variation experiment with deuterium labelling, combined with the fact that neutrons are non-ionizing and hence less damaging than X-rays, has stimulated significant developments in SANS applications in structural biology. There is now a SEC-SANS capability at the Institut Laue-Langevin (on beam-line D22), where datasets can be acquired with exposure times that can be less than a minute and on relative small sample volumes (Jordan et al. 2016). In addition it is now possible to selectively perdeuterate individual domains within multi-domain proteins using sortase (Sonntag et al. 2017). In their elegant study of the three RNA recognition motif (RRM) domains in the RNA binding protein TIA-1, Sonntag et al. were able to precisely define relative domain arrangements using a segmental labelling strategy with SANS and contrast variation. This capability opens new possibilities for studying multi-domain proteins in solution and monitoring domain rearrangements, for example upon ligand binding or changes in physiological solution conditions.

## 7.2.2 Data Reduction and Error Propagation

Solution SAS data are recorded as counts on a detector, which is often two-dimensional (2D) and records an isotropic scattering pattern that is generally circularly averaged to maximize counts in the 1D intensity profile,  $I(q)$  vs  $q$ . Depending on the details of the instrument, corrections may be applied to account for detector non-linearity and sensitivity, and approaches to error propagation will vary based on detector characteristics (*e.g.* detectors may count individual X-rays or neutrons, or may be proportional counters). Accurate solvent subtraction to obtain  $I(q)$  vs  $q$  for the particle of interest requires precise normalization of the scattered intensity to constant counts on sample and solvent blank, which today can be better than 0.1%. Practice has been that data may or may not be placed on an absolute scale (in units of  $\text{cm}^{-1}$ ). Absolute scaling provides the opportunity to directly compare the results from different instruments, including X-ray and neutron instruments, and also allows for determination of  $M$  for the scattering particle from  $I(0)$  without reference to another protein, as was historically done but which introduces unnecessary additional errors.

Each of the details of data reduction to  $I(q)$  vs  $q$ , solvent subtraction and error propagation are often invisible to the experimenter, especially with the high levels of automation on SAS beam-lines today. These details, however, can have significant implications for the accuracy of intensities that can impact the derived structural parameters, and on the accuracy of propagated errors that affect the most commonly used model validation parameter,  $\chi^2$  (see Sect. 7.2.4). It is therefore important for beam-line scientists to provide details of their data acquisition and reduction protocols to experimenters in a format that makes complete recording and reporting of the experimental parameters easy. Auto-processing pipelines for data reduction to  $I(q)$  vs  $q$  also cannot substitute entirely for user engagement in validating their final solvent subtracted SAS profiles are accurate and suitable for structural interpretation.

An informal group of SAS instrument scientists and experimenters, who have adopted the acronym canSAS (collective action for nomadic Small Angle Scatterers, <http://www.cansas.org/>), works cooperatively to provide the SAS user community with shared tools and information. Their Reproducibility and Reliability working group supports round-robin measurements for calibration and comparison of results at different SAXS and SANS beam-lines. This working group is also considering the handling of different sources of error in SAS data, including systematic and statistical errors. This kind of volunteer community effort to address reproducibility and reliability and to establish standard data formats is important as the SAS field matures. Increased transparency and standardization in data reduction and error propagation protocols are essential for SAS researchers to be able to adequately report their results and archive data in a form that can support hybrid methods structural biology (see 7.4.2). Significant ongoing effort is required, particularly among instrument scientists and programmers at synchrotron and neutron beamlines, to achieve these important goals.

### 7.2.3 Data Analysis and Validation

There are a number of basic analyses of SAS data that are essential for data validation. These include Guinier (Guinier 1939) and  $P(r)$  analyses (Glatter 1977) and determination of  $V$  for the scattering particle using the Porod approximation ( $V_p$ ) (Porod 1951) which should be compared with the  $M$  determination from  $I(0)$ . In addition, the SAS profile must be assessed for indications of the degree of foldedness or flexibility using the Kratky (Kratky 1982) or dimensionless Kratky (Bizien et al. 2016; Durand et al. 2010) plots, or Porod-Debye plots (Rambo and Tainer 2011). Each of these analyses for assessing flexibility is critically dependent on accurate solvent subtraction, which as noted above hinges on having an exact solvent blank and accurate normalization of sample and solvent measurements to constant counts on sample.

For illustration purposes, these basic analyses are presented in Fig. 7.1 with derived structural parameters in Table 7.1 for an example protein: the intra-cellular

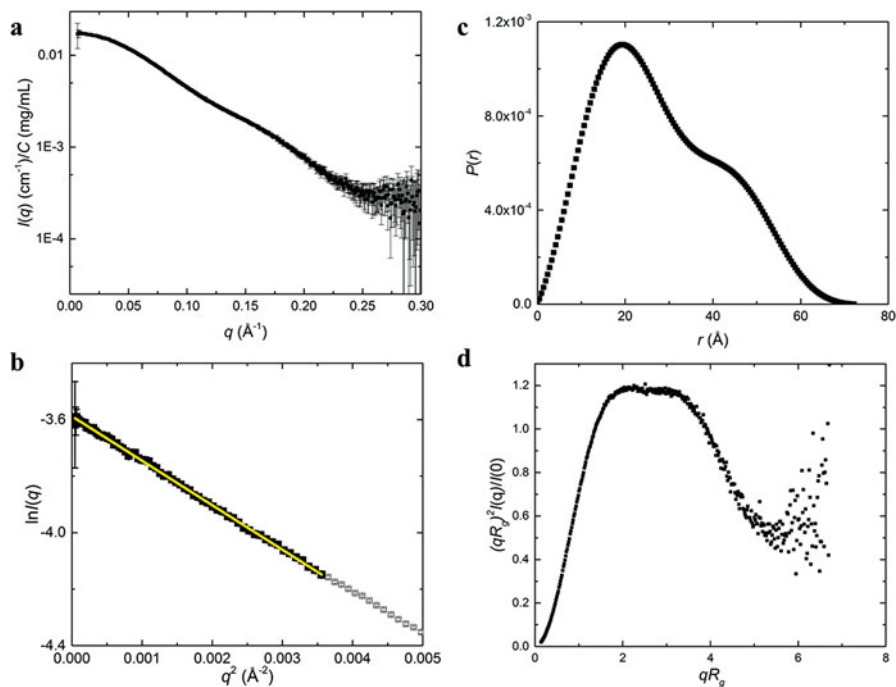
**Table 7.1** Derived structural parameters for calmodulin example

|   |                    |
|---|--------------------|
| Guinier analysis  |                    |
| $I(0)$ (cm <sup>-1</sup> )  | 0.0554 ± 0.00008   |
| $R_g$ (Å)   | 21.74 ± 0.06       |
| $q_{min}$ (Å <sup>-1</sup> )  | 0.007              |
| $qR_g$ max ( $q_{min} = 0.0066$ Å <sup>-1</sup> )                       | 1.3                |
| Coefficient of correlation, $R^2$                                       | 0.999              |
| $M$ from $I(0)$ <sup>a</sup> (ratio to predicted)                       | 21944 (1.31)       |
| $P(r)$ analysis   |                    |
| $I(0)$ (cm <sup>-1</sup> )  | 0.0533 ± 0.00006   |
| $R_g$ (Å)   | 22.2 ± 0.06        |
| $d_{max}$ (Å)   | 72                 |
| $q$ range (Å <sup>-1</sup> )  | 0.0074-0.3104      |
| $\chi^2$ (total estimate from GNOM)                                     | 0.855 (0.91)       |
| $M$ from $I(0)$ (ratio to predicted value)                              | 21,718 (1.29)      |
| Porod Volume (Å <sup>-3</sup> ) (ratio $V_p$ /calculated $M$ )          | 25,200 (1.5)       |
| $V, M$ using the Fischer method <sup>b</sup> (ratio of $M$ to expected) | 21,550,17.7 (1.05) |

<sup>a</sup>  $M = \frac{I(0)N_A}{C\Delta\rho_M^2}$  where  $\Delta\rho_M = \vartheta$  and  $\vartheta$  is the partial specific volume of CaM and  $\Delta\rho$  the scattering density difference between the solvent and CaM (Orthaber et al. 2000).  $C$  was calculated using a calculated extinction coefficient of 0.178 (for A280 0.1% w/v, 1 cm) (Gasteiger et al. 2005),  $\vartheta$  and  $\Delta\rho$  were calculated using MULCh (Whitten et al. 2008) based on volumes of the chemical constituents of CaM and its solvent (25 mM MOPS, 250 mM NaCl, 50 mM KCl, 2 mM TCEP, 0.1% NaN<sub>3</sub>, pH 7.5)

<sup>b</sup> Fisher et al. (2009)

Ca<sup>2+</sup>-receptor calmodulin (CaM), a 16.842 kDa calcium-binding protein (human isoform, Uniprot sequence P62155 (2–149)). The data are drawn from the example set described in full in (Trehwella et al. 2017), an open access article for which the CaM data are publicly available under the uniform resource identifier <https://creativecommons.org/licenses/by/2.0/uk/legalcode>. The data are also deposited in the SAS Biological Data Base (SASBDB) (deposition identifier SASDCQ2). The SAS intensity scale covers several orders of magnitude and so the  $I(q)$  vs  $q$  profile is presented as a log-linear plot (Fig. 7.1a). As expected for a monodisperse solution the Guinier plot (Fig. 7.1b) is linear and yields an  $R_g$  value consistent with previous observations (Heidorn and Trehwella 1988). The  $M$  value calculated from  $V_p$  using ATSAS (Franke et al. 2017) or the Fischer method (Fischer et al. 2009) agrees with the expected value from chemical composition, while the value of  $M$  derived from  $I(0)$  (Orthaber et al. 2000) is ~30% high. This latter high value can be attributed to the relatively large errors in the CaM concentration determination from UV measurement for non-tryptophan containing proteins that have very small extinction coefficients, and in partial specific volumes calculated from the volumes of chemical constituents for small proteins (<20 kDa). Determining  $M$  directly from the SAS profile using the different available methods, and understanding the origin of any observed differences is one important validation step to demonstrate the scattering



**Fig. 7.1** SAXS data for CaM. **(a)** Log-linear plot of solvent subtracted  $I(q)$  vs  $q$ , on an absolute scale and normalized to unit CaM concentration in  $\text{mg mL}^{-1}$ . **(b)** Guinier plot for the data in **a** with the linear fit (yellow line) (filled symbols indicate the Guinier region,  $qR_g < 1.3$ ). **(c)**  $P(r)$  vs  $r$  calculated as the indirect transform of the data in **a** using GNOM (as implemented in ATSAS 2.8.0 (Franke et al. 2017)). **(d)** Dimensionless Kratky plot

represents the form factor of the particle of interest. The crystal structure of CaM shows two globular domains connected by an extended  $\alpha$ -helix, while solution SAXS data previously showed that the globular domains were on average closer together than the crystal structure (Heidorn and Trehwella 1988). Subsequent NMR relaxation experiments revealed a 4-residue region in the helix connecting the two domains to be highly mobile (Barbato et al. 1992). Consistent with these results the  $P(r)$  function (Fig. 7.1c) is well behaved, approaching zero smoothly at  $r = 0$  and  $77 \text{ \AA}$  ( $d_{max}$ ) with a maximum at  $\sim 20 \text{ \AA}$  and a shoulder at  $\sim 45 \text{ \AA}$  consisted with a two-lobed elongated CaM structure. The dimensionless Kratky plot (Fig. 7.1d) shows a somewhat higher maximum than the usual 1.1 that is also shifted to  $qR_g = 2$  from the usual 1.75 value (Durand et al. 2010) with a shallow oscillation between 2.5 and 3.5 in  $qR_g$  reflecting the two-lobed elongated structure. Flexibility arising from mobile residues in the helix connecting the two domains is indicated by the increase in intensity for  $qR_g > 6$ .

The availability of easy to use SAXS and SANS data interpretation tools, including those facilitating the basic analyses described above in addition to 3D structural modelling, has helped grow structural biology SAS applications. A number of excellent program suites are freely available and well-documented. For



example, the BIOISIS web site (<http://www.bioisis.net/welcome>) offers scÅtter, a JAVA-based application for basic analysis of SAXS datasets along with tutorial material aimed at new and general users of biological SAXS. The much cited and broadly used ATSAS data acquisition and analysis package (Franke et al. 2017) provides a comprehensive set of SAXS- and SANS-data interpretation tools, including an extensive suite of 3D modeling programs, which are freely available to academic researchers. The US-SOMO suite of programs (<http://www.sas.uthscsa.edu/index.php>) includes SAXS and SANS modules to compute various hydrodynamic parameters and SAS profiles from biomolecular models, and an HPLC-SAXS module (Brookes et al. 2016) to deconvolve multiple species in the SEC-SAS profile for analysis of separated components. The MULCh suite of programs (available for download and as a web-based tool at <http://smb-research.smb.usyd.edu.au/NCVWeb/index.jsp> (Whitten et al. 2008)) is available to aid in planning and interpreting a SANS contrast variation series where a complex of biomolecular components having different mean scattering densities is measured in a series of solvents with different levels of deuteration. MULCh includes three modules: *Contrast* calculates the dependence of  $I(0)$  on contrast for X-rays and neutrons for a given solvent composition and/or deuteration levels in the biomolecular components and solvent; *R<sub>g</sub>* performs Stuhrmann and parallel axis theorem analyses that give the  $R_g$  values and separation distances for components having different mean scattering contrasts in the complex; *Compost* extracts component scattering functions that contain the shape information for individual components and a cross term that encodes information about their dispositions (Compost module).

### 7.2.4 3D Structural Modelling and Model Validation

With the basic analyses and data validation steps completed so that the SAS data can be judged suitable for 3D structural modelling, a SAS modelling strategy can be chosen: *e.g.* bead modelling to obtain basic shape information, rigid body modelling where domain of subunit structures are known and their positions and dispositions are optimized to fit the SAS profile(s), ensemble or multi-state modelling. The majority of 3D SAS modelling programs, optimize the model fit by minimizing, in some form,  $\chi^2$  where:

$$\chi^2 = \frac{1}{N-1} \sum_{i=1}^N \left[ \frac{I_{exp}(q_i) - cI_{mod}(q_i)}{\sigma(q_i)} \right]^2 \quad (7.1)$$

and  $N$  is the number of data points,  $I_{exp}(q_i)$  and  $I_{mod}(q_i)$  are the experimental and model intensities with  $c$  an adjustable scaling constant, and  $\sigma(q_i)$  the experimental errors. Assuming the errors have been accurately propagated from Poisson counting statistics and there are no systematic errors, a model that fits the data will have a  $\chi^2$  value near 1. In practice, reported  $\chi^2$  values for model fits can be anything from

a few tenths to 10's in magnitude as a result of the  $\sigma(q_i)$  in the denominator of Eq. 7.1 combined with substantial over or underestimation of the propagated counting statistics. In addition, as a global fit parameter over a rapidly decreasing intensity profile where the relative errors increase markedly with increasing  $q$ , the minimized  $\chi^2$  value can mask significant systematic mis-fit in important  $q$ -regions, *e.g.* the mid- $q$  region that is most sensitive to domain dispositions. As a result while  $\chi^2$  is useful for comparing model fits to the same data set, it is rendered essentially meaningless for comparing the model fits for different data sets and it is essential to use additional measures to validate a model.

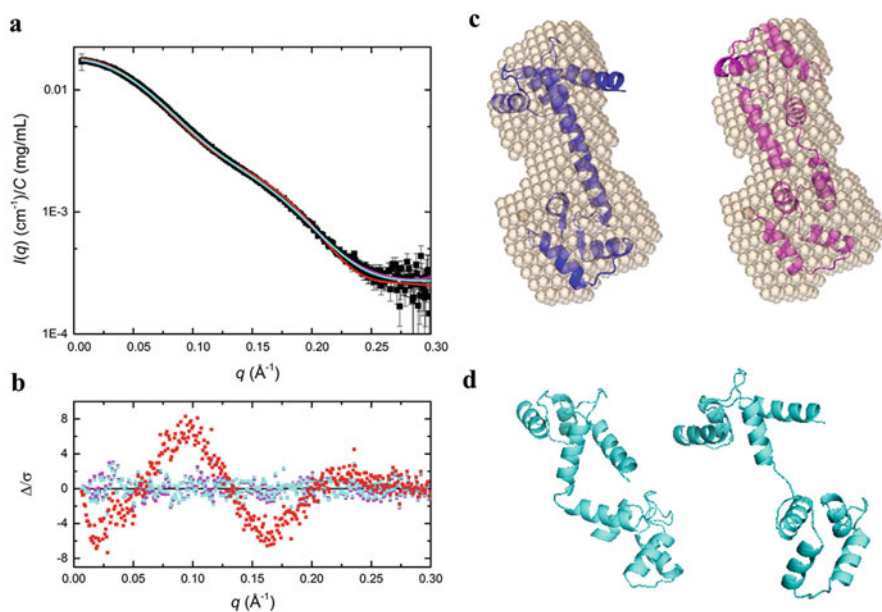
A simple and straightforward way to assess the quality of a model fit to a SAS profile is an error weighted residual plot. This plot will highlight any regions of systematic mis-fit, and the error weighting prevents the plot being dominated by areas of weaker scattering and high errors.

The CaM example provides an illustration of model fitting to SAS data using a simple uniform scattering density bead model or rigid-body modelling (Figs. 7.2 and 7.3) (data and models from (Trehwella et al. 2017) an open access article for which the CaM data and models are publicly available under the uniform resource identifier <https://creativecommons.org/licenses/by/2.0/uk/legalcode> and deposited in the SAS Biological Data Base (SASBDB) (deposition identifier SASDCQ2). Rigid body modelling used the CaM domains from the crystal structure with or without the flexible linker connecting them as identified by NMR relaxation measurements. The different CaM model fits are shown superimposed on the standard log-linear plot of the SAS data (Fig. 7.2a) and it is immediately evident that the error weighted residual difference plot (Fig. 7.2b) highlights much more clearly differences between models and data. The wave-like systematic deviation in the mid- $q$  region for the crystal structure fit is diagnostic of the fact that the average dispositions of the globular domains are not consistent with the crystal structure. Superposition of the crystal structure onto the bead-model shows significant parts of the crystal structure extending beyond its limits (Fig. 7.2c, left). Adding the flexible linker connecting the domains, the multi-state modelling program MultiFoXS can fit the SAS profile much better with either a 1-state or 2-state model (Fig. 7.2c, d, respectively), the latter having the lowest  $\chi^2$  value. The one state model is a better overall fit within the bead model envelop (Fig. 7.2c, right). Table 7.2 summarizes the CaM modeling parameters and  $\chi^2$  values, which for the best model fits are near 1 as expected for propagated counting statistics.

To address the cases where the errors are not true Poisson counting statistics, Franke et al. (2015) have developed an approach to model validation that does not depend on the magnitude of the specified errors. Their approach uses an all data point variance and covariance correlation matrix (or Correlation Map, CorMap) with a probability assessment for data-model fits. In simple terms the method assigns a probability in the form of a  $P$ -value (based on a 1-tailed Schilling test) for finding the longest string of experimental data points that lie systematically above (+1) or below (-1) the model profile. The  $P$ -value lies between 0-1 and a significance threshold,  $\alpha$ , is chosen below which the model fit is judged to show systematic deviation from experiment. As implemented in the most recent ATSAS package

(Franke et al. 2017), CorMap assigns significance to  $\alpha$ -values in the typical range statisticians use to indicate significant deviation, 0.01–0.05. The program generates a 2D plot with X and Y axes running from  $q_{min} - q_{max}$  and  $q_{max} - q_{min}$  and assigns the point to be black ( $-1$ ) or white ( $+1$ ) depending on whether it is above or below the model fit. The largest region of difference is identified by green, yellow or red; red indicates the  $P$ -value is  $<0.01$ , yellow is for  $0.01 < P < 0.05$  and green for  $P > 0.05$ . The higher the  $P$ -value, the more uniformly gray the correlation map appears as the white  $+1$  and black  $-1$  areas become small.

Correlation maps for the CaM models of Fig. 7.2 are illustrated in Fig. 7.3 and the corresponding  $P$ -values are given in Table 7.2. The poor fit of the crystal structure is boldly evident in the large red region that indicates a long stretch of 95 points (of a total of 390) that fall one side of the model profile (Fig. 7.2a), while the bead model



**Fig. 7.2** SAXS modelling results for CaM. **(a)** Log-linear plot of the solvent subtracted  $I(q)$  vs  $q$  profile (black squares) with model profiles calculated for the crystal structure of CaM (red squares) using CRY SOL with default parameters and PDB coordinates 1CLL, and for 1- and 2-state CaM models described in Table 7.2 (magenta and cyan, respectively) calculated using MultiFoXS (Schneidman-Duhovny et al. 2016) and assuming residues 77–81 are flexible. **(b)** Error-weighted residual plot for the models in **a** and using the same color key;  $\frac{\Delta}{\sigma} = \frac{I_{exp}(q_i) - cI_{mod}(q_i)}{\sigma(q_i)}$  where  $I_{exp}(q_i)$  and  $I_{mod}(q_i)$  are the experimental and model data points respectively,  $\sigma(q_i)$  are the experimental errors and  $c$  a scaling constant. **(c)** Gray spheres represent the bead model for CaM (calculated using DAMMIN and the  $P(r)$  profile in Fig. 7.1c) superimposed with cartoon representations of the crystal structure (left) and the 1-state model from **a**. **(d)** Cartoon representations of the 2-state CaM model. DAMMIN and CRY SOL programs used were as implemented in ATSAS 2.8.0 (Franke et al. 2017). PyMOL was used to generate images in figures **c** and **d**

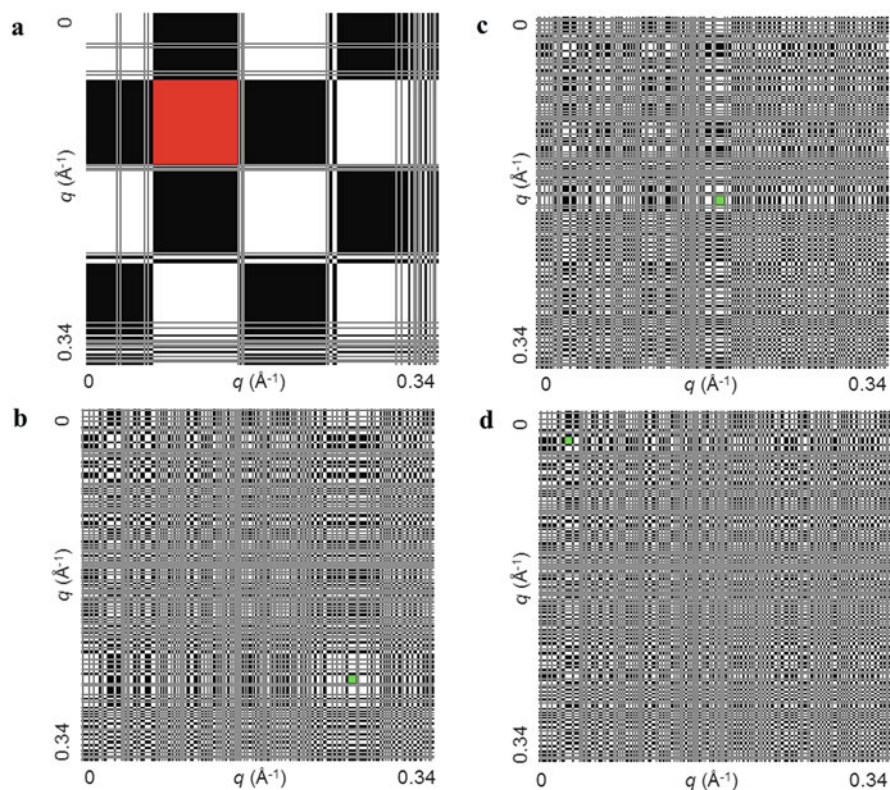
**Table 7.2** Model fitting parameters for calmodulin example

|   |   |
|---|---|
| Shape model fitting   |   |
| Program and parameters  | DAMMIN (default parameters)   |
| $q$ range for fitting ( $\text{\AA}^{-1}$ )                       | 0.007–0.310   |
| Symmetry, anisotropy assumptions                                  | P1  |
| $\chi^2$ , CorMap P values, constant adjustment to intensities    | 0.844, 0.53,<br>$1.877 \times 10^{-4}$                                  |
| Atomistic model fitting   |   |
| <i>From a single coordinate file</i>                              |   |
| Program and parameters  | CRY SOL <sup>a</sup> (default parameters, constant subtraction allowed) |
| Structure coordinates   | PDB:1C LL+ <sup>b</sup>   |
| $q$ -range for all modelling                                      | 0.007–0.310   |
| $\chi^2$ , P-value  | 12.62, 0.00   |
| Predicted $R_g$ ( $\text{\AA}$ )                                  | 22.11   |
| Vol ( $\text{\AA}$ ), Ra ( $\text{\AA}$ ), Dro ( $e/\text{\AA}$ ) | 22012, 1.40, 0.055  |
| <i>Multi-state/ensemble modelling</i>                             |   |
| Program and parameters  | MultiFoXS <sup>c</sup> (10,000 models in starting set)                  |
| Starting coordinates  | PDB:1C LL+  |
| Flexible residues   | 1–3 (ADQ), 77–87 (KDTDS)  |
| Number of states  | 1   |
| $\chi^2$ , CorMap P values  | 0.85, 0.31  |
| $c_1, c_2$  | 1.05, 0.99  |
| $R_g$ values of each state ( $\text{\AA}$ )                       | 21.03   |
| Weights, $w_n$  | 1   |
| Number of states  | 2   |
| $\chi^2$ , CorMap P values  | 0.79, 0.79  |
| $c_1, c_2$  | 1.02, 1.50  |
| $R_g$ values of each state ( $\text{\AA}$ )                       | 22.32, 19.47  |
| Weights, $w_n$  | 0.70, 0.30  |
| Number of states  | 3   |
| $\chi^2$ , CorMap P values  | 0.79, 0.79  |
| $c_1, c_2$  | 1.02, 1.52  |
| $R_g$ values of each state ( $\text{\AA}$ )                       | 22.32, 30.25, 19.00   |
| Weights, $w_n$  | 0.68, 0.13, 0.18  |

<sup>a</sup>In CRY SOL the adjustable parameters are excluded volume (Vol in  $\text{\AA}^3$ ), optimal atomic radius (Ra in  $\text{\AA}$ ) and Dro (optimal contrast of the hydration shell  $e/\text{\AA}^3$ )

<sup>b</sup>PDB:1C LL+ is PDB:1C LL plus the missing ADQ at the N-terminal and C-terminal K missing in the crystal structure

<sup>c</sup>MultiFoXS uses FoXS to calculate model profiles with  $c_1$  and  $c_2$  are the same for all states in a set, the scale factor  $c$  is then optimized for each state and a relative weight  $w_n$  for each state  $n$  is output. The parameters  $c_1$  and  $c_2$  form FoXS the adjustable parameters  $c_1$  and  $c_2$  are adjustments for excluded volume and hydration density.  $c_1$  can vary by 5% (0.95–1.05) and the maximum hydration adjustment  $c_2 = 4.0$  corresponds to  $\sim 0.388$  electrons/ $\text{\AA}^3$  (compared to bulk solvent density  $\rho = 0.334$  electrons/ $\text{\AA}^3$ .)



**Fig. 7.3** 2D correlation maps for SAXS data and CaM models. Models are those in Fig. 7.2. (a) Crystal structure of CaM. (b) Bead model for CaM. (c and d) 1- and 2-state CaM models, respectively. Red or green highlights the longest contiguous set of data points lying one side of the model profile. Red indicates the associated  $P$ -value is  $<0.01$  and below the threshold set for a random distribution of points about the model profile. Green indicates the associated  $P$ -value is  $>0.05$  and above the threshold. CorMaps were calculated using the implementation in ATSAS 2.8.0. (Franke et al. 2017)

(Fig. 7.2b) and 1- and 2-state models (Fig. 7.2c, d, respectively) each have  $P$ -values above the 0.05 threshold. The 2-state model has the highest  $P$ -values (0.85 with just 8 contiguous points falling on one side of the model fit). Comparing the  $P$ - and  $\chi^2$ -values in Table 7.2 there is the expected strong negative correlation (the higher the  $P$ -value, the lower  $\chi^2$ ), but with significantly greater model discrimination in the  $P$ -values. The CorMap analysis is relatively new and experience is needed for it to gain full understanding and broad acceptance. With a smooth model profile, it is possible for a set of contiguous data points to fall very slightly to one side of the model fit and the setting the threshold thus remains somewhat contentious. The significance of a range of  $P$ -values above the threshold for a given data set is not yet calibrated. Nevertheless, CorMap is a very useful complement to  $\chi^2$  and the error weighted

residual plot as it provides a quantitative assessment of the quality of a SAS model fit that is independent of the magnitude of the propagated statistical errors.

### 7.3 Areas for Further Research

An important area of ongoing research in relation to 3D structural modelling of SAS data is predicting  $I(q)$  v  $q$  from a set of atomic coordinates. A significant complication for this calculation arises from the hydration layer surrounding the biomolecule of interest. For SAXS and for SANS measurements in  $H_2O$  the hydration layer contributes significantly to the scattering (Zhang et al. 2012; Kim and Gabel 2015). The effects are largest for SAXS, where the scattering contrast of the hydration layer for a biomolecule in a typical aqueous solvent is similar to that of the protein, thus making the protein appear larger than the atomic coordinates alone would predict. There are a number of approaches to modelling the hydration layer using a uniform density layer approximation or explicit water models (*e.g.* as implemented in *CRY SOL* and *CRY SON* (Svergun et al. 1995; Svergun et al. 1998), *FoXS* (Schneidman-Duhovny et al. 2013), *AQUASAXS* (Poitevin et al. 2011)), and *AXES* (Grishaev et al. 2010), pepsi-SAXS (Grudin et al. 2017)). All of the approaches one way or another effectively add free parameters when fitting the model to experimental data. A number of the developers of these methods have done comparisons of different approaches, however there is no systematic study using high quality experimental data from a set of representative, well-characterized biomolecular systems that could serve for benchmarking. In a recent study, Kim et al. combined SAXS and SANS measurements on mutants of green-fluorescent protein having highly variable net charge to provide evidence of density modifications in the hydration layer that result from the residue-specific attraction of ions from the bulk solvent in combination with structural rearrangements in their vicinity (Kim et al. 2016). Clearly, additional experimental data is needed to fully explore the parameters that affect the hydration layers surrounding biomolecules in solution and their contribution to the total scattering. A comprehensive bench-marking study of the different methods is called for. A potentially fruitful project would be to compare different approaches using an agreed set of exemplar experimental SAS data for a starting set of relatively rigid proteins where high resolution crystal structure coordinates are available. A range of solution conditions would need to be evaluated. A more challenging problem would be to consider highly charged poly-nucleotides and the effects on the scattering profile of the ion cloud they can attract.

Outstanding questions of ongoing research in regard to SAS contributions to hybrid atomistic modelling are a part of the more general questions regarding determining and reporting model uncertainty, accuracy and precision, and how to ascertain that the conformational search space is adequately sampled within the context of the specific set of spatial constraints used (Schneidman-Duhovny et al. 2014). With regard to SAS data, multiple 3D models can fit the same 1D SAS data set. Typically, the question of uniqueness of the model solution has been handled by

performing multiple optimizations of either *ab initio* bead or rigid-body modelling against a SAS profile, or profiles in the case of contrast variation data sets. A cluster analysis can then be used to discriminate potential classes of models and provide some measure of model ambiguity and uncertainty.

If there is an ensemble of conformers present or flexibility, the measured profile represents the population weighted average structure over the measurement period. There are a multitude of multi-domain proteins with flexible linkers and/or hinges that are important for their biological function (e.g. in enzyme catalysis (Henzler-Wildman et al. 2007; Kim et al. 2015), DNA damage signaling and repair (Perry et al. 2010), DNA binding and allosteric signaling (Taraban et al. 2008), mechanical properties in the giant protein muscle protein titin (Improta et al. 1998; Kruger and Kotter 2016), target recognition by CaM (Tidow and Nissen 2013), ubiquitin-mediated regulatory mechanisms (Berndsen and Wolberger 2014; Hershko and Ciechanover 1998). The flexible linkers generally are a challenge for crystallization, and in the crystal form information regarding the solution ensemble is lost. These multi-domain proteins are also most often too large for NMR solution structure techniques and present ambiguous results for microscopy techniques. Given their abundance and the difficulty in characterizing them, ensemble or multi-state modelling against SAS data has been an increasingly popular choice (see reviews (Hammel 2012; Kikhney and Svergun 2015; Rambo and Tainer 2010). However, the problems arising from the limited information content of the SAS profile are many times amplified with the ensemble model. An ensemble model will have many more degrees of freedom than a single 3D model. As a result, ensemble modelling against a SAS profile is much more vulnerable to over-fitting and over-interpretation, even with limits to the conformational space to be sampled within a set of restraints (e.g. knowledge of domain structures, specific flexible regions, contact information from NMR, cross-linking or FRET measurements, etc.).

There are many different approaches to multi-state/ensemble modelling against SAS data, the majority of which optimize by minimizing  $\chi^2$  (e.g. Ensemble Optimization Method EOM (Bernado et al. 2007; Tria et al. 2015), MultiFOXS (Schneidman-Duhovny et al. 2016), and BILBOMD (Pelikan et al. 2009)), ASTEROIDS (Huang et al. 2014), ENSEMBLE (Krzeminski et al. 2013)). Different underlying philosophies are evident in the different methods: e.g. finding the minimal ensemble that fits the data (as in MultiFoXS or BILBOMD), or assuming that flexible regions will sample a continuous distribution of flexible conformations (as in EOM). Other method developers have employed specific strategies to avoid overfitting, e.g. SES (Berlin et al. 2013) uses a linear least squares with a regularization term to obtain a sparse ensemble of conformations, EROS uses a maximum entropy principle as guiding principle to avoid overfitting (Rozycki et al. 2011), BSS-SAXS (Antonov et al. 2016) uses a probabilistic model with Bayesian ensemble inference to model intrinsically ordered proteins. Bayesian methods have seen a recent surge in popularity for ensemble modelling, their appeal being that they seek to limit the solution to the number of conformers that are justified given the model evidence (Potrzebowski et al. [in press](#)). Ongoing research for modelling conformational ensembles requires the collaboration of computational, theoretical

and experimental scientists to consider how much data, what kinds of data, what kinds of representations, what theory and what computational methods need to come together to make progress.

## 7.4 Standards and Publication Guidelines

The increased utilisation of SAS data in hybrid structural modelling to study complexes and assemblies (reviewed in (Schneidman-Duhovny et al. 2012; Vestergaard 2016; Mertens and Svergun 2017; Trehella 2016)) combined with sophisticated software tools designed to be easy to use by non-expert modellers and SAS experimenters makes it imperative to have clear and agreed publication practices with a standard reporting framework and archiving of data in an accessible, searchable data bank.

### 7.4.1 *Establishing Guidelines Through Community Engagement*

Standardization in any field justifiably raises community concerns that there may be unintended consequences that restrict opportunities for publishing. There is also the concern that standards will be too narrow, unreasonable or even misguided. To overcome these natural concerns, there must be ample opportunity for broad community engagement in the process of first developing publication guidelines that can become embedded as standard practice and evolve as the field advances.

The process for developing publication guidelines requires a commitment to two way communication, structured planning, and formal reporting of progress in open access articles. Most importantly there must be leadership from experts across the international community, including providers and developers of instrumentation and analysis tools. Finally, time must be allowed to embed new practices and obtain the resources required to support new norms.

The biomolecular small-angle scattering community has been working toward the establishment of publication guidelines for more than a decade. Supporting the process have been the IUCr through its Commissions for Small-Angle Scattering (CSAS) and Journals (JSAS) and the wwPDB through the establishment of the SASvtf. The meetings of the IUCr Congress and Assembly, as well as the triennial SAS meetings (most recently SAS2012 in Sydney, Australia, SAS2015 in Berlin Germany, and SAS2018 in Traverse City, USA), provided excellent opportunities to report on and ask for community input into the developing recommendations of the CSAS and SASvtf. Commentary pieces made the case for the importance of a community agreed reporting framework for biomolecular SAS (*e.g.* (Jacques et al. 2012)) and there were interim reports outlining preliminary recommended guidelines (Jacques et al. 2012; Trehella et al. 2013).



Most recently, 22 leading SAS experimenters, instrument scientists, SAS analysis program developers as well as experts in crystallography and NMR from around the world, came together to develop a consensus set publication guidelines for biomolecular SAS (Trehella et al. 2017). The guidelines provide a detailed reporting framework that enables readers to “independently assess the quality of the data and the basis for any interpretations presented.” Further, the recommendations were developed to explicitly satisfy recommendation 4 of the 2013 SASvtf report that community agreed “criteria [were] needed for the assessment of the quality of deposited data and the accuracy of SAS-derived models, and the extent to which a given model fits the SAS data” (Trehella et al. 2013). The 2017 guidelines are comprehensive and include recommendations regarding: sample details; data acquisition and reduction; data presentation, analysis and validation; and structure modelling. The reporting guidelines are then applied to a set of example including the CaM example discussed above, where a subset of the reporting framework is used to illustrate essential steps required before choosing a modelling strategy and then approaches to model validation.

### 7.4.2 Archiving SAS Data and Hybrid Models

Recommendations 1–3 of the SASvtf report (Trehella et al. 2013) concerned (1) making SAS data available in a standard format via a searchable and freely accessible archive, (2) developing a dictionary of terms for collecting and managing SAS data, and (3) providing options for depositing SAS-derived models along with specific information on uniqueness and uncertainty, and the protocol used to obtain it.

A SAS data archive requires a standard dictionary of terms with precise definitions enabling the collection and management SAS data. The sasCIF, first established in 2000 (Malfois and Svergun 2000), is an extension of the widely used IUCr Crystallographic Information Framework (CIF). In response to the recommendations of the SASvtf, the sasCIF was further developed and extended as a dictionary that would include experimental information, results and models, including relevant metadata for SAS data analysis and for deposition into a database (Kachala et al. 2016). Importantly, the CIF format is infinitely extensible and as such the sasCIF can be updated to include new terms and definitions, for example from the 2017 guidelines and any future recommended additions. A set of processing tools for sasCIF files has also been developed and made available as standalone open-source programs and integrated into the SAS Biological Data Bank (SASBDB) (<https://www.sasbdb.org/> (Valentini et al. 2015)). These tools enable the export and import of data entries as sasCIF files, thus enabling potential data exchange between SAS databases, *e.g.* between the SASBDB and the data and models held in BIOISIS <http://www.bioisis.net/welcome>.

Recommendations 5 and 6 from the SASvtf report (Trehwella et al. 2013) were that: 5) with the increasing diversity of structural biology data and models being generated, archiving options for models derived from diverse data will be required; and 6) thought leaders from the various structural biology disciplines should jointly define what to archive in the PDB and what complementary archives might be needed (taking into account both scientific needs and funding). In response to these recommendations the Integrative/Hybrid Methods (I/HM) workshop was held in Hinxton (United Kingdom) in October of 2014, bringing together 38 leading structural biologists who came to five consensus recommendations (Sali et al. 2015). These recommendations focused on the importance of being able to archive hybrid/integrative models with complete data and meta-data, a necessarily broader capacity for varied model representations, the importance of providing information regarding what is likely to be variable uncertainty in a given model, agreed model validation tools, and establishing standards for publication of hybrid models. Of particular relevance to the work done to develop the sasCIF, the recommendations included that a “federation of model and data archives should be created” to support the archiving of models derived from hybrid data sets. The sasCIF enables seamless data exchange and interoperation with such a federated system that includes wwPDB. The SAS community is thus well placed to participate in and support this vision for integrative/hybrid methods.

A small but significant step toward the envisioned federated system is a collaborative project between the wwPDB European partner (PDBe) and SASBDB to establish a protocol in the wwPDB OneDep system for hybrid NMR/SAXS structure depositions where the SAS data and meta data are held in the SASBDB and the models in the wwPDB. The co-refinement of SAXS and NMR data is a notable example of hybrid structural modelling. The short-range distance and orientational restraints from NMR combined with the long range distance and translational restraints from SAXS have proven a powerful combination for substantially improving the accuracy of solution NMR structures (Grishaev et al. 2010; Grishaev et al. 2008; Grishaev et al. 2005; Schwieters and Clore 2014; Madl et al. 2011). The combination of SANS and NMR data with crystal structures has been especially powerful in structural modelling of protein RNA assemblies (Lapinaite et al. 2013; Hennig et al. 2014; Gabel 2015). Providing public access to the complete experimental data sets with associated meta-data is essential to the future of hybrid methods structural studies. For the relatively small sized NMR/SAXS structures, many have been deposited in the PDB, but to date the SAXS data were either not included, or included in an *ad hoc* way that makes them difficult to find. A OneDep protocol for hybrid NMR/SAXS structure depositions linked with SASBDB addresses this deficiency.

Addressing the more ambitious hybrid/integrative structural biology challenge of large complexes and assemblies requires bringing together many more disparate data types, new computational methods, visualization tools and yet to be understood tools for model validation. The work done by the biomolecular SAS community to agree publication guidelines with data quality and model validation tools means we are well positioned to participate in this larger vision as part of the wwPDB-

led project that is now developing a prototype model archive system for large-scale structures determined by hybrid methods (Burley et al. 2017).

## 7.5 Conclusion

The SAS experiment is conceptually simple and yet technical demanding. Furthermore, the limited information content in the data can lead to over-interpretation and even mis-interpretation. In many ways, SAS can be most powerful by itself in proving a model inadequate or incomplete (as in comparing the solution and crystal structures for our CaM example). Otherwise, it can be a very powerful restraint in 3D structural modelling when combined with sufficient complementary data. In all applications, SAS data validation requires information beyond what is contained within the scattering profile itself, and validation of the optimal model profile fit and evaluation of model uncertainty and uniqueness require multiple approaches, and even new research.

The kinds of cooperative, volunteer efforts as exemplified by the canSAS working groups, the IUCr CSAS and the wwPDB SASvtf are critically important as the biomolecular SAS field continues to mature and embed standard practices with regard to data and model validation. It is in some ways a fortunate confluence of events and timing that has led to the current state where biomolecular SAS is positioned with many of the tools and guidelines in place to be able to contribute to the developments in hybrid structure determination. This readiness is a reflection of much work and concerted efforts over more than a decade.

**Acknowledgements** The community building and developing of agreed publication guidelines for biomolecular SAS that are an important focus in this chapter was made possible by the collaborative spirit of colleagues who served with me for the past 12 years on the CSAS of the IUCr, the wwPDB SASvtf, and other colleagues around the world who contributed positively to achieve a consensus set of recommended guidelines. I especially wish to acknowledge: J. Mitchell Guss who with his editor's hat on, first suggested to me the importance of writing down some guidelines that would be useful for editors and reviewers dealing with manuscripts containing biomolecular SAS data; David Jacques and Dmitri Svergun who joined Mitchell and I to co-author the preliminary publication guidelines; Wayne Hendrickson, Andrej Sali, Torsten Schwede and John Tainer who joined me in establishing the SASvtf to write the first report that expanded on the preliminary guidelines and recommended the establishment of a SAS data archive; Lois Pollack, Dina Schneidman-Duhovny, Masaaki Sugiyama, and Patrice Vachette who subsequently joined the SASvtf to review and update the preliminary guidelines; Anthony P. Duff, Dominique Durand, Frank Gabel, Greg L. Hura, Nigel M. Kirby, Ann H. Kwan, Javier Pérez, Timothy M. Ryan, John Westbrook, Andrew E. Whitten who joined the effort and contributed to our paper "2017 Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update."

## References

- Antonov LD, Olsson S, Boomsma W, Hamelryck T (2016) Bayesian inference of protein ensembles from SAXS data. *Phys Chem Chem Phys* PCCP 18:5832–5838
- Barbato G, Ikura M, Kay LE, Pastor RW, Bax A (1992) Backbone dynamics of calmodulin studied by  $^{15}\text{N}$  relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. *Biochemistry* 31:5269–5278
- Berlin K, Castaneda CA, Schneidman-Duhovny D, Sali A, Nava-Tudela A, Fushman D (2013) Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *J Am Chem Soc* 135:16595–16609
- Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* 129:5656–5664
- Berndsen CE, Wolberger C (2014) New insights into ubiquitin E3 ligase mechanism. *Nat Struct Mol Biol* 21:301–307
- Bizien T, Durand D, Roblina P, Thureau A, Vachette P, Perez J (2016) A Brief Survey of State-of-the-Art BioSAXS. *Protein Pept Lett* 23:217–231
- Blanchet CE, Spilotros A, Schwemmer F, Graewert MA, Kikhney A, Jeffries CM, Franke D, Mark D, Zengerle R, Cipriani F, Fiedler S, Roessle M, Svergun DI (2015) Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY). *J Appl Crystallogr* 48:431–443
- Brennich ME, Round AR, Hutin S (2017) Online Size-exclusion and Ion-exchange Chromatography on a SAXS Beamline. *J Vis Exp*
- Brookes E, Vachette P, Rocco M, Perez J (2016) US-SOMO HPLC-SAXS module: dealing with capillary fouling and extraction of pure component patterns from poorly resolved SEC-SAXS data. *J Appl Crystallogr* 49:1827–1841
- Burley SK, Kurisu G, Markley JL, Nakamura H, Velankar S, Berman HM, Sali A, Schwede T, Trehwella J (2017) PDB-Dev: a Prototype System for Depositing Integrative/Hybrid Structural Models. *Structure* 25:1317–1318
- David G, Perez J (2009) Combined sampler robot and high-performance liquid chromatography: a fully automated system for biological small-angle X-ray scattering experiments at the Synchrotron SOLEIL SWING beamline. *J Appl Crystallogr* 42:892–900
- Durand D, Vives C, Cannella D, Perez J, Pebay-Peyroula E, Vachette P, Fieschi F (2010) NADPH oxidase activator p67(phox) behaves in solution as a multidomain protein with semi-flexible linkers. *J Struct Biol* 169:45–53
- Fischer H, de Oliveira Neto M, Napolitano HB, Polikarpov I, Craievich AF (2009) The molecular weight of proteins in solution can be determined from a single SAXS measurement on a relative scale. *J Appl Crystallogr* 43:101–109
- Franke D, Jeffries CM, Svergun DI (2015) Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. *Nat Methods* 12:419–422
- Franke D, Petoukhov MV, Konarev PV, Panjkovich A, Tuukkanen A, Mertens HDT, Kikhney AG, Hajizadeh NR, Franklin JM, Jeffries CM, Svergun DI (2017) ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J Appl Crystallogr* 50:1212–1225
- Gabel F (2015) Small-angle Neutron scattering for structural biology of Protein-RNA Complexes. *Methods Enzymol* 558:391–415
- Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the Expasy Server. In: Walker JM (ed) *The proteomics protocols handbook*. Humana Press, New York, pp 571–607
- Glatter O (1977) A new method for the evaluation of small-angle scattering data. *J Appl Crystallogr* 10:307–315

- Graewert MA, Franke D, Jeffries CM, Blanchet CE, Ruskule D, Kuhle K, Flieger A, Schafer B, Tartsch B, Meijers R, Svergun DI (2015) Automated pipeline for purification, biophysical and x-ray analysis of biomacromolecular solutions. *Sci Rep* 5:10734
- Grishaev A, Wu J, Trehwella J, Bax A (2005) Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. *J Am Chem Soc* 127:16621–16628
- Grishaev A, Tugarinov V, Kay LE, Trehwella J, Bax A (2008) Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *J Biomol NMR* 40:95–106
- Grishaev A, Guo L, Irving T, Bax A (2010) Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. *J Am Chem Soc* 132:15484–15486
- Grudin S, Garkavenko M, Kazennov A (2017) Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallogr Sect D Struct Biol* 73:449–464
- Guinier A (1939) La diffraction des rayons x aux très faibles angles: Applications à l'étude de phénomènes ultra-microscopiques. *Ann Phys Paris* 12:161–237
- Guinier A, Fournet G (1955) Small-angle scattering of X-rays. Wiley, New York
- Hammel M (2012) Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS). *Eur Biophys J EBJ* 41:789–799
- Heidorn DB, Trehwella J (1988) Comparison of the crystal and solution structures of calmodulin and troponin C. *Biochemistry* 27:909–915
- Hennig J, Militti C, Popowicz GM, Wang I, Sonntag M, Geerlof A, Gabel F, Gebauer F, Sattler M (2014) Structural basis for the assembly of the Sxl-Unr translation regulatory complex. *Nature* 515:287–290
- Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D (2007) A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* 450:913–916
- Hershko A, Ciechanover A (1998) The ubiquitin system. *Ann Rev Biochem* 67:425–479
- Huang JR, Warner LR, Sanchez C, Gabel F, Madl T, Mackereth CD, Sattler M, Blackledge M (2014) Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. *J Am Chem Soc* 136:7068–7076
- Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL, 2nd, Tsutakawa SE, Jenney FE, Jr, Classen S, Frankel KA, Hopkins RC, Yang SJ, Scott JW, Dillard BD, Adams MW, Tainer JA (2009) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Methods* 6:606–612
- Improta S, Krueger JK, Gautel M, Atkinson RA, Lefevre JF, Moulton S, Trehwella J, Pastore A (1998) The assembly of immunoglobulin-like modules in titin: implications for muscle elasticity. *J Mol Biol* 284:761–777
- Jacques DA, Trehwella J (2010) Small-angle scattering for structural biology—expanding the frontier while avoiding the pitfalls. *Protein Sci* 19:642–657
- Jacques DA, Guss JM, Trehwella J (2012) Reliable structural interpretation of small-angle scattering data from bio-molecules in solution—the importance of quality control and a standard reporting framework. *BMC Struct Biol* 12:9
- Jordan A, Jacques M, Merrick C, Devos J, Forsyth VT, Porcar L, Martel A (2016) SEC-SANS: size exclusion chromatography combined in situ with small-angle neutron scattering. *J Appl Crystallogr* 49:2015–2020
- Kachala M, Westbrook J, Svergun D (2016) Extension of the sasCIF format and its applications for data processing and deposition. *J Appl Crystallogr* 49:302–310
- Kikhney AG, Svergun DI (2015) A practical guide to small angle X-ray Scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* 589:2570–2577

- Kim HS, Gabel F (2015) Uniqueness of models from small-angle scattering data: the impact of a hydration shell and complementary NMR restraints. *Acta Crystallogr Sect D Biol Crystallogr* 71:57–66
- Kim HS, Martel A, Girard E, Moulin M, Hartlein M, Madern D, Blackledge M, Franzetti B, Gabel F (2016) SAXS/SANS on supercharged proteins reveals residue-specific modifications of the hydration shell. *Biophys J* 110:2185–2194
- Kim J, Masterson LR, Cembran A, Verardi R, Shi L, Gao J, Taylor SS, Veglia G (2015) Dysfunctional conformational dynamics of protein kinase A induced by a lethal mutant of phospholamban hinder phosphorylation. *Proc Natl Acad Sci USA* 112:3716–3721
- Koch MH, Vachette P, Svergun DI (2003) Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q Rev Biophys* 36:147–227
- Kratky O (1982) Natural high polymers in the dissolved and solid state. In: Glatter O, Kratky O (eds) *Small-angle X-ray scattering*. Academic, London, pp 361–386
- Kruger M, Kötter S (2016) Titin, a central mediator for hypertrophic signaling, exercise-induced mechanosignaling and skeletal muscle remodeling. *Front Physiol* 7:76
- Krzeminski M, Marsh JA, Neale C, Choy WY, Forman-Kay JD (2013) Characterization of disordered proteins with ENSEMBLE. *Bioinformatics* 29:398–399
- Lapinaite A, Simon B, Skjaerven L, Rakwalska-Bange M, Gabel F, Carlomagno T (2013) The structure of the box C/D enzyme reveals regulation of RNA methylation. *Nature* 502:519–523
- Madl T, Gabel F, Sattler M (2011) NMR and small-angle scattering-based structural analysis of protein complexes in solution. *J Struct Biol* 173:472–482
- Malfois M, Svergun DI (2000) sasCIF: an extension of core Crystallographic Information File for SAS. *J Appl Cryst* 33:812–816
- Mathew E, Mirza A, Menhart N (2004) Liquid-chromatography-coupled SAXS for accurate sizing of aggregating proteins. *J Synchrotron Radiat* 11:314–318
- Mertens HDT, Svergun DI (2017) Combining NMR and small angle X-ray scattering for the study of biomolecular structure and dynamics. *Arch Biochem Biophys* 628:33–41
- Orthaber D, Bergmann A, Glatter O (2000) SAXS experiments on absolute scale with Kratky systems using water as a secondary standard. *J Appl Cryst* 33:218–225
- Pelikan M, Hura GL, Hammel M (2009) Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen Physiol Biophys* 28:174–189
- Perry JJ, Cotner-Gohara E, Ellenberger T, Tainer JA (2010) Structural dynamics in DNA damage signaling and repair. *Curr Opin Struct Biol* 20:283–294
- Poitevin F, Orland H, Doniach S, Koehl P, Delarue M (2011) AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformly hydrated atomic models. *Nucleic Acids Res* 39:W184–W189
- Porod G (1951) Die Röntgenkleinwinkelstreuung von dichtgepackten kolloidalen Systemen. *Kolloid Z* 124:83–114
- Potrzebowski W, Trehwella J, Andre I (in press) Bayesian inference of protein conformational ensembles from limited structural data. *PLOS Comput Biol*
- Rambo RP, Tainer JA (2010) Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle X-ray scattering. *Curr Opin Struct Biol* 20:128–137
- Rambo RP, Tainer JA (2011) Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* 95:559–571
- Round A, Felisaz F, Fodinger L, Gobbo A, Huet J, Villard C, Blanchet CE, Pernot P, McSweeney S, Roessle M, Svergun DI, Cipriani F (2015) BioSAXS Sample Changer: a robotic sample changer for rapid and reliable high-throughput X-ray solution scattering experiments. *Acta Crystallogr Sect D Biol Crystallogr* 71:67–75
- Rozycki B, Kim YC, Hummer G (2011) SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* 19:109–116

- Ryan TM, Trehwella J, Murphy J, Keown JR, Casey L, Pearce FG, Goldstone DC, Chen K, Luo Z, Kobe B, McDevitt CA, Watkin SA, Hawley AM, Mudie ST, Samardzic-Boban V, Kirby N. (2017) An optimized SEC-SAXS system enabling high X-ray dose for rapid SAXS assessment with correlated UV measurements for biomolecular structure analysis. *J Appl Cryst* (in review)
- Sali A, Berman HM, Schwede T, Trehwella J, Kleywegt G, Burley SK, Markley J, Nakamura H, Adams P, Bonvin AM, Chiu W, Peraro MD, Di Maio F, Ferrin TE, Grunewald K, Gutmanas A, Henderson R, Hummer G, Iwasaki K, Johnson G, Lawson CL, Meiler J, Marti-Renom MA, Montelione GT, Nilges M, Nussinov R, Patwardhan A, Rappasilber J, Read RJ, Saibil H, Schroder GF, Schwieters CD, Seidel CA, Svergun D, Topf M, Ulrich EL, Velankar S, Westbrook JD (2015) Outcome of the first wwPDB Hybrid/Integrative methods task force workshop. *Structure* 23:1156–1167
- Schneidman-Duhovny D, Kim SJ, Sali A (2012) Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct Biol* 12:17
- Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2013) Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys J* 105:962–974
- Schneidman-Duhovny D, Pellarin R, Sali A (2014) Uncertainty in integrative structural modeling. *Curr Opin Struct Biol* 28:96–104
- Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2016) FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res* 44:W424–W429
- Schwieters CD, Clore GM (2014) Using small angle solution scattering data in Xplor-NIH structure calculations. *Prog Nucl Magn Reson Spectrosc* 80:1–11
- Sonntag M, Jagtap PKA, Simon B, Appavou MS, Geerlof A, Stehle R, Gabel F, Hennig J, Sattler M (2017) Segmental, domain-selective perdeuteration and small-angle neutron scattering for structural analysis of multi-domain proteins. *Angew Chem Int Ed Engl* 56:9322–9325
- Svergun D, Barberato C, Koch MHJ (1995) CRY SOL - a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr* 28:768–773
- Svergun DI, Richard S, Koch MH, Sayers Z, Zaccai G (1998) Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc Natl Acad Sci USA* 95:2267–2272
- Svergun DI, Koch MHJ, Timmins PA, May RP (2013) Small-angle X-ray and neutron scattering from biological macromolecules. Oxford University Press, Oxford
- Taraban M, Zhan H, Whitten AE, Langley DB, Matthews KS, Swint-Kruse L, Trehwella J (2008) Ligand-induced conformational changes and conformational dynamics in the solution structure of the lactose repressor protein. *J Mol Biol* 376:466–481
- Tidow H, Nissen P (2013) Structural diversity of calmodulin binding to its target sites. *FEBS J* 280:5551–5565
- Trehwella J (2016) Small-angle scattering and 3D structure interpretation. *Curr Opin Struct Biol* 40:1–7
- Trehwella J, Hendrickson WA, Kleywegt GJ, Sali A, Sato M, Schwede T, Svergun DI, Tainer JA, Westbrook J, Berman HM (2013) Report of the wwPDB small-angle scattering task force: data requirements for biomolecular modeling and the PDB. *Structure* 21:875–881
- Trehwella J, Duff AP, Durand D, Gabel F, Guss JM, Hendrickson WA, Hura GL, Jacques DA, Kirby NM, Kwan AH, Pérez J, Pollack L, Ryan TM, Sali A, Schneidman-Duhovny D, Schwede T, Svergun DI, Sugiyama M, Tainer JA, Vachette P, Westbrook J, Whitten AE (2017) 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update. *Acta Crystallogr Sect D Struct Biol* 73:710–728
- Tria G, Mertens HD, Kachala M, Svergun DI (2015) Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* 2:207–217
- Valentini E, Kikhney AG, Previtali G, Jeffries CM, Svergun DI (2015) SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res* 43:D357–D363
- Vestergaard B (2016) Analysis of biostructural changes, dynamics, and interactions - Small-angle X-ray scattering to the rescue. *Arch Biochem Biophys* 602:69–79

- Whitten AE, Cai SZ, Trehella J (2008) MULCh: modules for the analysis of small-angle neutron contrast variation data from biomolecular assemblies. *J Appl Crystallogr* 41:222–226
- Whitten AE, Trehella J (2009) Small-angle scattering and neutron contrast variation for studying bio-molecular complexes. *Methods Mol Biol* 544:307–323
- Zaccai G, Jacrot B (1983) Small angle neutron scattering. *Annu Rev Biophys Bioeng* 12:139–157
- Zaccai NR, Sandlin CW, Hoopes JT, Curtis JE, Fleming PJ, Fleming KG, Krueger S (2016) Deuterium labeling together with contrast variation small-angle neutron scattering suggests how skp captures and releases unfolded outer membrane proteins. *Methods Enzymol* 566:159–210
- Zhang F, Roosen-Runge F, Skoda MW, Jacobs RM, Wolf M, Callow P, Frielinghaus H, Pipich V, Prevost S, Schreiber F (2012) Hydration and interactions in protein solutions containing concentrated electrolytes studied by small-angle scattering. *Phys Chem Chem Phys PCCP* 14:2483–2493



# Chapter 8

## Structural Investigation of Proteins and Protein Complexes by Chemical Cross-Linking/Mass Spectrometry



Christine Piotrowski and Andrea Sinz

**Abstract** During the last two decades, cross-linking combined with mass spectrometry (MS) has evolved as a valuable tool to gain structural insights into proteins and protein assemblies. Structural information is obtained by introducing covalent connections between amino acids that are in spatial proximity in proteins and protein complexes. The distance constraints imposed by the cross-linking reagent provide information on the three-dimensional arrangement of the covalently connected amino acid residues and serve as basis for *de-novo* or homology modeling approaches. As cross-linking/MS allows investigating protein 3D-structures and protein-protein interactions not only *in-vitro*, but also *in-vivo*, it is especially appealing for studying protein systems in their native environment. In this chapter, we describe the principles of cross-linking/MS and illustrate its value for investigating protein 3D-structures and for unraveling protein interaction networks.

**Keywords** Cross-linking · Mass spectrometry · Protein 3D-structure · Protein-protein interactions

### 8.1 Introduction

Proteins play pivotal roles in all biological processes. As the structure of a protein dictates its function, investigating the 3D-structure of a protein and clarifying its interactions with other proteins is one of the most important tasks to elucidate biological processes. While the 3D-structural analysis of proteins is commonly achieved by nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, and cryo-electron microscopy (cryo-EM), protein-protein interactions might be

---

C. Piotrowski · A. Sinz (✉)

Department of Pharmaceutical Chemistry & Bioanalytics, Institute of Pharmacy, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany  
e-mail: [andrea.sinz@pharmazie.uni-halle.de](mailto:andrea.sinz@pharmazie.uni-halle.de)

identified by co-immunoprecipitation or Förster resonance energy transfer (FRET) (Operana and Tukey 2007).

To date, NMR and X-ray crystallography are still the dominating techniques to determine high-resolution protein structures as is indicated by the large number of structures available in the PDB (~120,000 structures obtained by X-ray crystallography versus ~12,000 structures obtained by NMR spectroscopy). Limitations of both high-resolution techniques, however, persist in the investigation of very large and transient protein complexes as well as membrane proteins. Cryo-EM overcomes some of these limitations as structural analysis can be performed at rather low protein concentrations (less than 1  $\mu\text{M}$ ) and highly complex protein assemblies can be targeted (Li et al. 2013).

Cross-linking/MS is an approach that complements the high-resolution 3D-structural techniques and has emerged as promising tool for the structural investigation of proteins and protein complexes in the last years (Young et al. 2000). Especially the combination of cryo-EM with cross-linking-MS has proven beneficial to provide insights into large protein assemblies (Greber et al. 2014; Weisz et al. 2017; Benda et al. 2014) that cannot be obtained by X-ray crystallography or NMR spectroscopy.

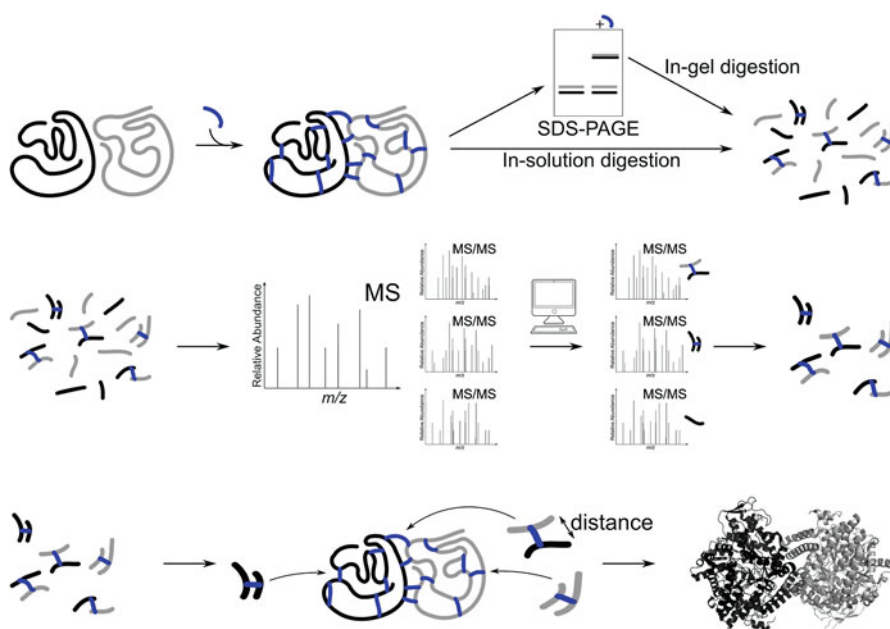
Cross-linking/MS relies on introducing covalent connections between functional groups of amino acid side chains by a chemical reagent. This cross-linker possesses a defined length and connects only these amino acids that are in the appropriate distance to be cross-linked. Usually, the analysis of the cross-linked amino acids is performed in a classical proteomics “*bottom-up*” approach where the cross-linked protein(s) are enzymatically digested and the peptide mixtures are analyzed by liquid chromatography tandem mass spectrometry (LC/MS/MS). This highly sensitive method allows examining as low as femto- to attomole amounts of proteins. The distance constraints that are derived from the cross-linked amino acids are subsequently employed for *de-novo* or homology modeling approaches (Leitner et al. 2016; Rappsilber 2011; Sinz 2014; Walzthoeni et al. 2013; Politis et al. 2014).

Importantly, the cross-linking/MS approach is not only applicable to the 3D-structural analysis of purified proteins, but it also allows elucidating protein-protein interaction networks (Häupl et al. 2016; Schweppe et al. 2017). Protein-protein interaction studies are often based on an affinity enrichment of a tagged bait protein to a specific matrix (Puig et al. 2001; Gavin et al. 2002), together with its interaction partners. Here, the washing procedure applied to remove non-interacting proteins is a crucial step, which however harbors the risk of losing transiently or weakly bound protein interaction partners. Due to the covalent fixation of proteins in the cross-linking/MS approach, the loss of weakly bound proteins during the washing procedure is circumvented.

In this chapter, we give an introduction into the principles of cross-linking/MS and present examples for successful applications of this approach to derive 3D-structural information of proteins and to identify protein interaction networks.

## 8.2 The Cross-Linking/MS Strategy

Once the 3D-structure of a protein or protein complex is covalently fixed by a cross-linking reagent *in-vitro* or *in-vivo*, the identification of the cross-linked amino acids will ultimately give insights into the spatial organization of the protein system under investigation. After the cross-linking reaction, the reaction mixture is analyzed by one-dimensional gel electrophoresis (SDS-PAGE) to visualize the result of the cross-linking reaction and to eventually optimize the reaction conditions (Sinz 2006; Rappsilber 2011) (Fig. 8.1). As mentioned above, the analysis of the cross-linked amino acids is commonly achieved by a “bottom-up” approach, including enzymatic digestion and LC/electrospray ionization (ESI)-MS/MS analysis of the resulting peptide mixture. Proteolysis is realized either by *in-gel* or *in-solution* digestion. Applying the *in-gel* approach, the band containing the protein or protein complex of interest is excised and digested within the gel. Alternatively, proteolytic cleavage can be carried out directly *in-solution* without previous separation of the proteins. The resulting peptide mixture is highly complex as it not only contains



**Fig. 8.1** Cross-linking/MS workflow. A protein or protein complex is stabilized by introducing a covalent bond with a cross-linking reagent. Separation of the cross-linked protein(s) by SDS-PAGE is followed by enzymatic *in-gel* or *in-solution* digestion, resulting in a peptide mixture containing cross-linked and non-cross-linked (linear) peptides. Applying the peptide mixture to MS analysis enables the identification of cross-linked peptides by customized software tools that match MS/MS spectra to potential cross-linking candidates. The cross-links identified provide distance information for modeling protein 3D-structures or for identifying protein interaction partners

non-cross-linked, i.e., linear, and cross-linked peptides of the target proteins, but also peptides of possible contaminants or the protease used for digestion. In subsequent LC/MS/MS analysis, mass spectra and fragment ion mass spectra are recorded, followed by the identification of cross-linked peptides by specific software tools, such as xQuest (Rinner et al. 2008), pLink (Yang et al. 2012), StavroX (Götze et al. 2012a) or Kojak (Hoopmann et al. 2015), that automatically match MS/MS spectra of cross-linked peptides to cross-linking candidates. The cross-links deliver (i) distance information for 3D-structural computational modeling and (ii) insights into the identity of protein interaction partners and protein-protein interaction sites.

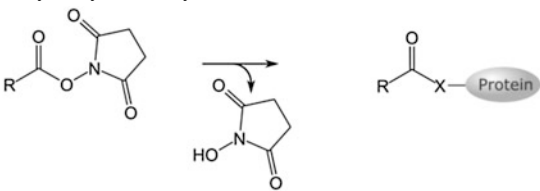
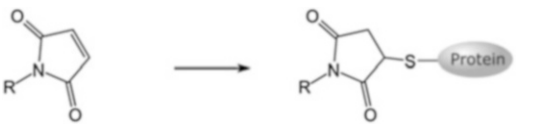
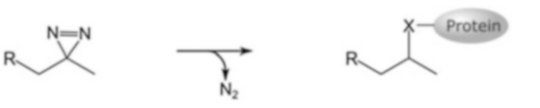
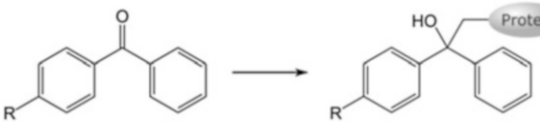
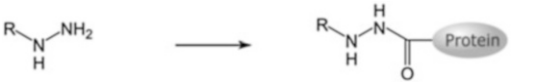
## 8.3 Experimental Design of the Cross-Linking/MS Workflow

### 8.3.1 Cross-Linker Design and Reactivity

The commonly used cross-linkers comprise two reactive head groups that are separated by a spacer with a defined length (Sinz 2006). The spacer determines the distance between the amino acids to be covalently connected and as such serves as “molecular ruler” within a protein or protein complex. Cross-linking reagents are categorized into homobifunctional cross-linkers, comprising identical head groups, or heterobifunctional cross-linkers with non-identical reactive sites (Table 8.1). The most frequently employed reactive groups are *N*-hydroxysuccinimidyl (NHS) esters targeting primary amines in lysines and protein *N*-termini. For NHS esters, an additional reactivity for hydroxy groups in serines, threonines, and tyrosines has been observed (Mädler et al. 2009; Kalkhof and Sinz 2008). Heterobifunctional linkers often contain NHS esters as one of the reactive groups (Hermanson 1996). The second reactive site can be a maleimide, targeting cysteine residues, or a photo-reactive group, such as diazirines or benzophenones. Photo-reactive moieties react in a non-specific manner, potentially connecting all 20 amino acids that are in spatial proximity. For diazirines, a preference for acidic amino acids was observed (Ziemianowicz et al. 2017; Jumper et al. 2012; Iacobucci et al. 2018), while benzophenones target mostly methionines (Wittelsberger et al. 2006). Cross-linking of acidic amino acids is still a challenging task at experimental conditions that do not interfere with the native protein structure. Hydrazines react with aspartic and glutamic acid residues as well as with the *C*-terminus of proteins upon activation with 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) (Novak and Kruppa 2008) or 4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methylmorpholinium chloride (DMTMM) (Leitner et al. 2014a). The activated acidic amino acid might then also react with primary amines of lysines and the *N*-terminus, forming a direct connection between a carboxylic acid and an amine group (Schwarz et al. 2016).

In addition to the cross-linking reagents that are externally introduced (Table 8.1) photo-reactive, unnatural amino acids are available that are directly incorporated into proteins (Suchanek et al. 2005; Piotrowski et al. 2015). These photo-reactive

**Table 8.1** Functional groups in cross-linking reagents

| Reactive group  | Targeted amino acid(s)  |
|---|---|
| <p><i>N</i>-hydroxysuccinimidyl (NHS) ester</p>  | Lysine, <i>N</i> -terminus, serine, threonine, tyrosine X = NH, O   |
| <p>Maleimide</p>                                 | Cysteine  |
| <p>Diazirine</p>                                 | All amino acids, <i>N</i> - and <i>C</i> -terminus, acidic residues (aspartate, glutamate) preferred X = CH <sub>2</sub> , NH, O or S |
| <p>Benzophenone</p>                              | All amino acids, <i>N</i> - and <i>C</i> -terminus, methionine preferred  |
| <p>Hydrazine</p>                                | Acidic amino acids (aspartate, glutamate), <i>C</i> -terminus   |

An overview of commercially available cross-linking reagents is provided at <https://www.thermofisher.com/de/en/home/life-science/protein-biology/protein-labeling-crosslinking/protein-crosslinking.html>

amino acid analogues contain photo-reactive groups, such as benzophenones or diazirines, and they are incorporated into proteins during the translation process in living cells. In general, two strategies are applied, where the photo-reactive amino acids are incorporated into the protein(s) either in a site-specific or a non-directed fashion. The site-specific incorporation of the photo-reactive amino acid makes use of the amber stop codon that is placed at the desired position in the DNA. The amber stop codon encodes for the photo-reactive amino acid, e.g. *para*-benzoylphenylalanin (Bpa) (Ryu and Schultz 2006; Schwarz et al. 2016). To incorporate the photo-reactive amino acid, a specific transfer RNA (tRNA) is needed, which is encoded by an additional plasmid to be transformed or transfected into the cell. The tRNA binds the photo-reactive amino acid and incorporates it into the protein at the amber stop codon position. The major advantage of this approach is that the cross-linking reaction will specifically take place at the desired position




within the Bpa-labeled protein. On the other hand, the non-directed incorporation of the photo-reactive amino acid exploits the translation machinery of the cell to incorporate photo-reactive amino acids. As such, photo-methionine (photo-Met) or photo-leucine (photo-Leu) can be incorporated into proteins by the respective tRNAs for methionine and leucine (Suchanek et al. 2005; Piotrowski et al. 2015; Lössl et al. 2014; Häupl et al. 2017; Iacobucci et al. 2013). Efficient incorporation of photo-Met into proteins has been shown for different cell types (*E. coli*, HEK 293 and HeLa cells) with incorporation rates of 30–35% (Piotrowski et al. 2015). A detailed protocol of a cross-linking approach using the complementary cross-linking principles of BS<sup>2</sup>G and photo-reactive amino acids is provided in (Lössl and Sinz 2016).

### 8.3.2 Identification of Cross-Linked Peptides

As shown in the cross-linking/MS workflow, cross-linked peptides are generated by enzymatic cleavage of the cross-linked proteins by a specific protease (Fig. 8.1). The most prominent protease is trypsin that cleaves proteins C-terminally to basic amino acids (lysine and arginine residues). Cross-linking/MS however differs from the usual proteomics workflow as the use of one single protease is in some cases not sufficient. As two peptides are covalently connected, high molecular weight products are generated exceeding the optimal range of peptide MS detection. Applying a protease additionally to trypsin, such as GluC (cleaving C-terminally to glutamate and aspartate residues), will decrease the molecular weight of cross-linked peptides (Piotrowski et al. 2015). Another approach to generate peptides with lower molecular weight is to conduct proteolysis by an unspecific protease, such as proteinase K (Petrotchenko et al. 2012).

In general, cross-linked peptides are categorized into three different classes as type 0, type 1, and type 2 cross-links (Table 8.2) (Schilling et al. 2003). Type 0 (“dead-end” or “mono-link”) describes a peptide, in which one amino acid is modified by a cross-linker reagent. Here, only one reactive group of the cross-linker has reacted with an amino acid, while the other one has been hydrolyzed or has reacted with the reagent that was used for quenching the cross-linking reaction. “Dead-end” cross-links can deliver insights into the solvent-accessible surface of specific amino acids and as such, give information on the overall topology of the protein under investigation. Type 1 (intrapeptide or “loop-link”) describes the connection of two neighboring amino acids within one peptide. Only

**Table 8.2** Nomenclature of cross-linked peptides

| Type 0  | Type 1  | Type 2  |
|---|---|---|
| “Dead-end”, “mono-link”   | Intrapeptide (“loop”-link)  | Interpeptide  |
|  |  |  |

limited information on the protein's tertiary structure is provided by these cross-links. Type 2 (interpeptide) cross-links connect two peptides originating from one protein or interacting proteins. This class represents the most valuable cross-linked products that yield information on the structural proximity of specific amino acid residues and allow deducing 3D-structural information. According to the systematic nomenclature provided by Schilling et al., the higher molecular weight peptide is termed “ $\alpha$ -peptide” whereas the peptide with the lower molecular weight is referred to as “ $\beta$ -peptide” (Schilling et al. 2003).

The number of software tools for identifying cross-linked peptides from MS data is steadily increasing and not easy to review. Table 8.3 provides an overview comprising several of the so far developed software applications. The general workflow of these software tools includes an *in-silico* digestion of proteins. Subsequently, potential cross-link candidates are automatically compared to the recorded MS/MS spectra and matching cross-link candidates are reported. In addition to the software tools available for cross-link identification, an increasing number of software applications is available to further examine the cross-links, such as xVis (Grimm et al. 2015) or Xlink-DB (Zheng et al. 2013). These applications provide a visualization of the identified cross-links either in a schematic fashion

**Table 8.3** Selected software tools for identifying cross-linked peptides

| Software                                | References  |
|---|---|
| CLPM                                    | Tang et al. (2005)  |
| Crux                                    | McIlwain et al. (2014)  |
| DXMSMS                                  | Petrotchenko et al. (2014)  |
| ECL/ECL2                                | Yu et al. (2016, 2017)  |
| FINDX                                   | Soderberg et al. (2012)   |
| Kojak                                   | Hoopmann et al. (2015)  |
| MassAI (CrossWork)                      | Rasmussen et al. (2011)   |
| MassMatrix                              | Xu et al. (2008)  |
| MassSpecStudio                          | Sarpe et al. (2016)   |
| MS-bridge (included in USCF prospector) | <a href="http://prospector.ucsf.edu">http://prospector.ucsf.edu</a> |
| PeptideMap                              | Fenyo (1997)  |
| pLink                                   | Yang et al. (2012)  |
| Pro-cross-link                          | Gao et al. (2006)   |
| ProteinXXX (included in GPMAW)          | Nielsen et al. (2007)   |
| SIM-XL                                  | Lima et al. (2015)  |
| StavroX/MeroX                           | Götze et al. (2012b, 2015)  |
| Xi                                      | Fischer et al. (2013) and Giese et al. (2016b)                      |
| Xilmass                                 | Yilmaz et al. (2016)  |
| Xlink analyzer                          | Kosinski et al. (2015)  |
| Xlink-identifier                        | Du et al. (2011)  |
| XlinkX/XlinkX 2.0                       | Liu et al. (2015) and Liu et al. (2017)                             |
| XLPM                                    | Jaiswal et al. (2014)   |
| xComb                                   | Panchaud et al. (2010)  |
| xQuest                                  | Rinner et al. (2008)  |

or by mapping them into published PDB structures. Also, there are software tools available for quantifying cross-linked peptides, e.g. xTract (Walzthoeni et al. 2015) or XiQ (Fischer et al. 2013).

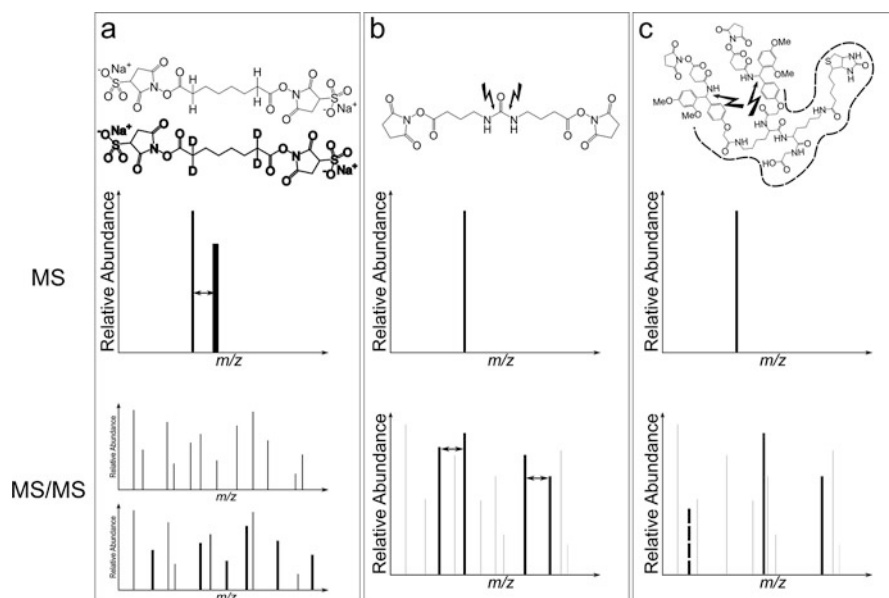
### 8.3.3 *Facilitated Analysis of Cross-Linked Products*

The identification of cross-linked products is still challenging due to their low abundance in the peptide mixtures generated after enzymatic digestion (Fig. 8.1). Specifically, we see two main difficulties: (i) Cross-linked peptides might be missed during MS analysis, (ii) false-positive identifications of cross-links might occur due to their great variability. To overcome these problems, sample complexity can be reduced or cross-linked products can be enriched. Alternatively, cross-linkers carry an isotope signature (usually by introducing deuterium atoms) or create specific fragment ion patterns during MS/MS experiments for automated data analysis.

Strong cation exchange (SCX) and size-exclusion chromatography (SEC) are the methods of choice to enrich cross-linked peptides (Leitner et al. 2010, 2014b; Schmidt and Sinz 2017). Also, an affinity enrichment of cross-linked products is based on a biotin tag that specifically binds to avidin. Biotin is either incorporated in the cross-linking reagent (Tang and Bruce 2010) or is introduced after the cross-linking reaction by click chemistry (Nury et al. 2015). In order to utilize SCX enrichment of cross-linked peptides, enzymatic cleavage has to be performed with a protease cleaving C-terminally to basic amino acids, such as trypsin. Consequently, every peptide carries a positive charge at the C-terminus that can be utilized for SCX enrichment. As cross-linked products are composed of two peptides, they accommodate a higher number of positive charges than linear peptides. Thus, the cross-linked peptides can be enriched by SCX (Leitner et al. 2010; Fritzsche et al. 2012; Schmidt and Sinz 2017; Tinnefeld et al. 2017). As cross-linked peptides usually possess higher molecular weights than non-cross-linked peptides, they can also be enrichment by SEC (Herzog et al. 2012; Rampler et al. 2015). A major advantage of SEC is its applicability for all peptide mixtures, independent of the protease used for digestion, but on the other hand, low-molecular weight cross-linked peptides might get lost during SEC enrichment.

For an unambiguous identification of cross-linked products, cross-linking reagents with unique characteristics have been designed. The first class of cross-linkers contains isotope labels, in most cases deuterium atoms, such as bis(sulfosuccinimidyl)suberate (BS<sup>3</sup>) D<sub>0</sub>/D<sub>4</sub> (Fig. 8.2a) (Müller et al. 2001; Schmidt et al. 2005). The deuterated and non-deuterated version of the cross-linker are mixed in a 1:1 ratio and added to the protein solution. Hence, every cross-linked product is visible in mass spectra as a specific doublet of signals. Both species generate nearly identical MS/MS spectra, but differ in the fragment ions containing the deuterated or non-deuterated cross-linker. Acquiring MS/MS spectra from both isotope species helps in unambiguously identifying cross-linked peptides as two spectra with the specific mass shift are only obtained for cross-linked products, but not for linear peptides.





**Fig. 8.2** Strategies to facilitate the analysis of cross-linked peptides. Structures of respective cross-linkers are presented in the upper panel. Specific cleavage sites of the cross-linkers upon collisional activation inside the mass spectrometer are indicated. The middle panel (MS) displays how the cross-linked peptides appear in the mass spectrum, while the lower panel (MS/MS) shows characteristics of the cross-linked products in fragment ion mass spectra. (a) For the isotope-labeled cross-linker BS<sup>3</sup>(D<sub>0</sub>/D<sub>4</sub>), two signals are detected in the mass spectrum for one cross-linked peptide pair differing by the isotope label (4 amu). Subsequently, two MS/MS spectra for one cross-linked peptide pair are recorded, black – BS<sup>3</sup>D<sub>0</sub>; bold – BS<sup>3</sup>D<sub>4</sub>. (b) For the MS-cleavable cross-linker DSBU, characteristic fragment ion signatures of the linker (two doublets) are visible in MS/MS spectra. (c) PIR cross-linkers exhibit characteristic fragment ions (dashed line) and peptides modified with cross-linker fragments (bold lines) in MS/MS spectra

A highly attractive approach that is currently gaining more and more importance is employing cross-linkers with an MS-cleavable moiety (Sinz 2017). The cross-linker is cleaved during collisional activation in the gas phase inside the mass spectrometer resulting in specific fragment ions that contain parts of the cross-linker. In Fig. 8.2b, the MS-cleavable cross-linker disuccinimidyl dibutyric urea (DSBU, formerly BuUrBu) is presented. DSBU comprises two NHS esters as reactive sites and a cleavable urea moiety (Müller et al. 2010). After fragmentation of the linker, two doublet signals are visible for each interpeptide (type 2) cross-linked product in the MS/MS spectrum. These doublets result from cleavage of one of the two NH–CO bonds of the urea moiety. Thus, two pairs of asymmetric fragments are generated, exhibiting a specific mass difference of 25.979 amu that allows an unambiguous identification of a cross-linked product.

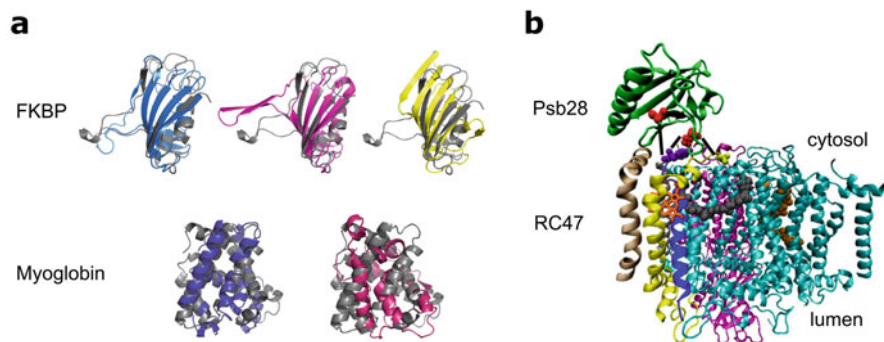
There are other MS-cleavable cross-linkers available that generate characteristic fragment ions, such as a class of reagents termed “protein interaction reporters

(PIR)” (Fig. 8.2c) (Tang and Bruce 2010). PIR cross-linkers comprise two cleavage sites releasing a specific part of the cross-linker after collisional activation inside the mass spectrometer. This specific fragment ion as well as peptides containing the remaining fragment of the cross-linker are present in every MS/MS spectrum of a cross-linked product and allow its unambiguous assignment.

## 8.4 Structural Investigation of Purified Proteins and Large Protein Assemblies

As outlined above, the distance constraints derived from the cross-linked amino acids serve as basis for a computational modeling of purified proteins or protein complexes. The distance constraints are provided as  $C_{\alpha}$ - $C_{\alpha}$  or  $C_{\beta}$ - $C_{\beta}$  distances and defined by the length of the side chains of the cross-linked residues plus the cross-linker spacer length (Merkley et al. 2014; Hofmann et al. 2015). In principle, two strategies are employed to implement these constraints into the modeling process: (i) modeling of protein structures using the cross-linking distances as input, (ii) filtering the theoretical models by the experimentally determined cross-linking distances. A number of different software tools are available for modeling protein structures, such as ROSETTA (Kaufmann et al. 2010), I-TASSER (Zhang 2009), PEP-FOLD (Maupetit et al. 2009) or Abalone (<http://www.biomolecular-modeling.com/Abalone/index.html>). The optimal cross-linker spacer lengths for protein modeling have been evaluated by combining and analyzing simulated and experimentally observed cross-linking constraints for various proteins (Hofmann et al. 2015). A specific equation was developed to predict the ideal spacer length in correlation to the size of the targeted protein. As a result, the optimal spacer length of a cross-linker to study the 35-kDa human phosphatase activator protein (PDB-ID: 2IXM) was determined to be 12.5 Å (Hofmann et al. 2015).

The investigation of a purified protein aims at elucidating its tertiary structure and the assembly state of eventually present oligomeric forms. Furthermore, newly developed cross-linking reagents are usually evaluated using small proteins, such as myoglobin, human serum albumin or bovine serum albumin (Brodie et al. 2017; Belsom et al. 2016, 2017; Giese et al. 2016a; Iacobucci et al. 2017). Small proteins, such as myoglobin and the FK506 binding protein (FKBP), have also been used for studying the impact of cross-linking applied to structural modeling with a discrete molecular dynamics (DMD) simulation based on cross-linking constraints (Brodie et al. 2017). Five short-range cross-linkers with various functional groups targeting different amino acids were employed to derive distance constraints of both proteins. The cross-linking data obtained were subsequently used as input for the DMD. Clustering of the generated models identified three clusters for FKBP and two for myoglobin. Representative model structures of each cluster were similar to the known PDB structures (Fig. 8.3a), underlining the strength of the cross-linking/MS approach to obtain native-like conformations.



**Fig. 8.3** Structural investigation of single proteins and a large protein complex. **(a)** Comparison of cross-link-based modeled structures to available PDB structures of FK506 binding protein (FKBP) and myoglobin. The best scored structures of the largest clusters are superimposed with the PDB structures (dark grey). Figure is adapted with permission from (Brodie et al. 2017). **(b)** Structure of Psb28 docked to the RC47 subcomplex of the photosystem II of the cyanobacterium *Synechocystis* sp. (Figure is adapted with permission from Weisz et al. 2017)

Applying cross-linking to large protein complexes illustrates that the size of the complexes of interest is unrestricted for cross-linking/MS approaches, while NMR or X-ray crystallography are limited in protein size by sample preparation and data acquisition. Cross-linking/MS is able to deliver structural information on small protein assemblies, such as nidogen-1/laminin  $\gamma$ 1 (Lössl et al. 2014) or chaperone Hsc70/ $\alpha$ -synuclein complexes (Nury et al. 2015), but also on large protein systems, such as the mammalian mitochondrial ribosome (Greber et al. 2014) or the ribosome post-recycling complex (Kiosze-Becker et al. 2016).

To simplify the study of protein-protein interactions, representative peptides can be employed that harbor known interaction sites, as predicted by preceding biochemical studies or computational approaches. As an example, Munc13 peptides containing the respective calmodulin (CaM) binding site were synthesized to investigate presynaptic CaM/Munc13 complexes (Dimova et al. 2009; Lipstein et al. 2012). For this, the unnatural photo-reactive amino acid Bpa was incorporated during peptide synthesis into Munc13 peptides and the peptides were applied for photo-cross-linking experiments. Additionally, the heterobifunctional cross-linker *N*-succinimidyl-*p*-benzoyldihydrocinamate (SBC), containing a NHS ester and a benzophenone group, as well as BS<sup>3</sup> and bis(sulfosuccinimidyl)-2,2,4,4-glutarate (BS<sup>2</sup>G) were applied to obtain complementary cross-linking data on the interaction between CaM and Munc13 peptides. The cross-linking constraints were then subjected to computational modeling of the CaM/Munc13 peptide complexes using the PatchDock and ROSETTADock software applications. The resulting structures revealed all Munc13 isoforms to bind similarly to CaM, indicating a common CaM-binding motif of all four Munc13 isoforms (Dimova et al. 2009; Lipstein et al. 2012).

Photosystem II intermediate complexes from cyanobacterium *Synechocystis* sp present an impressive example for applying cross-linking/MS to large protein

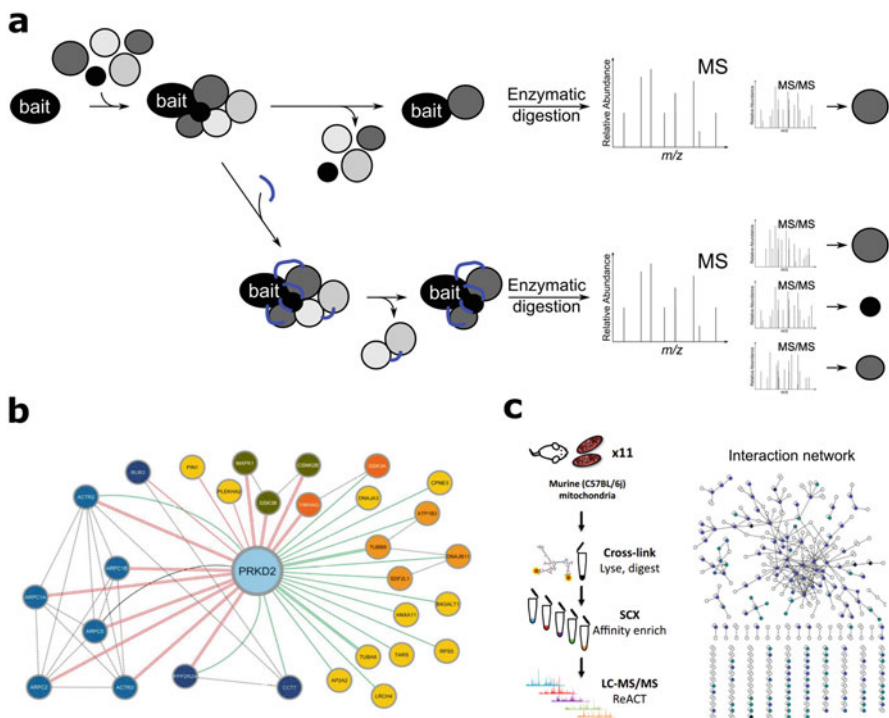
assemblies (Weisz et al. 2017). To study the interaction of the cytoplasmic photosystem II protein Psb28 with the membrane bound photosystem II component, complete photosystem II was purified from *Synechocystis sp.* Structural analysis was carried out with the homobifunctional, amine-reactive cross-linking reagent BS<sup>3</sup> ( $D_0/D_{12}$ ). The cross-linking data revealed an interaction of Psb28 with cytochrome b559, PsbE, and PsbF subunits of the photosystem II included in the RC47 subcomplex. Subsequent docking of Psb28 to the RC47 subcomplex of the photosystem II, which is known to interact with Psb28, was performed with the DOT 2.0 docking software. The resulting structures were validated by comparing the cross-linking data with the generated models. The final model of the protein complex is displayed in Fig. 8.3b, showing that Psb28 is interacting with PsbE and PsbF of the RC47 subcomplex as confirmed by the cross-linking data.

Currently, the majority of cross-linking studies is performed *in-vitro* as these studies allow a specific targeting of the proteins of interest. Deriving structural information on protein and protein complexes *in-vivo* is still a daunting task due to the enormous complexity of cellular samples. In many studies, cross-linking experiments are combined with immunoblotting. Examples include studies of the COX-2/mPGES complex and TS3-regulating proteins LcrG and LcrV (Henderson and Nilles 2017) as well as investigating the assembly state of  $\alpha$ -synuclein (Corbille et al. 2016). The analysis of  $\alpha$ -synuclein assemblies in living cells was performed by disuccinimidylglutarate (DSG) and dithiobis(succinimidyl)propionate (DSP). Both cross-linkers are NHS esters targeting amine groups in proteins, but they differ in spacer length. Additionally, the cross-linker DSP contains a disulfide bond that can be cleaved under reducing conditions (Corbille et al. 2016). Cross-linking was induced by adding the cross-linkers directly to the cell suspension, followed by disruption of the cells. Analysis of the complexes was then performed by immunoblotting using an anti- $\alpha$ -synuclein antibody revealing the presence of  $\alpha$ -synuclein dimers and pentamers in the cells. This result was verified by the cleavable DSP cross-linker as the pentamer disappeared after reduction of the disulfide bond in DSP.

## 8.5 Identification of Protein-Protein Interaction Networks

The identification of protein-protein interaction networks is the key to understanding biological processes and cross-linking/MS can make here major contributions by identifying interacting proteins as well as defining their interaction sites. Often, the result of the cross-linking reaction is monitored via SDS-PAGE or immunoblotting, giving insights only into these interaction partners for which antibodies are available (Hetu et al. 2008; Maadi et al. 2017; Henderson and Nilles 2017). The combination of cross-linking with MS will give more detailed insights, giving a more comprehensive picture on protein interaction networks.

*In-vitro* MS based procedures usually include an affinity-based identification of protein interaction partners. The starting point for these studies is the immobilization



**Fig. 8.4** Elucidation of protein-protein interaction networks. **(a)** The immobilized bait protein is incubated with a cell lysate, followed by a washing procedure to remove the non-binding proteins. LC/MS/MS analysis identifies the interacting proteins. Upper panel: common affinity-based strategy, lower panel: cross-linking-based strategy. **(b)** Identified interaction partners of protein kinase D2 applying BS<sup>2</sup>G to capture interacting proteins by the strategy presented in a) lower panel. Figure is adapted with permission from (Häupl et al. 2016). **(c)** Workflow for the identification of protein-protein interaction network from murine mitochondria. Circles display proteins, lines indicate cross-links. The color depth of each dot is proportional to the frequency of the respective protein within the eleven samples. (Figure is adapted with permission from Schweppe et al. 2017)

of a bait protein on a matrix, followed by incubation with cell lysates or cellular fractions. Several washing steps are performed to remove non-interacting proteins, followed by enzymatic digestion to identify protein binding partners (Fig. 8.4a, upper panel). Unfortunately, the washing procedure can prevent the detection of weakly or transiently bound protein-protein interaction. Applying a cross-linking reagent for a covalent fixation of interacting proteins prior to the washing procedure prevents losing potential protein binding partners (Fig. 8.4a, lower panel). Afterwards, the enriched interacting proteins are enzymatically digested and analyzed by LC/MS/MS. Cross-linked peptides as well as non-cross-linked peptides are used to identify the interaction partners. Non-cross-linked peptides identify the binding proteins, while cross-linked peptides additionally yield information on the protein interfaces.

A combined affinity purification cross-linking/MS strategy was applied to identify partners of protein kinase D2 (PKD2) from Golgi preparations and whole cell lysates. For these studies, the external cross-linker BS<sup>2</sup>G (*D*<sub>0</sub>/*D*<sub>4</sub>) as well as the unnatural amino acids, photo-Leu and photo-Met, incorporated into proteins during translation in HeLa cells, were applied (Häupl et al. 2016, 2017). To investigate PKD2 interaction partners, glutathione-S-transferase (GST)-tagged PKD2 was immobilized on GSH sepharose beads and incubated with cell lysate or a Golgi preparation. PKD2-interacting proteins were covalently bound by adding the amine-reactive cross-linker BS<sup>2</sup>G (*D*<sub>0</sub>/*D*<sub>4</sub>) to the reaction mixture or by inducing photo-cross-linking of the unnatural amino acids by UV-A irradiation. LC/MS/MS analysis allowed identifying the covalently fixed PKD2 interaction partners. The results obtained by the BS<sup>2</sup>G cross-linking are illustrated in Fig. 8.4b (Häupl et al. 2016). With the photo-reactive amino acids, similar PKD2 interaction partners were identified, but the complementary reactivity and shorter spacer length of the photo-reactive amino revealed additional interacting proteins (Häupl et al. 2017). A similar approach, exclusively based on BS<sup>2</sup>G (*D*<sub>0</sub>/*D*<sub>4</sub>), was employed for investigating protein interaction partners of tissue-type plasminogen activator (t-PA), an established tumor marker in various cancers (Bosse et al. 2016). Proteins secreted by erlotinib-sensitive (PC9) and erlotinib-resistant (PC9ER) non-small cell lung cancer (NSCLC) cells were investigated, indicating differences of t-PA interacting proteins between erlotinib-sensitive and -resistant cells.

To enable protein interaction partner studies *in-vivo*, cross-linking reagents are added to cell cultures or cell suspensions, which are crossing the cell membrane to react with target proteins within the cell (Weisbrod et al. 2013; de Jong et al. 2017). In case of photo-reactive amino acid incorporation, the reactive groups enabling cross-linking reactions are already incorporated during cell growth. Desired cross-linking is afterwards induced by exposure of the cells to UV-A light (Yang et al. 2016a). Subsequent proteolysis of whole cells or cell lysates results in enormously complex peptide mixtures that hamper a thorough identification of cross-linked peptides. Consequently, an enrichment of cross-linked peptides is mandatory before performing LC/MS/MS analyses. Affinity strategies can be applied if the protein of interest contains a tag for the specific isolation of the desired protein complexes (Walker-Gray et al. 2017). Alternatively, the biotin label is contained in the cross-linker. As described above, the biotin label can either be incorporated in the cross-linker itself (Tang and Bruce 2010; Tan et al. 2016; Yang et al. 2016b) or it is added after to the cross-linking reaction by a click-reaction (Nury et al. 2015). The latter approach is based on orthogonal chemistry strategies developed for proteomic analyses (Speers and Cravatt 2005; Weerapana et al. 2007).

Abovementioned PIR cross-linkers (Fig. 8.2c) have been used for *in-vivo* studies identifying protein-protein interactions in mitochondria – cell organelles that are comprised of more than 1000 proteins (Schweppe et al. 2017). Cross-linking experiments were conducted on active mitochondria isolated from mouse heart and allowed the identification of protein interaction partners as well as the 3D-structural investigation of the respective protein complexes. The PIR cross-linker used for this study was membrane-permeable, and comprised two NHS esters as reactive

head groups as well as a biotin group for an enrichment of cross-links (Fig. 8.4c). After the cross-linking reaction, mitochondria were disrupted and the proteins were enzymatically digested. Cross-linked peptides were first fractionated by SCX and further enriched by affinity chromatography. Subsequent LC/MS/MS analysis by identified 327 proteins and 2427 cross-linked peptides, which additionally allowed gaining insights into the 3D-structures of selected protein complexes.

## 8.6 Conclusion

Cross-linking/MS has matured as a valuable technique in structural biology that complements existing techniques, such as X-ray crystallography, NMR spectroscopy, and cryo-EM. The major applications of the cross-linking/MS approach are to derive 3D-structural information of purified proteins and protein complexes, providing distance information for computational modeling studies, and to elucidate protein-protein interaction networks from cell lysates or even in intact cells. Although cross-linking of proteins *in-vivo* is still a challenging task, innovative approaches have been developed and are continuously being improved. A comprehensive analysis of protein-protein interaction networks in the cellular environment has become feasible.

**Acknowledgments** AS acknowledges financial support by the DFG (project Si 867/15-2) and the region of Saxony-Anhalt.

## References

- Belsom A, Schneider M, Fischer L, Brock O, Rappsilber J (2016) Serum albumin domain structures in human blood serum by mass spectrometry and computational biology. *Mol Cell Proteomics* 15(3):1105–1116. <https://doi.org/10.1074/mcp.M115.048504>
- Belsom A, Mudd G, Giese S, Auer M, Rappsilber J (2017) Complementary benzophenone cross-linking/mass spectrometry photochemistry. *Anal Chem* 89(10):5319–5324. <https://doi.org/10.1021/acs.analchem.6b04938>
- Benda C, Ebert J, Scheltema RA, Schiller HB, Baumgartner M, Bonneau F, Mann M, Conti E (2014) Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. *Mol Cell* 56(1):43–54. <https://doi.org/10.1016/j.molcel.2014.09.002>
- Bosse K, Haneder S, Arlt C, Ihling CH, Seufferlein T, Sinz A (2016) Mass spectrometry-based secretome analysis of non-small cell lung cancer cell lines. *Proteomics* 16(21):2801–2814. <https://doi.org/10.1002/pmic.201600297>
- Brodie NI, Popov KI, Petrotchenko EV, Dokholyan NV, Borchers CH (2017) Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations. *Sci Adv* 3(7):e1700479. <https://doi.org/10.1126/sciadv.1700479>
- Corbille AG, Neunlist M, Derkinderen P (2016) Cross-linking for the analysis of alpha-synuclein in the enteric nervous system. *J Neurochem* 139(5):839–847. <https://doi.org/10.1111/jnc.13845>
- de Jong L, de Koning EA, Roseboom W, Buncherd H, Wanner MJ, Dapic I, Jansen PJ, van Maarseveen JH, Corthals GL, Lewis PJ, Hamoen LW, de Koster CG (2017) In-culture cross-linking of bacterial cells reveals large-scale dynamic protein-protein interactions at the peptide level. *J Proteome Res* 16(7):2457–2471. <https://doi.org/10.1021/acs.jproteome.7b00068>

- Dimova K, Kalkhof S, Pottratz I, Ihling C, Rodriguez-Castaneda F, Liepold T, Griesinger C, Brose N, Sinz A, Jahn O (2009) Structural insights into the calmodulin-Munc13 interaction obtained by cross-linking and mass spectrometry. *Biochemistry* 48(25):5908–5921. <https://doi.org/10.1021/bi900300r>
- Du X, Chowdhury SM, Manes NP, Wu S, Mayer MU, Adkins JN, Anderson GA, Smith RD (2011) Xlink-identifier: an automated data analysis platform for confident identifications of chemically cross-linked peptides using tandem mass spectrometry. *J Proteome Res* 10(3):923–931. <https://doi.org/10.1021/pr100848a>
- Fenyo D (1997) A software tool for the analysis of mass spectrometric disulfide mapping experiments. *Comp Appl Biosci* 13(6):617–618
- Fischer L, Chen ZA, Rappsilber J (2013) Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *J Proteome* 88:120–128. <https://doi.org/10.1016/j.jprot.2013.03.005>
- Fritzsche R, Ihling CH, Gotze M, Sinz A (2012) Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis. *Rap Commun Mass Spectrom* 26(6):653–658. <https://doi.org/10.1002/Rcm.6150>
- Gao Q, Xue S, Doneanu CE, Shaffer SA, Goodlett DR, Nelson SD (2006) Pro-CrossLink. Software tool for protein cross-linking and mass spectrometry. *Anal Chem* 78(7):2145–2149. <https://doi.org/10.1021/ac051339c>
- Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelman A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415(6868):141–147. <https://doi.org/10.1038/415141a>
- Giese SH, Belsom A, Rappsilber J (2016a) Optimized fragmentation regime for Diazirine photo-cross-linked peptides. *Anal Chem* 88(16):8239–8247. <https://doi.org/10.1021/acs.analchem.6b02082>
- Giese SH, Fischer L, Rappsilber J (2016b) A study into the collision-induced dissociation (CID) behavior of cross-linked peptides. *Mol Cell Proteomics* 15(3):1094–1104. <https://doi.org/10.1074/mcp.M115.049296>
- Götze M, Pettelkau J, Schaks S, Bosse K, Ihling CH, Krauth F, Fritzsche R, Kuhn U, Sinz A (2012a) StavroX—a software for analyzing crosslinked products in protein interaction studies. *J Am Soc Mass Spectrom* 23(1):76–87. <https://doi.org/10.1007/s13361-011-0261-2>
- Götze M, Pettelkau J, Schaks S, Bosse K, Ihling CH, Krauth F, Fritzsche R, Kühn U, Sinz A (2012b) StavroX-A software for analyzing crosslinked products in protein interaction studies. *J Am Soc Mass Spectrom* 23(1):76–87. <https://doi.org/10.1007/s13361-011-0261-2>
- Götze M, Pettelkau J, Fritzsche R, Ihling CH, Schafer M, Sinz A (2015) Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis. *J Am Soc Mass Spectrom* 26(1):83–97. <https://doi.org/10.1007/s13361-014-1001-1>
- Greber BJ, Boehringer D, Leitner A, Bieri P, Voigts-Hoffmann F, Erzberger JP, Leibundgut M, Aebersold R, Ban N (2014) Architecture of the large subunit of the mammalian mitochondrial ribosome. *Nature* 505(7484):515–519. <https://doi.org/10.1038/nature12890>
- Grimm M, Zimniak T, Kahraman A, Herzog F (2015) xVis: a web server for the schematic visualization and interpretation of crosslink-derived spatial restraints. *Nucleic Acids Res* 43(W1):W362–W369. <https://doi.org/10.1093/nar/gkv463>
- Häupl B, Ihling CH, Sinz A (2016) Protein interaction network of human protein kinase D2 revealed by chemical cross-linking/mass spectrometry. *J Proteome Res* 15:3686–3699. <https://doi.org/10.1021/acs.jproteome.6b00513>
- Häupl B, Ihling CH, Sinz A (2017) Combining affinity enrichment, cross-linking with photo-amino acids, and mass spectrometry for probing protein kinase D2 interactions. *Proteomics*. <https://doi.org/10.1002/pmic.201600459>



- Henderson TA, Nilles ML (2017) In vivo photo-cross-linking to study T3S interactions demonstrated using the *Yersinia pestis* T3S system. *Methods Mol Biol* 1531:47–60. [https://doi.org/10.1007/978-1-4939-6649-3\\_4](https://doi.org/10.1007/978-1-4939-6649-3_4)
- Hermanson GT (1996) Bioconjugate techniques. Academic, San Diego
- Herzog F, Kahraman A, Boehringer D, Mak R, Bracher A, Walzthoeni T, Leitner A, Beck M, Hartl FU, Ban N, Malmstrom L, Aebersold R (2012) Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science* 337(6100):1348–1352. <https://doi.org/10.1126/science.1221483>
- Hetu PO, Ouellet M, Falguyet JP, Ramachandran C, Robichaud J, Zamboni R, Riendeau D (2008) Photo-crosslinking of proteins in intact cells reveals a dimeric structure of cyclooxygenase-2 and an inhibitor-sensitive oligomeric structure of microsomal prostaglandin E2 synthase-1. *Arch Biochem Biophys* 477(1):155–162. <https://doi.org/10.1016/j.abb.2008.04.038>
- Hofmann T, Fischer AW, Meiler J, Kalkhof S (2015) Protein structure prediction guided by crosslinking restraints – a systematic evaluation of the impact of the crosslinking spacer length. *Methods* 89:79–90. <https://doi.org/10.1016/j.ymeth.2015.05.014>
- Hoopmann MR, Zelter A, Johnson RS, Riffle M, MacCoss MJ, Davis TN, Moritz RL (2015) Kojak: efficient analysis of chemically cross-linked protein complexes. *J Proteome Res* 14(5):2190–2198. <https://doi.org/10.1021/pr501321h>
- Iacobucci C, Reale S, De Angelis F (2013) Photoactivable amino acid bioisosteres and mass spectrometry: snapshots of in vivo 3D protein structures. *Chembiochem Eur J Chem Biol* 14(2):181–183. <https://doi.org/10.1002/cbic.201200742>
- Iacobucci C, Hage C, Schafer M, Sinz A (2017) A novel MS-cleavable azo cross-linker for peptide structure analysis by free radical initiated peptide sequencing (FRIPS). *J Am Soc Mass Spectrom* 28:2039–2053. <https://doi.org/10.1007/s13361-017-1744-6>
- Iacobucci C, Götz M, Piotrowski C, Arlt C, Rehkamp A, Ihling C, Hage C, Sinz A (2018) Carboxyl- and photo-reactive, MS-cleavable cross-linkers: unveiling the real nature of diazirine-based reagents. *Anal Chem* 90(4):2805–2809
- Jaiswal M, Crabtree N, Bauer MA, Hall R, Raney KD, Zybailov BL (2014) XLPM: efficient algorithm for the analysis of protein-protein contacts using chemical cross-linking mass spectrometry. *BMC bioinformatics* 15 Suppl 11:S16. <https://doi.org/10.1186/1471-2105-15-S11-S16>
- Jumper CC, Bomgardner R, Rogers J, Etienne C, Schriemer DC (2012) High-resolution mapping of carbene-based protein footprints. *Anal Chem* 84(10):4411–4418. <https://doi.org/10.1021/ac300120z>
- Kalkhof S, Sinz A (2008) Chances and pitfalls of chemical cross-linking with amine-reactive N-hydroxysuccinimide esters. *Anal Bioanal Chem* 392(1–2):305–312. <https://doi.org/10.1007/s00216-008-2231-5>
- Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, Meiler J (2010) Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* 49(14):2987–2998. <https://doi.org/10.1021/bi902153g>
- Kiosze-Becker K, Ori A, Gerovac M, Heuer A, Nurenberg-Goloub E, Rashid UJ, Becker T, Beckmann R, Beck M, Tampe R (2016) Structure of the ribosome post-recycling complex probed by chemical cross-linking and mass spectrometry. *Nat Commun* 7:13248. <https://doi.org/10.1038/ncomms13248>
- Kosinski J, von Appen A, Ori A, Karius K, Muller CW, Beck M (2015) Xlink analyzer: software for analysis and visualization of cross-linking data in the context of three-dimensional structures. *J Struct Biol* 189(3):177–183. <https://doi.org/10.1016/j.jsb.2015.01.014>
- Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, Beck M, Aebersold R (2010) Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Molecular & cellular proteomics : MCP* 9(8):1634–1649. <https://doi.org/10.1074/mcp.R000001-MCP201>
- Leitner A, Joachimiak LA, Unverdorben P, Walzthoeni T, Frydman J, Forster F, Aebersold R (2014a) Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes. *Proc Natl Acad Sci U S A* 111(26):9455–9460. <https://doi.org/10.1073/pnas.1320298111>

- Leitner A, Walzthoeni T, Aebersold R (2014b) Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline. *Nat Protoc* 9(1):120–137. <https://doi.org/10.1038/nprot.2013.168>
- Leitner A, Faini N, Stengel F, Aebersold R (2016) Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. *Trends Biochem Sci* 41(1):20–32. <https://doi.org/10.1016/j.tibs.2015.10.008>
- Li X, Mooney P, Zheng S, Booth CR, Braunfeld MB, Gubbens S, Agard DA, Cheng Y (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat Methods* 10(6):584–590. <https://doi.org/10.1038/nmeth.2472>
- Lima DB, de Lima TB, Balbuena TS, Neves-Ferreira AGC, Barbosa VC, Gozzo FC, Carvalho PC (2015) SIM-XL: a powerful and user-friendly tool for peptide cross-linking analysis. *J Proteome* 129:51–55. <https://doi.org/10.1016/j.jprot.2015.01.013>
- Lipstein N, Schaks S, Dimova K, Kalkhof S, Ihling C, Kolbel K, Ashery U, Rhee J, Brose N, Sinz A, Jahn O (2012) Nonconserved  $Ca_2+$ /calmodulin binding sites in Munc13s differentially control synaptic short-term plasticity. *Mol Cell Biol* 32(22):4628–4641. <https://doi.org/10.1128/Mcb.00933-12>
- Liu F, Rijkers DT, Post H, Heck AJ (2015) Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat Methods* 12(12):1179–1184. <https://doi.org/10.1038/nmeth.3603>
- Liu F, Lossl P, Scheltema R, Viner R, Heck AJR (2017) Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat Commun* 8:15473. <https://doi.org/10.1038/ncomms15473>
- Lössl P, Sinz A (2016) Combining amine-reactive cross-linkers and photo-reactive amino acids for 3D-structure analysis of proteins and protein complexes. *Methods Mol Biol* 1394:109–127. [https://doi.org/10.1007/978-1-4939-3341-9\\_9](https://doi.org/10.1007/978-1-4939-3341-9_9)
- Lössl P, Kölbl K, Tänzler D, Nannemann D, Ihling CH, Keller MV, Schneider M, Zaucke F, Meiler J, Sinz A (2014) Analysis of Nidogen-1/laminin gamma1 interaction by cross-linking, mass spectrometry, and computational modeling reveals multiple binding modes. *PLoS One* 9(11):e112886. <https://doi.org/10.1371/journal.pone.0112886>
- Maadi H, Nami B, Wang Z (2017) Dimerization assessment of epithelial growth factor family of receptor tyrosine kinases by using cross-linking reagent. *Methods Mol Biol* 1652:101–108. [https://doi.org/10.1007/978-1-4939-7219-7\\_6](https://doi.org/10.1007/978-1-4939-7219-7_6)
- Mädler S, Bich C, Touboul D, Zenobi R (2009) Chemical cross-linking with NHS esters: a systematic study on amino acid reactivities. *J Mass Spectrom* 44(5):694–706. <https://doi.org/10.1002/jms.1544>
- Maupetit J, Derreumaux P, Tuffery P (2009) PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic acids research* 37 (web server issue):W498–503. <https://doi.org/10.1093/nar/gkp323>
- McIlwain S, Tamura K, Kertesz-Farkas A, Grant CE, Diamant B, Frewen B, Howbert JJ, Hoopmann MR, Kall L, Eng JK, MacCoss MJ, Noble WS (2014) Crux: rapid open source protein tandem mass spectrometry analysis. *J Proteome Res* 13(10):4488–4491. <https://doi.org/10.1021/pr500741y>
- Merkley ED, Rysavy S, Kahraman A, Hafen RP, Daggett V, Adkins JN (2014) Distance restraints from crosslinking mass spectrometry: mining a molecular dynamics simulation database to evaluate lysine-lysine distances. *Prot Sci* 23(6):747–759. <https://doi.org/10.1002/pro.2458>
- Müller DR, Schindler P, Towbin H, Wirth U, Voshol H, Hoving S, Steinmetz MO (2001) Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis. *Anal Chem* 73(9):1927–1934
- Müller MQ, Schafer M, Dreier F, Ihling CH, Sinz A (2010) Cleavable cross-linker for protein structure analysis: reliable identification of cross-linking products by tandem MS. *Anal Chem* 82(16):6958–6968. <https://doi.org/10.1021/ac101241t>
- Nielsen T, Thaysen-Andersen M, Larsen N, Jorgensen FS, Houen G, Hojrup P (2007) Determination of protein conformation by isotopically labelled cross-linking and dedicated software: application to the chaperone, calreticulin. *Int J Mass Spectrom* 268(2–3):217–226. <https://doi.org/10.1016/j.jms.2007.06.019>

- Novak P, Kruppa GH (2008) Intra-molecular cross-linking of acidic residues for protein structure studies. *Eur J Mass Spectrom* 14(6):355–365. <https://doi.org/10.1255/ejms.963>
- Nury C, Redeker V, Dautrey S, Romieu A, van der Rest G, Renard PY, Melki R, Chamot-Rooke J (2015) A novel bio-orthogonal cross-linker for improved protein/protein interaction analysis. *Anal Chem* 87(3):1853–1860. <https://doi.org/10.1021/ac503892c>
- Operana TN, Tukey RH (2007) Oligomerization of the UDP-glucuronosyltransferase 1A proteins: homo- and heterodimerization analysis by fluorescence resonance energy transfer and co-immunoprecipitation. *J Biol Chem* 282(7):4821–4829. <https://doi.org/10.1074/jbc.M609417200>
- Panchaud A, Singh P, Shaffer SA, Goodlett DR (2010) xComb: a cross-linked peptide database approach to protein-protein interaction analysis. *J Proteome Res* 9(5):2508–2515. <https://doi.org/10.1021/pr9011816>
- Petrochenko EV, Serpa JJ, Hardie DB, Berjanskii M, Suriyamongkol BP, Wishart DS, Borchers CH (2012) Use of proteinase K nonspecific digestion for selective and comprehensive identification of interpeptide cross-links: application to prion proteins. *Mol Cell Proteom* 11(7):M111 013524. <https://doi.org/10.1074/mcp.M111.013524>
- Petrochenko EV, Makepeace KA, Borchers CH (2014) DXMSMS match program for automated analysis of LC-MS/MS data obtained using isotopically coded CID-cleavable cross-linking reagents. *Curr Prot Bioinform* 48:8.18.11–19. doi:<https://doi.org/10.1002/0471250953.bi0818s48>
- Piotrowski C, Ihling CH, Sinz A (2015) Extending the cross-linking/mass spectrometry strategy: facile incorporation of photo-activatable amino acids into the model protein calmodulin in *Escherichia coli* cells. *Methods* 89:121–127. <https://doi.org/10.1016/j.ymeth.2015.02.012>
- Politis A, Stengel F, Hall Z, Hernandez H, Leitner A, Walzthoeni T, Robinson CV, Aebersold R (2014) A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nat Methods* 11(4):403–406. <https://doi.org/10.1038/nmeth.2841>
- Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Seraphin B (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* 24(3):218–229. <https://doi.org/10.1006/meth.2001.1183>
- Rampl E, Stranzl T, Orban-Nemeth Z, Hollenstein DM, Hudecz O, Schloegelhofer P, Mechtler K (2015) Comprehensive cross-linking mass spectrometry reveals parallel orientation and flexible conformations of plant HOP2-MND1. *J Proteome Res* 14(12):5048–5062. <https://doi.org/10.1021/acs.jproteome.5b00903>
- Rappsilber J (2011) The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J Struct Biol* 173(3):530–540. <https://doi.org/10.1016/j.jsb.2010.10.014>
- Rasmussen MI, Refsgaard JC, Peng L, Houen G, Hojrup P (2011) CrossWork: software-assisted identification of cross-linked peptides. *J Proteome* 74(10):1871–1883. <https://doi.org/10.1016/j.jprot.2011.04.019>
- Rinner O, Seebacher J, Walzthoeni T, Mueller LN, Beck M, Schmidt A, Mueller M, Aebersold R (2008) Identification of cross-linked peptides from large sequence databases. *Nat Methods* 5(4):315–318. <https://doi.org/10.1038/nmeth.1192>
- Ryu Y, Schultz PG (2006) Efficient incorporation of unnatural amino acids into proteins in *Escherichia coli*. *Nat Methods* 3(4):263–265 nmeth864 [pii] <https://doi.org/10.1038/nmeth864>
- Sarpe V, Rafiei A, Hepburn M, Ostan N, Schryvers AB, Schriemer DC (2016) High sensitivity crosslink detection coupled with integrative structure modeling in the mass spec studio. *Mol Cell Proteomics* 15(9):3071–3080. <https://doi.org/10.1074/mcp.O116.058685>
- Schilling B, Row RH, Gibson BW, Guo X, Young MM (2003) MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *J Am Soc Mass Spectrom* 14(8):834–850. [https://doi.org/10.1016/S1044-0305\(03\)00327-1](https://doi.org/10.1016/S1044-0305(03)00327-1)
- Schmidt R, Sinz A (2017) Improved single-step enrichment methods of cross-linked products for protein structure analysis and protein interaction mapping. *Anal Bioanal Chem* 409(9):2393–2400. <https://doi.org/10.1007/s00216-017-0185-1>

- Schmidt A, Kalkhof S, Ihling C, Cooper DM, Sinz A (2005) Mapping protein interfaces by chemical cross-linking and Fourier transform ion cyclotron resonance mass spectrometry: application to a calmodulin/adenylyl cyclase 8 peptide complex. *Eur J Mass Spectrom* 11(5):525–534. <https://doi.org/10.1255/ejms.748>
- Schwarz R, Tanzler D, Ihling CH, Sinz A (2016) Monitoring solution structures of peroxisome proliferator-activated receptor beta/delta upon ligand binding. *PLoS One* 11(3):e0151412. <https://doi.org/10.1371/journal.pone.0151412>
- Schweppe DK, Chavez JD, Lee CF, Caudal A, Kruse SE, Stuppard R, Marcinek DJ, Shadel GS, Tian R, Bruce JE (2017) Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *Proc Natl Acad Sci U S A* 114(7):1732–1737. <https://doi.org/10.1073/pnas.1617220114>
- Sinz A (2006) Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom Rev* 25(4):663–682. <https://doi.org/10.1002/mas.20082>
- Sinz A (2014) The advancement of chemical cross-linking and mass spectrometry for structural proteomics: from single proteins to protein interaction networks. *Expert Rev Proteomics* 11(6):733–743. <https://doi.org/10.1586/14789450.2014.960852>
- Sinz A (2017) Divide and conquer: cleavable cross-linkers to study protein conformation and protein-protein interactions. *Anal Bioanal Chem* 409(1):33–44. <https://doi.org/10.1007/s00216-016-9941-x>
- Soderberg CA, Lambert W, Kjellstrom S, Wiegandt A, Wulff RP, Mansson C, Rutsdottir G, Emanuelsson C (2012) Detection of crosslinks within and between proteins by LC-MALDI-TOF/TOF and the software FINDX to reduce the MSMS-data to acquire for validation. *PLoS One* 7(6):e38927. <https://doi.org/10.1371/journal.pone.0038927>
- Speers AE, Cravatt BF (2005) A tandem orthogonal proteolysis strategy for high-content chemical proteomics. *J Am Chem Soc* 127(28):10018–10019. <https://doi.org/10.1021/ja0532842>
- Suchanek M, Radzikowska A, Thiele C (2005) Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells. *Nat Methods* 2(4):261–267. <https://doi.org/10.1038/nmeth752>
- Tan D, Li Q, Zhang MJ, Liu C, Ma C, Zhang P, Ding YH, Fan SB, Tao L, Yang B, Li X, Ma S, Liu J, Feng B, Liu X, Wang HW, He SM, Gao N, Ye K, Dong MQ, Lei X (2016) Trifunctional cross-linker for mapping protein-protein interaction networks and comparing protein conformational states. *eLife* 5. <https://doi.org/10.7554/eLife.12509>
- Tang X, Bruce JE (2010) A new cross-linking strategy: protein interaction reporter (PIR) technology for protein-protein interaction studies. *Mol BioSyst* 6(6):939–947. <https://doi.org/10.1039/b920876c>
- Tang Y, Chen Y, Lichti CF, Hall RA, Raney KD, Jennings SF (2005) CLPM: a cross-linked peptide mapping algorithm for mass spectrometric analysis. *BMC bioinformatics* 6 Suppl 2:S9. <https://doi.org/10.1186/1471-2105-6-S2-S9>
- Tinnefeld V, Venne AS, Sickmann A, Zahedi RP (2017) Enrichment of cross-linked peptides using charge-based fractional diagonal chromatography (ChaFRADIC). *J Proteome Res* 16:459–469. <https://doi.org/10.1021/acs.jproteome.6b00587>
- Walker-Gray R, Stengel F, Gold MG (2017) Mechanisms for restraining cAMP-dependent protein kinase revealed by subunit quantitation and cross-linking approaches. *Proc Natl Acad Sci U S A* 114(39):10414–10419. <https://doi.org/10.1073/pnas.1701782114>
- Walzthoeni T, Leitner A, Stengel F, Aebersold R (2013) Mass spectrometry supported determination of protein complex structure. *Curr Opin Struct Biol* 23(2):252–260. <https://doi.org/10.1016/j.sbi.2013.02.008>
- Walzthoeni T, Joachimiak LA, Rosenberger G, Rost HL, Malmstrom L, Leitner A, Frydman J, Aebersold R (2015) xTract: software for characterizing conformational changes of protein complexes by quantitative cross-linking mass spectrometry. *Nat Methods* 12(12):1185–1190. <https://doi.org/10.1038/nmeth.3631>

- Weerapana E, Speers AE, Cravatt BF (2007) Tandem orthogonal proteolysis-activity-based protein profiling (TOP-ABPP)—a general method for mapping sites of probe modification in proteomes. *Nat Protoc* 2(6):1414–1425. <https://doi.org/10.1038/nprot.2007.194>
- Weisbrod CR, Chavez JD, Eng JK, Yang L, Zheng C, Bruce JE (2013) In vivo protein interaction network identified with a novel real-time cross-linked peptide identification strategy. *J Proteome Res* 12(4):1569–1579. <https://doi.org/10.1021/pr3011638>
- Weisz DA, Liu H, Zhang H, Thangapandian S, Tajkhorshid E, Gross ML, Pakrasi HB (2017) Mass spectrometry-based cross-linking study shows that the Psb28 protein binds to cytochrome b559 in photosystem II. *Proc Natl Acad Sci U S A* 114(9):2224–2229. <https://doi.org/10.1073/pnas.1620360114>
- Wittelsberger A, Thomas BE, Mierke DF, Rosenblatt M (2006) Methionine acts as a “magnet” in photoaffinity crosslinking experiments. *Febs Lett* 580(7):1872–1876. <https://doi.org/10.1016/j.febslet.2006.02.050>
- Xu H, Zhang L, Freitas MA (2008) Identification and characterization of disulfide bonds in proteins and peptides from tandem MS data by use of the MassMatrix MS/MS search engine. *J Proteome Res* 7(1):138–144. <https://doi.org/10.1021/pr070363z>
- Yang B, Wu YJ, Zhu M, Fan SB, Lin J, Zhang K, Li S, Chi H, Li YX, Chen HF, Luo SK, Ding YH, Wang LH, Hao Z, Xiu LY, Chen S, Ye K, He SM, Dong MQ (2012) Identification of cross-linked peptides from complex samples. *Nat Methods* 9(9):904–906. <https://doi.org/10.1038/nmeth.2099>
- Yang Y, Song H, Chen PR (2016a) Genetically encoded photocrosslinkers for identifying and mapping protein-protein interactions in living cells. *IUBMB Life* 68(11):879–886. <https://doi.org/10.1002/iub.1560>
- Yang Y, Song H, He D, Zhang S, Dai S, Lin S, Meng R, Wang C, Chen PR (2016b) Genetically encoded protein photocrosslinker with a transferable mass spectrometry-identifiable label. *Nat Commun* 7:12299. <https://doi.org/10.1038/ncomms12299>
- Yilmaz S, Drepper F, Hulstaert N, Cernic M, Gevaert K, Economou A, Warscheid B, Martens L, Vandermarliere E (2016) Xilmass: a new approach toward the identification of cross-linked peptides. *Anal Chem* 88(20):9949–9957. <https://doi.org/10.1021/acs.analchem.6b01585>
- Young MM, Tang N, Hempel JC, Oshiro CM, Taylor EW, Kuntz ID, Gibson BW, Dollinger G (2000) High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc Natl Acad Sci U S A* 97(11):5802–5806. <https://doi.org/10.1073/pnas.090099097>
- Yu F, Li N, Yu W (2016) ECL: an exhaustive search tool for the identification of cross-linked peptides using whole database. *BMC Bioinform* 17(1):217. <https://doi.org/10.1186/s12859-016-1073-y>
- Yu F, Li N, Yu W (2017) Exhaustively identifying cross-linked peptides with a linear computational complexity. *J Proteome Res* 16(10):3942–3952. <https://doi.org/10.1021/acs.jproteome.7b00338>
- Zhang Y (2009) I-TASSER: fully automated protein structure prediction in CASP8. *Proteins* 77(Suppl 9):100–113. <https://doi.org/10.1002/prot.22588>
- Zheng C, Weisbrod CR, Chavez JD, Eng JK, Sharma V, Wu X, Bruce JE (2013) XLink-DB: database and software tools for storing and visualizing protein interaction topology data. *J Proteome Res* 12(4):1989–1995. <https://doi.org/10.1021/pr301162j>
- Ziemanowicz DS, Bomgarden R, Etienne C, Schriemer DC (2017) Amino acid insertion frequencies arising from photoproducts generated using aliphatic Diazirines. *J Am Soc Mass Spectrom* 28:2011–2021. <https://doi.org/10.1007/s13361-017-1730-z>

# Chapter 9

## Prediction of Structures and Interactions from Genome Information



Sanzo Miyazawa

**Abstract** Predicting three dimensional residue-residue contacts from evolutionary information in protein sequences was attempted already in the early 1990s. However, contact prediction accuracies of methods evaluated in CASP experiments before CASP11 remained quite low, typically with <20% true positives. Recently, contact prediction has been significantly improved to the level that an accurate three dimensional model of a large protein can be generated on the basis of predicted contacts. This improvement was attained by disentangling direct from indirect correlations in amino acid covariations or cosubstitutions between sites in protein evolution. Here, we review statistical methods for extracting causative correlations and various approaches to describe protein structure, complex, and flexibility based on predicted contacts.

**Keywords** Contact prediction · Direct coupling · Amino acid covariation · Amino acid cosubstitution · Partial correlation · Maximum entropy model · Inverse Potts model · Markov random field · Boltzmann machine · Deep neural network

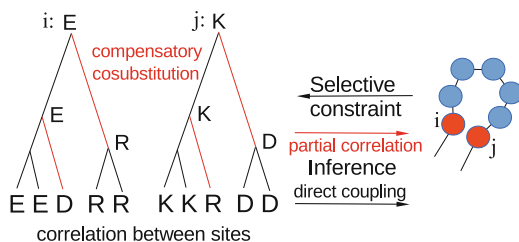
### 9.1 Introduction

The evolutionary history of protein sequences is a valuable source of information in many fields of science not only in evolutionary biology but even to understand protein structures. Residue-residue interactions that fold a protein into a unique three-dimensional (3D) structure and make it play a specific function impose structural and functional constraints in varying degrees on each amino acid. Selective constraints on amino acids are recorded in amino acid orders in homologous protein sequences and also in the evolutionary trace of amino acid substitutions. Negative

---

S. Miyazawa (✉)  
Gunma University, Kiryu, Japan

**Fig. 9.1** Amino acids at sites  $i$  and  $j$  in a MSA are shown with a phylogenetic tree. Causative correlations between sites in protein evolution are extracted from the MSA or phylogenetic tree, and utilized to infer close residue pairs



effects caused by mutations at one site must be compensated by successive mutations at other sites (Yanovsky et al. 1964; Fitch and Markowitz 1970; Maisnier-Patin and Andersson 2004), causing covariations/cosubstitutions/coevolution between sites (Tufféry and Darlu 2000; Fleishman et al. 2004; Dutheil et al. 2005; Dutheil and Galtier 2007), otherwise most negative mutants will be eliminated from a gene pool and never reach fixation in population. Such structural and functional constraints arise from interactions between sites mostly in close spatial proximity. Thus, it has been suggested and also shown that the types of amino acids (Lapedes et al. 1999, 2002, 2012; Russ et al. 2005; Skerker et al. 2008; Burger and van Nimwegen 2008; Weigt et al. 2009; Halabi et al. 2009; Burger and van Nimwegen 2010; Morcos et al. 2011; Marks et al. 2011) and amino acid substitutions (Altschuh et al. 1988; Göbel et al. 1994; Shindyalov et al. 1994; Pollock and Taylor 1997; Pollock et al. 1999; Atchley et al. 2000; Fariselli et al. 2001; Fodor and Aldrich 2004; Fleishman et al. 2004; Dutheil et al. 2005; Martin et al. 2005; Fares and Travers 2006; Doron-Faigenboim and Pupko 2007; Dutheil and Galtier 2007; Dunn et al. 2008; Poon et al. 2008; Dutheil 2012; Gulyás-Kovács 2012) are correlated between sites that are close in a protein 3D structure. However, until CASP11, contact prediction accuracy remained quite low, typically with  $\leq 20\%$  true positives for top- $L/5$  long-range contacts in free modeling targets (Kosciolek and Jones 2016);  $L$  denotes protein length. Recently contact prediction has been significantly improved to the level that an accurate three dimensional model of a large protein ( $\approx 250$  residues) can be generated on the basis of predicted contacts (Moult et al. 2016). These improvements were attained primarily by disentangling direct from indirect correlations in amino acid covariations or cosubstitutions between sites in protein evolution, and secondarily by reducing phylogenetic biases in a multiple sequence alignment (MSA) or removing them on the basis of a phylogenetic tree; see Fig. 9.1.

Here, we review statistical methods for extracting causative correlations in amino acid covariations/cosubstitutions between sites, and various approaches to describe protein structure, complex and flexibility based on predicted contacts. Mathematical formulation of each statistical method is concisely described in the unified manner in an appendix, the full version of which will be found in the article (Miyazawa 2017a) submitted to arXiv.

## 9.2 Statistical Methods to Extract Causative Interactions Between Sites

The primary task to develop a robust method toward contact prediction is to detect causative correlations, which reflect evolutionary constraints, in amino acid covariations between sites in a multiple sequence alignment (MSA) or in amino acid cosubstitutions between sites in branches of a phylogenetic tree; see Table 9.1. The former was called direct coupling analysis (DCA) (Morcos et al. 2011).

**Table 9.1** Statistical methods for disentangling direct from indirect correlations between sites

| Category  |  |
|---|--|
| Method name   | Method/algorithm   |
| (A) Direct coupling analysis of amino acid covariations between sites in a MSA                      |  |
| Boltzmann machine   | Markov chain Monte Carlo to calculate marginal probabilities and gradient descent to estimate fields and couplings   |
| CMI (Lapedes et al. 2012)   | Boltzmann machine to estimate conditional mutual information   |
| mpDCA (Weigt et al. 2009)   | Message-passing algorithm to estimate marginal probabilities and gradient descent to estimate fields and couplings   |
| mfDCA (Morcos et al. 2011; Marks et al. 2011)   | Mean field approximation to estimate the partition function  |
| PSICOV (Jones et al. 2012)  | Graphical lasso (Gaussian approximation with an exponential prior) with a shrinkage method for a covariance matrix   |
| GaussDCA (Baldassi et al. 2014)   | A multivariate Gaussian model with a normal-inverse-Wishart prior  |
| plmDCA (Ekeberg et al. 2013, 2014)  | Pseudo-likelihood maximization with Gaussian priors ( $\ell_2$ regularizers)   |
| GREMLIN (Balakrishnan et al. 2011; Kamisetty et al. 2013)   | Pseudo-likelihood maximization with $\ell_1$ regularization terms (Balakrishnan et al. 2011) or with Gaussian priors (Kamisetty et al. 2013) which depend on site pair |
| ACE (Cocco and Monasson 2011, 2012; Barton et al. 2016)   | Adaptive cluster expansion of cross-entropy with Gaussian priors   |
| Persistent VI & Fadeout   | Variational inference with sparsity-inducing prior, horseshoe (Ingraham and Marks 2016)  |
| Sutto et al. (2015)   | Boltzmann machine with $\ell_2$ regularization terms   |
| DI (Taylor and Sadowski 2011)   | Partial correlation of normalized mutual informations between sites  |
| (B) Partial correlation analysis of amino acid cosubstitutions between sites in a phylogenetic tree |  |
| pcSV (Miyazawa 2013)  | Partial correlation coefficients of coevolutionary substitutions between sites within branches in a phylogenetic tree  |



## 9.2.1 Direct Coupling Analysis for Amino Acid Covariations Between Sites in a Multiple Sequence Alignment

The direct coupling analysis is based on the maximum entropy model for the distribution of protein sequences, which satisfies the observed statistics in a MSA.

### 9.2.1.1 Maximum Entropy Model for the Distribution of Protein Sequences

Let us consider probability distributions  $P(\sigma)$  of amino acid sequences,  $\sigma \equiv (\sigma_1, \dots, \sigma_L)^T$  with  $\sigma_i \in \{\text{amino acids, deletion}\}$ , single-site and two-site marginal probabilities of which are equal to a given frequency  $P_i(a_k)$  of amino acid  $a_k$  at each site  $i$  and a given frequency  $P_{ij}(a_k, a_l)$  of amino acid pair  $(a_k, a_l)$  for site pair  $(i, j)$ , respectively.

$$P(\sigma_i = a_k) \equiv \sum_{\sigma} P(\sigma) \delta_{\sigma_i a_k} = P_i(a_k) \quad (9.1)$$

$$P(\sigma_i = a_k, \sigma_j = a_l) \equiv \sum_{\sigma} P(\sigma) \delta_{\sigma_i a_k} \delta_{\sigma_j a_l} = P_{ij}(a_k, a_l) \quad (9.2)$$

where  $a_k \in \{\text{amino acids, deletion}\}$ ,  $k = 1, \dots, q$ ,  $q \equiv |\{\text{amino acids, deletion}\}| = 21$ ,  $i, j = 1, \dots, L$ , and  $\delta_{\sigma_i a_k}$  is the Kronecker delta. The distribution  $P_{\text{ME}}$  with the maximum entropy is

$$P_{\text{ME}}(\sigma|h, J) \quad (9.3)$$

$$\begin{aligned} &= \arg \max_{P(\sigma)} [-\sum_{\sigma} P(\sigma) \log P(\sigma) + \lambda(\sum_{\sigma} P(\sigma) - 1) \\ &+ \sum_i [h_i(a_k)(\sum_{\sigma} P(\sigma) \delta_{\sigma_i a_k} - P_i(a_k))] \\ &+ \sum_i \sum_{j>i} [J_{ij}(a_k, a_l)(\sum_{\sigma} P(\sigma) \delta_{\sigma_i a_k} \delta_{\sigma_j a_l} - P_{ij}(a_k, a_l))] ] = \frac{1}{Z} e^{-H_{\text{Potts}}(\sigma|h, J)} \end{aligned} \quad (9.4)$$

where  $\lambda$ ,  $h_i(a_k)$ , and  $J_{ij}(a_k, a_l)$  are Lagrange multipliers, and a Hamiltonian  $H_{\text{Potts}}$ , which is called that of the Potts model for  $q > 2$  (or the Ising model for  $q = 2$ ), and a partition function  $Z$  are defined as

$$-H_{\text{Potts}}(\sigma|h, J) = \sum_i h_i(\sigma_i) + \sum_{i<j} J_{ij}(\sigma_i, \sigma_j), \quad Z = \sum_{\sigma} e^{-H_{\text{Potts}}(\sigma|h, J)} \quad (9.5)$$

where  $h_i(a_k)$  and  $J_{ij}(a_k, a_l)$  are interaction potentials called fields and couplings.

Although pairwise frequencies  $P_{ij}(a_k, a_l)$  reflect not only direct but indirect correlations in amino acid covariations between sites, couplings  $J_{ij}(a_k, a_l)$  reflect causative correlations only. Thus, it is essential to estimate fields and couplings from marginal probabilities. This model is called the inverse Potts model.

### 9.2.1.2 Log-Likelihood and Log-Posterior-Probability

Log-posterior-probability and log-likelihood for the Potts model are

$$\log P_{\text{post}}(h, J | \{\sigma\}) \propto \ell_{\text{Potts}}(\{P_i\}, \{P_{ij}\} | h, J) + \log P_0(h, J) \quad (9.6)$$

$$\ell_{\text{Potts}}(\{P_i\}, \{P_{ij}\} | h, J) = B \sum_{\sigma} P_{\text{obs}}(\sigma) \log P_{\text{ME}}(\sigma | h, J) \quad (9.7)$$

where  $P_{\text{obs}}(\equiv \sum_{\tau=1}^B \delta_{\sigma\sigma\tau} / B)$  is the observed distribution of  $\sigma$  specified with  $\{P_i(a_k)\}$  and  $\{P_{ij}(a_k, a_l)\}$ , and  $B$  is the number of instances; sequences  $\sigma^\tau$  are assumed here to be independently and identically distributed samples in sequence space.  $P_0(h, J)$  is a prior probability of  $(h, J)$ .

Let us define cross entropy (Cocco and Monasson 2012) as the negative log-posterior-probability per instance.

$$\begin{aligned} S_0(h, J | \{P_i\}, \{P_{ij}\}) &\propto -(\log P_{\text{post}}(h, J | \{\sigma\})) / B \\ &\equiv S_{\text{Potts}}(h, J | \{P_i\}, \{P_{ij}\}) + R(h, J) \end{aligned} \quad (9.8)$$

where the cross entropy  $S_{\text{Potts}}$ , which is the negative log-likelihood per instance for the Potts model, and the negative log-prior per instance  $R$  are defined as follows.

$$S_{\text{Potts}}(h, J | \{P_i\}, \{P_{ij}\}) \equiv -\ell_{\text{Potts}}(\{P_i\}, \{P_{ij}\} | h, J) / B \quad (9.9)$$

$$= \log Z(h, J) - \sum_i \sum_k h_i(a_k) P_i(a_k) - \sum_i \sum_k \sum_{j>i} \sum_l J_{ij}(a_k, a_l) P_{ij}(a_k, a_l) \quad (9.10)$$

$$R(h, J) \equiv -\log(P_0(h, J)) / B \quad (9.11)$$

The maximum likelihood estimates of  $h$  and  $J$ , which minimize the cross entropy with  $R = 0$ , satisfy the following equations.

$$\frac{\partial \log Z(h, J)}{\partial h_i(a_k)} = P_i(a_k), \quad \frac{\partial \log Z(h, J)}{\partial J_{ij}(a_k, a_l)} = P_{ij}(a_k, a_l) \quad (9.12)$$

It is, however, hardly tractable to computationally evaluate the partition function  $Z(h, J)$  for any reasonable system size as a function of  $h$  and  $J$ . Thus, approximate maximization of the log-likelihood or minimization of the cross entropy is needed to estimate  $h$  and  $J$ .

The minimum of the cross entropy with  $R = 0$  for the Potts model is just the Legendre transform of  $\log Z(h, J)$  from  $(h, J)$  to  $(\{P_i\}, \{P_{ij}\})$ , (Eq. 9.10), and is equal to the entropy of the Potts model satisfying Eqs. 9.1 and 9.2;

$$S_{\text{Potts}}(\{P_i\}, \{P_{ij}\}) \equiv \min_{h, J} S_{\text{Potts}}(h, J | \{P_i\}, \{P_{ij}\}) = \sum_{\sigma} -P(\sigma) \log P(\sigma) \quad (9.13)$$

The cross entropy  $S_{\text{Potts}}(h, J | \{P_i\}, \{P_{ij}\})$  in Eq. 9.10 is invariant under a certain transformation of fields and couplings,  $J_{ij}(a_k, a_l) \rightarrow J_{ij}(a_k, a_l) - J_{ij}^1(a_k) - J_{ji}^1(a_l) + J_{ij}^0, h_i(a_k) \rightarrow h_i(a_k) - h_i^0 + \sum_{j \neq i} J_{ij}^1(a_k)$  for any  $J_{ij}^1(a_k)$ ,  $J_{ij}^0$  and  $h_i^0$ . This gauge-invariance reduces the number of independent variables in the Potts model to  $(q - 1)L$  fields and  $(q - 1)L \times (q - 1)L$  couplings.

A prior  $P_0(h, J)$  yields regularization terms for  $h$  and  $J$  (Cocco and Monasson 2012). If a Gaussian distribution is employed for the prior, then it will yield  $\ell_2$  norm regularization terms.  $\ell_1$  norm regularization corresponds to the case of exponential priors. Given marginal probabilities, the estimates of fields and couplings are those minimizing the cross entropy.

$$(h, J) = \arg \min_{(h, J)} S_0(h, J | \{P_i\}, \{P_{ij}\}), \quad S_0(\{P_i\}, \{P_{ij}\}) \equiv \min_{(h, J)} S_0(h, J | \{P_i\}, \{P_{ij}\}) \quad (9.14)$$

Since  $S_0(\{P_i\}, \{P_{ij}\})$  is the Legendre transform of  $(\log Z(h, j) + R(h, J))$  from  $(h, J)$  to  $(\{P_i\}, \{P_{ij}\})$ , these optimum  $h$  and  $J$  can also be calculated from

$$h_i(a_k) = -\frac{\partial S_0(\{P_i\}, \{P_{ij}\})}{\partial P_i(a_k)}, \quad J_{ij}(a_k, a_l) = -\frac{\partial S_0(\{P_i\}, \{P_{ij}\})}{\partial P_{ij}(a_k, a_l)} \quad (9.15)$$

In most methods for contact prediction, residue pairs are predicted as contacts in the decreasing order of score ( $\mathcal{S}_{ij}$ ) calculated from fields  $\{J_{ij}(a_k, a_l) | 1 \leq k, l < q\}$ ; see Eq. 9.47.

### 9.2.1.3 Inverse Potts Model

The problem of inferring interactions from observations of instances has been studied as inverse statistical mechanics, particularly inverse Potts model for Eq. 9.4, in the field of statistical physics, as a Markov random field, Markov network or undirected graphical model in the domain of physics, statistics and information science, and as Boltzmann machine in the field of machine learning.

The maximum-entropy approach to the prediction of residue-residue contacts toward protein structure prediction from residue covariation patterns was first described in 2002 by Lapedes and collaborators (Giraud et al. 1999; Lapedes et al. 1999, 2002, 2012). They estimated conditional mutual information (CMI), which was employed as a score for residue-residue contacts, for each site pair by Boltzmann learning with Monte Carlo importance sampling to calculate equilibrium

averages and gradient descent to minimize the cross entropy and successfully predicted contacts for 11 small proteins.

Calculating marginal probabilities for given fields and couplings by Monte Carlo simulations in Boltzmann machine is very computationally intensive. To reduce a computational load, the message passing algorithm, which is exact for a tree topology of couplings but approximate for the present model, is employed instead in mpDCA (Weigt et al. 2009). Because even the message passing algorithm is too slow to be applied to a large-scale analysis across many protein families, the mean field approximation is employed in mfDCA (Morcos et al. 2011; Marks et al. 2011);  $J^{MF} = -C^{-1}$ , where  $C_{ij}(a_k, a_l) \equiv P_{ij}(a_k, a_l) - P_i(a_k)P_j(a_l)$ . In the mean field approximation, a bottleneck in computation is the calculation of the inverse of a covariance matrix  $C$  that is a  $(q - 1)L \times (q - 1)L$  matrix. In the mean field approximation, a prior distribution in Eq. 9.11 is ignored and pseudocount is employed instead of regularization terms to make the covariance matrix invertible.

The Gaussian approximation (a continuous multivariate Gaussian model) for the probability distribution of sequences is employed together with an exponential prior (an  $\ell_1$  regularization term) in PSICOV (Jones et al. 2012), and with a normal-inverse-Wishart (NIW) prior, which is a conjugate distribution of the multivariate Gaussian, in GaussDCA (Baldassi et al. 2014). The use of NIW prior has a merit that fields and couplings can be analytically formulated; see Eqs. 9.30 and 9.31.

All methods based on the Gaussian approximation employ the analytical formula for couplings,  $J \simeq -C^{-1} = -\Theta$ , which are essentially as same as the mean field approximation with a difference that the covariance matrix ( $C$ ) or precision matrix ( $\Theta$ ) is differently estimated based on the various priors. The mean field and Gaussian approximations may be appropriate to systems of dense and weak couplings but questionable for sparse and strong couplings that is the characteristic of residue-residue contact networks. Although the mean field and Gaussian approximations successfully predict residue-residue contacts in proteins, it has been shown (Barton et al. 2016; Cocco et al. 2017) that they do not give the accurate estimates of fields and couplings in proteins.

A pseudo-likelihood with Gaussian priors ( $\ell_2$  regularization terms) is maximized to estimate fields and couplings in plmDCA (Ekeberg et al. 2013, 2014) for the Potts model with sparse interactions as well as reducing computational time; see Eq. 9.38 for the symmetric plmDCA and Eq. 9.41 for the asymmetric plmDCA. The asymmetric plmDCA method (Ekeberg et al. 2014) requires less computational time and fits particularly with parallel computing.

GREMLIN (Kamisetty et al. 2013) employs together with pseudo-likelihood Gaussian priors that depend on site pair, although its earlier version (Balakrishnan et al. 2011) employed  $\ell_1$  regularizers, which may be more appropriate to systems of sparse couplings. The  $\ell_1$  regularizers appear to learn parameters that are closer to their true strength, but the  $\ell_2$  regularizers appear to be as good as the  $\ell_1$  regularizers for the task of contact prediction that requires the relative ranking of the interactions and not their actual values (Kamisetty et al. 2013).

One of approaches to surpass the pseudo-likelihood approximation for systems of sparse couplings may be the adaptive cluster expansion (ACE) of cross

entropy (Cocco and Monasson 2011, 2012; Barton et al. 2016), in which cross entropy is approximately minimized by taking account of only site clusters the incremental entropy (cluster entropy) of which by adding one more site is significant. In this method (Barton et al. 2016), a Boltzmann machine is employed to refine fields and couplings and also to calculate model correlations such as single-site and pairwise amino acid frequencies under given fields and couplings. The results of the Boltzmann machine for both biological and artificial models showed that ACE outperforms plmDCA in recovering single-site marginals (amino acid frequencies at each site) and the distribution of the total dimensionless energies ( $H_{\text{Potts}}(\sigma)$ ) (Barton et al. 2016); those models were a lattice protein, trypsin inhibitor, HIV p7 nucleocapsid protein, multi-electrode recording of cortical neurons, and Potts models on Eridös-Rényi random graphs. More importantly ACE could accurately recover the true fields  $h$  and couplings  $J$  corresponding to Potts states with  $P_i(a_k) \geq 0.05$  for Potts models ( $L = 50$ ) on Eridös-Rényi random graphs (Barton et al. 2016). On the other hand, plmDCA gave accurate estimates of couplings at weak regularization for well sampled single-site probabilities, but less accurate fields. Also, plmDCA yielded less well inferred fields and couplings for single-site and two-site probabilities not well sampled, indicating that not well populated states should be merged. As a result, the distribution of the total energies (Barton et al. 2016) and the distribution of mutations with respect to the consensus sequence were not well reproduced (Cocco et al. 2017). Similarly, the mean field approximation could not reproduce two-site marginals and even single-site marginals (Cocco et al. 2017) and the Gaussian approximation could not well reproduce the distribution of mutations with respect to the consensus sequence (Barton et al. 2016).

However, the less reproducibility of couplings does not necessarily indicate the less predictability of residue-residue contacts, probably because in contact prediction the relative ranking of scores (Eq. 9.47) based on couplings is more important than their actual values. ACE with the optimum regularization strength with respect to the reproducibility of fields and couplings showed less accurate contact prediction than plmDCA and mfDCA. For ACE to show comparable performance of contact prediction with plmDCA, regularization strength had to be increased from  $\gamma = 2/B = 10^{-3}$  to  $\gamma = 1$  for Trypsin inhibitor, making couplings strongly damped and then the generative properties of inferred models lost (Barton et al. 2016) (Table 9.2).

### ***9.2.2 Partial Correlation of Amino Acid Cosubstitutions Between Sites at Each Branch of a Phylogenetic Tree***

In the DCA analyses on residue covariations between sites in a multiple sequence alignment (MSA), phylogenetic biases, which are sequence biases due to phylogenetic relations between species, in the MSA must be removed as well as indirect

**Table 9.2** Free softwares/servers for the direct coupling analysis

| Name   | Methods        | URL  |
|--|----------------|--|
| EVcouplings (Marks et al. 2011)                                    | mfDCA          | <a href="http://evfold.org">http://evfold.org</a>  |
| EVcouplings, plmc (Toth-Petroczy et al. 2016; Weinreb et al. 2016) | mf/plmDCA      | <a href="https://github.com/debbiemarkslab">https://github.com/debbiemarkslab</a>  |
| DCA (Morcos et al. 2011; Marks et al. 2011)                        | mfDCA          | <a href="http://dca.rice.edu/portal/dca/home">http://dca.rice.edu/portal/dca/home</a>  |
| GaussDCA (Baldassi et al. 2014)                                    | GaussDCA       | <a href="http://areeweb.polito.it/ricerca/cmp/code">http://areeweb.polito.it/ricerca/cmp/code</a>  |
| FreeContact (Kaján et al. 2014)                                    | mfDCA, PSICONV | <a href="http://rostlab.org/owiki/index.php/FreeContact">http://rostlab.org/owiki/index.php/FreeContact</a>  |
| plmDCA (Ekeberg et al. 2013, 2014)                                 | plmDCA         | <a href="http://plmdca.csc.kth.se/">http://plmdca.csc.kth.se/</a><br><a href="https://github.com/pagnani/plmDCA">https://github.com/pagnani/plmDCA</a> |
| CCMpred (Seemayer et al. 2014)                                     | plmDCA         | Performance-optimized software<br><a href="https://github.com/soedinglab/ccmpred">https://github.com/soedinglab/ccmpred</a>                            |
| GREMLIN (Balakrishnan et al. 2011; Kamisetty et al. 2013)          | GREMLIN        | <a href="http://gremlin.bakerlab.org/">http://gremlin.bakerlab.org/</a>  |
| ACE (Cocco and Monasson 2011, 2012; Barton et al. 2016)            | ACE            | <a href="https://github.com/johnbarton/ACE">https://github.com/johnbarton/ACE</a>  |
| Persistent-vi (Ingraham and Marks 2016)                            | Persistent VI  | <a href="https://github.com/debbiemarkslab">https://github.com/debbiemarkslab</a>  |

correlations between sites, but instead are reduced by taking weighted averages over homologous sequences in the calculation of single and pairwise frequencies of amino acids.

Needless to say, it is supposed that observed patterns of covariation were caused by molecular coevolution between sites. Whatever caused covariations found in the MSA, it has been confirmed that they can be utilized to predict residue pairs in close proximity in a three dimensional structure. Talavera et al. (2015) claimed, however, that covarying substitutions were mostly found on different branches of the phylogenetic tree, indicating that they might or might not be attributable to coevolution.

In order to remove phylogenetic biases and also to respond to such a claim above, it is meaningful to study covarying substitutions between sites in a phylogenetic tree-dependent manner. Such an alternative approach was taken to infer coevolving site pairs from direct correlations between sites in concurrent and compensatory substitutions within the same branches of a phylogenetic tree (Miyazawa 2013). In this method, substitution probability and mean changes of physico-chemical properties of side chain accompanied by amino acid substitutions at each site in each branch of the tree are estimated with the likelihood of each substitution to detect concurrent and compensatory substitutions. Then, partial correlation coefficients of the vectors of their characteristic changes accompanied by substitutions, substitution probability and mean changes of physico-chemical properties, along branches between sites are calculated to extract direct correlations in coevolutionary

substitutions and employed as a score for residue-residue contact. The accuracy of contact prediction by this method was comparable with that by mfDCA (Miyazawa 2013). This method, however, has a drawback to be computationally intensive, because an optimum phylogenetic tree must be estimated.

### 9.3 Machine Learning Methods to Augment the Contact Prediction Accuracy Based on Amino Acid Coevolution

All the DCA methods such as mfDCA, plmDCA, GREMLIN, and PSICOV predict significantly nonoverlapping sets of contacts (Jones et al. 2015; Kosciolk and Jones 2016; Wuyun et al. 2016). Then, increasing prediction accuracy by combining their predictions together with other sequence/structure information have been attempted (Skwark et al. 2013, 2014, 2016; Kosciolk and Jones 2014, 2016; Jones et al. 2015; Wang et al. 2017; Shendure and Ji 2017); see Table 9.3.

PconsC (Skwark et al. 2013) combines the predictions of PSICOV and plmDCA into a machine learning method, random forests, and employs alignments with HHblits (Remmert et al. 2012) and jackHMMer (Johnson et al. 2010) at four different e-value cut-offs. Five-layer neural network is employed instead of random forests in PconsC2 (Skwark et al. 2014), and plmDCA and GaussDCA are employed in PconsC3 (Skwark et al. 2016). A receptive field consisting of  $11 \times 11$  predicted contacts around each residue pair is taken into account in each layer except the first one.

**Table 9.3** Machine learning methods that combine predicted direct couplings with other sequence/structure information

| Name   | Basic method                              | Post-processing  |
|--|---|--|
| PconsC3<br>(Skwark et al. 2016)                                  | plmDCA, GaussDCA                          | 5 layer DNN; <a href="http://c3.pcons.net">http://c3.pcons.net</a> .<br>PconsC (Skwark et al. 2013), PconsC2 (Skwark et al. 2014)  |
| MetaPSICOV<br>(Kosciolk and Jones 2014, 2016; Jones et al. 2015) | PSICOV, mfDCA, GREMLIN/CCMpred            | A two stage neural network predictor; CONSIP2 pipeline<br><a href="http://bioinf.cs.ucl.ac.uk/MetaPSICOV">http://bioinf.cs.ucl.ac.uk/MetaPSICOV</a>  |
| RaptorX<br>(Wang et al. 2017)                                    | CCMpred                                   | Ultra-deep learning model consisting of 1- and 2-dimensional convolutional residual neural networks<br><a href="http://raptorx.uchicago.edu/ContactMap/">http://raptorx.uchicago.edu/ContactMap/</a> |
| iFold (CASP12 2017)  |   | Deep neural network (DNN)  |
| EPSILON-CP   | PSICOV, GREMLIN, mfDCA, CCMpred, GaussDCA | 4 hidden layer neural network with 400-200-200-50 neurons (Shendure and Ji 2017)   |

MetaPSICOV (Jones et al. 2015; Kosciolok and Jones 2016) combines the predictions of PSICOV, mfDCA, and CCMpred/GREMLIN into the first stage of a two-stage neural network predictor together with a well-established “classic” machine learning contact predictor, which utilizes many features such as amino acid profiles, predicted secondary structure and solvent accessibility along with sequence separation predicted, as an additional source of information for a little depth of MSAs. The second stage analyses the output of the first stage to eliminate outliers and to fill in the gaps in the contact map. On a set of 40 target domains with a median family size of around 40 effective sequences in CASPII, CONSIP2 server achieved an average top- $L/5$  long-range contact precision of 27% (Kosciolok and Jones 2016).

Wang et al. (2017) have also shown that a ultra-deep neural network (RaptorX) can significantly improve contact prediction based on amino acid coevolution. They have modeled short-range and long-range correlations in sequential and structural features with respect to complex sequence-structure relationships in proteins by one-dimensional and two-dimensional deep neural networks (DNN), respectively. Both the DNNs are convolutional residual neural networks. The 1D DNN performs convolutional transformations, with respect to residue position, of sequential features such as position-dependent scoring matrix, predicted 3-state secondary structure and 3-state solvent accessibility. The 2D DNN does 2D convolutional transformations of pairwise features such as coevolutional information calculated by CCMpred, mutual information, pairwise contact potentials as well as the output of the 1D DNN converted by a similar operation to outer product. Residual neural networks are employed because they can pass both linear and nonlinear informations from initial input to final output, making their training relatively easy.

## 9.4 Performance of Contact Prediction

New statistical methods based on the direct coupling analysis are confirmed in various benchmarking studies (Moult et al. 2016; CASP12 2017; Kamisetty et al. 2013; Wuyun et al. 2016) to show remarkable accuracy of contact prediction, although deep, stable alignments are required. They can more accurately detect a higher number of contacts between residues, which are very distant along sequence (Morcos et al. 2011). The top-scoring residue couplings are not only sufficiently accurate but also well-distributed to define the 3D protein fold with remarkable accuracy (Marks et al. 2011); this observation was quantified by computing, from sequence alone, all-atom 3D structures of 15 test proteins from different fold classes, ranging in size from 50 to 260 residues, including a G-protein coupled receptor. The contact prediction performs relatively better on  $\beta$  proteins than on  $\alpha$  proteins (Miyazawa 2013). These initial findings on a limited number of proteins were confirmed as a general trend in a large-scale comparative assessment of contact prediction methods (Wuyun et al. 2016; Adhikari et al. 2016).



In CASP12, RaptorX performed the best in terms of F1 score for top  $L/2$  long- and medium-range contacts of 38 free-modeling (FM) targets; the total F1 score of RaptorX was better by about 7.6% and 10.0% than the second and third best servers, iFold\_1 and the revised MetaPSICOV, respectively (Wang et al. 2017; CASP12 2017). Tested on 105 CASP11 targets, 76 past CAMEO hard targets, and 398 membrane proteins, the average top  $L(L/10)$  long-range prediction accuracies of RaptorX are 0.47(0.77) in comparison with 0.30(0.59) for MetaPSICOV and 0.21(0.47) for CCMpred (Wang et al. 2017; CASP12 2017).

### 9.4.1 MSA Dependence of Contact Prediction Accuracy

In the direct-coupling-based methods, the accuracy of predicted contacts depends on the depth (Miyazawa 2013; Kamisetty et al. 2013; Wuyun et al. 2016) and quality of multiple sequence alignment (MSA) for a target.  $5 \times L$  (protein length) aligned sequences may be desirable for accurate contact predictions (Kamisetty et al. 2013), although attempts to improve prediction methods for fewer aligned sequences have been made (Skwark et al. 2013, 2014, 2016; Wang et al. 2017). PconsC3 can be used for families with as little as 100 effective sequence members (Skwark et al. 2016). Also, RaptorX (Wang et al. 2017) attained top-  $L/2$ -accuracy  $>0.3$  for long-range contacts even by using MSAs with 20 effective sequence members.

Deepest MSAs including a target sequence were built with various values of E-value cutoff (Skwark et al. 2013) and coverage parameters (Jones et al. 2015; Kosciolek and Jones 2016) in sequence search and alignment programs based on the hidden Markov models such as HHblits and jackHMMer. Although prediction performance tends to increase in general as alignment depth is deeper (Miyazawa 2013), it was reported (Kosciolek and Jones 2016) that in the case of transmembrane domains, building too deep alignments could result in unrelated sequences or drifted domains being included. To increase alignment quality, E-value and coverage parameters may be carefully tuned for each alignment (Kosciolek and Jones 2016). In the case of alignments that might contain regions of partial matches, a too stringent sequence coverage requirement could result in missing related sequences. On the other hand, a too permissive sequence coverage requirement could pick up unrelated sequences, permitting many partial matches. A trade-off is required between the effective number of sequences and sequence coverage, and an appropriate E-value must be chosen not to much decrease both alignment depth and sequence coverage (Hopf et al. 2012).

## 9.5 Contact-Guided de novo Protein Structure Prediction

It is a primary obstacle to de novo structure prediction that current methods and computers cannot make it feasible to adequately sample the vast conformational

space a protein might take in the process of folding into the native structure (Kim et al. 2009). Thus, it is critical whether residue-residue proximities inferred with direct coupling analysis can provide sufficient information to reduce a huge search space for a protein fold, without any known 3D structural information of the protein.

Algorithms are needed to fold proteins into native folds based on contact information; see Table 9.4. Distance geometry generation (Havel et al. 1983; Braun and Go 1985) of 3D structures, which may be followed by energy minimization and molecular dynamics, will be just the primary one. In EVfold (Marks et al. 2011), contacts inferred by direct coupling analysis and predicted secondary structure information are translated into a set of distance constraints for the use of a distance geometry algorithm in the Crystallography and NMR System (CNS) (Brünger 2007). It was confirmed that the evolutionary inferred contacts can sufficiently reduce a search space in the structure predictions of 15 test proteins from different fold classes (Marks et al. 2011), and of 11 unknown and 23 known transmembrane protein structures (Hopf et al. 2012). Because distance constraints from predicted contacts may be partial in a protein sequence, they should be embedded into *ab initio* structure prediction methods.

**Table 9.4** Contact-guided de novo protein structure prediction methods and servers

| Name  | Contact prediction                           |  |
|---|--|--|
| EVfold (Marks et al. 2011, 2012)/EVfold_membrane (Hopf et al. 2012) | mfDCA/plmDCA                                 | Using distance geometry algorithm (Havel et al. 1983) and simulated annealing of CNS (Brünger 2007); <a href="http://evfold.org/">http://evfold.org/</a>   |
| DCA-fold (Sufkowska et al. 2012)                                    | mfDCA  | Simulated annealing using a coarse-grained molecular dynamics for a C $_{\alpha}$ model  |
| FRAGFOLD/FILM3  | MetaPSICOV                                   | Combining fragment-based folding algorithm (Jones et al. 2005) with PSICOV (Kosciolek and Jones 2014) and with MetaPSICOV (Jones et al. 2015).<br>FILM3 (Nugent and Jones 2012) is employed instead of FRAGFOLD (Jones 2001) for transmembrane proteins. |
| CONFOLD (Adhikari et al. 2015)                                      | EVFOLD/FRAGFOLD (PSIPRED for 2nd structures) | Two-stage contact-guided de novo protein folding, using distance geometry simulated annealing protocol in a revised CNS v1.3. <a href="http://protein.rnet.missouri.edu/confold/">http://protein.rnet.missouri.edu/confold/</a>                          |
| Rosetta (Kim et al. 2004; Ovchinnikov et al. 2016)                  | GREMLIN                                      | Fragment assembly  |

Sulkowska et al. also showed that a simple hybrid method, called DCA-fold, integrating mfDCA-predicted contacts with an accurate knowledge of secondary structure is sufficient to fold proteins in the range of 1–3 Å resolution (Sulkowska et al. 2012). In this study, simulated annealing using a coarse-grained molecular dynamics model was employed for a  $C_\alpha$  chain model, in which  $C_\alpha$ s interact with each other with a contact potential approximated by a Gaussian function and a torsional potential depending on  $C_\alpha$  dihedral angles at each position.

Adhikari et al. (2015) studied a way to effectively encode secondary structure information into distance and dihedral angle constraints that complement long-range contact constraints, and revised the CNS v1.3 to effectively use secondary structure constraints together with predicted long-range constraints; CONFOLD (Adhikari et al. 2015) consists of two stages. In the first stage secondary structure information is converted into distance, dihedral angle, and hydrogen bond constraints, and then best models are selected by executing the distance geometry simulated annealing. In the second stage self-conflicting contacts in the best structure predicted in the first stage are removed, constraints based on the secondary structures are refined, and again the distance geometry simulated annealing is executed.

Baker group (Ovchinnikov et al. 2016) embedded contact constraints predicted by GREMLIN (Kamisetty et al. 2013) as sigmoidal constraints to overcome noise in the Rosetta (Kim et al. 2004) conformational sampling and refinement. They found that model accuracy will be generally improved, if more than 3 L (protein length) sequences are available, and that large topologically complex proteins can be modeled with close to atomic-level accuracy without knowledge of homologous structures, if there are enough homologous sequences available.

On the other hand, a fragment-based folding algorithm FRAGFOLD was combined with PSICOV (Kosciolek and Jones 2014) and with MetaPSICOV (Jones et al. 2015; Kosciolek and Jones 2016); In this approach, predicted contacts are converted into additional energy terms for FRAGFOLD in addition to the pairwise potentials of mean force and solvation (Jones et al. 2015; Kosciolek and Jones 2016). FILM3 (Nugent and Jones 2012), with constraints based on predicted contacts and ones approximating Z-coordinate values within the lipid membrane, is employed instead of FRAGFOLD for transmembrane proteins.

RaptorX (Wang et al. 2017) employed the CNS suite (Brünger 2007) to generate 3D models from predicted contacts and secondary structure converted to distance, angle and h-bond restraints, and could yield TMscore >0.6 for 203 of 579 test proteins, while using MetaPSICOV and CCMpred could do so for 79 and 62, respectively.

### **9.5.1 How Many Predicted Contacts Should Be Used to Build 3D Models?**

The number of feasible contacts surrounding a residue in a protein is about 6.3 (Miyazawa and Jernigan 1996), which corresponds to the maximum number of contacts per a protein,  $6.3L/2$ , where  $L$  denotes protein length. However, more than 50% of known 3D structures in the PDB have less than  $2L$  contacts, and in the test on 15 proteins in EVfold benchmark set, less than  $1.6L$  predicted contacts yielded best results (Adhikari et al. 2015). In the original EVfold, the optimal number of evolutionary constraints was in the order of  $0.5L$  to  $0.7L$  (Hopf et al. 2012). Because prediction accuracy tends to decrease as the rank of contact score increases, and different proteins need different numbers of predicted contacts to be folded well, protein folds were generated with a wide range of the number of predicted contacts, and then best folds were selected; from 30 to  $L$  in EVfold (Hopf et al. 2012), and from  $0.4L$  to  $2.2L$  in CONFOLD (Adhikari et al. 2015). In RaptorX, the top  $2L$  predicted contacts irrespective of site separation were converted to distance restraints (Wang et al. 2017). On the other hand, Jones group reported (Kosciolek and Jones 2014) that artificially truncating the list of predicted contacts was likely to remove useful information to fold a protein with FRAGFOLD and PSICOV, in which the weight of a given predicted contact is determined by its positive predictive value.

## **9.6 Evolutionary Direct Couplings Between Residues Not Contacting in a Protein 3D Structure**

Needless to say, evolutionary constraints do not only originate in intra-molecular contacts but also result from inter-molecular contacts/interactions. Even in the case of intra-molecular contacts, if there are structural variations including ones due to conformational changes in a protein family, evolutionary constraints will reflect the alternative conformations (Morcos et al. 2011; Hopf et al. 2012; Anishchenko et al. 2013). Also, intra-molecular residue couplings may contain useful information of ligand-mediated residue couplings (Morcos et al. 2011; Ovchinnikov et al. 2016). On the other hand, inter-molecular contacts may allow us to predict protein complexes, and are useful to build protein-protein interaction networks at a residue level.

### **9.6.1 Structural Variation Including Conformational Changes**

MSA contains information on all members of the protein family, and direct couplings between residues estimated from the MSA reflect the structures of all

members. It was shown (Anishchenko et al. 2013) that 74% of top  $L/2$  direct couplings residue pairs that are more than 5 Å apart in the target structures of 3883 proteins are less than 5 Å apart in at least one homolog structure.

Conformational change is an interesting case of structural variation. Many proteins adopt different conformations as part of their functions (Tokuriki and Tawfik 2009), indicating that protein flexibility is as important as structure on biological function. Protein flexibility around the energy minimum can be studied by sampling around the native structure in normal mode/principal component analysis, coarse-grained elastic network model, and short-timescale MD simulations. However, distant conformers that require large conformational transitions are difficult to predict. If conformational changes are essential on protein functions, evolutionary constraints will reflect the multiple conformations. Toth-Petroczy et al. (2016) showed that coevolutionary information may reveal alternative structural states of disorderd regions.

Morcos et al. (2011) found that some of top predicted contacts in the response-regulator DNA-binding domain family (GerE, PF00196) conflict with the structure (PDB ID 3C3W) of the full-length response-regulator DosR of *M. tuberculosis*, but are compatible with the structure (PDB ID 1JE8) of DNA-binding domain of *E. coli* NarL.

Sutto et al. (2015) combined coevolutionary data and molecular dynamics simulations to study protein conformational heterogeneity; the Boltzmann-learning algorithm with  $\ell_2$  regularization terms was employed to extract direct couplings between sites in homologous protein sequences, and a set of conformations consistent with the observed residue couplings were generated by exhaustive sampling simulations based on a coarse-grained protein model. Although the most representative structure was consistent with the experimental fold, the various regions of the sequence showed different stability, indicating conformational changes (Sutto et al. 2015).

Sfriso et al. (2016) made an automated pipeline based on discrete molecular dynamics guided by predicted contacts for the systematic identification of functional conformations in proteins, and identified alternative conformers in 70 of 92 proteins in a validation set of proteins in PDB; various conformational transitions are relevant to those conformers, such as open-closed, rotation, rotation-closed, concerted, and miscellanea of complex motions.

### 9.6.2 *Homo-Oligomer Contacts*

Intra-molecular contacts that conflict with the native fold may indicate homo-oligomer contacts (Anishchenko et al. 2013). Such a case was confirmed for homo-oligomer contacts in the ATPase domain of nitrogen regulatory protein C-like sigma-54 dependent transcriptional activators (Morcos et al. 2011) and between transmembrane helices (Hopf et al. 2012). It was pointed out (Hopf et al. 2012) that

the identification of evolutionary couplings due to homo-oligomerization is not only meaningful in itself but also useful because their removal improves the accuracy of the structure prediction for the monomer.

### ***9.6.3 Residue Couplings Mediated by Binding to a Third Agent***

Direct couplings between residues found by the DCA analysis can be mediated (Morcos et al. 2011) by their interactions with a third agent, i.e., ligands, substrates, RNA, DNA, and other metabolites. This indicates that binding sites with such a agent may be found as residue sites directly coupled but not in contact.

If interactions with a third agent requires too specific residue type at a certain site, then the residue type will be well conserved at the binding sites. This often occurs, and has been utilized to identify binding sites. However, the interactions for binding are less specific but certainly restricted, direct couplings between residues around the binding sites may occurs.

Hopf et al. (2012) devised a total evolutionary coupling score, which is defined as EC values summed over all high-ranking pairs involving a given residue and normalized by their average over all high-ranking pairs, and showed that residues with high total coupling scores line substrate-binding sites and affect signaling or transport in transmembrane proteins, Adrb2 and Opsd.

## **9.7 Heterogeneous Protein-Protein Contacts**

An application of the direct coupling analysis to predict the structures of protein complexes is straightforward. In place of a MSA of a single protein family, a single MSA that is built by concatenating the multiple MSAs of multiple protein families every species can be employed to extract direct couplings between sites of different proteins by removing indirect intra- and inter-protein couplings (Pazos et al. 1997; Skerker et al. 2008; Weigt et al. 2009; Hopf et al. 2012).

A critical requirement for sequences to be concatenated is, however, that respective sets of the protein sequences must have the same evolutionary history to coevolve. In other words, phylogenetic trees built from the respective sets of sequences employed for the protein families must have at least the same topology. One way to build a set of cognate pairs of protein sequences is to employ orthologous sequences for each protein family, the phylogenetic tree of which coincides with that of species. Thus, a genome-wide analysis of finding protein-protein interactions based on protein sequences is not so simple.

Weigt et al. (2009) successfully applied the direct coupling analysis to the bacterial two-component signal transduction system consisting of sensor kinase (SK) and response regulator (RR), which are believed (Skerker et al. 2008) to interact specifically with each other in most cases and often revealed by adjacency

in chromosomal location. This analysis is based on the fact that in prokaryotes cognate pairs are often encoded in the same operon. Genome-sequencing projects have revealed that most organisms contain large expansions of a relatively small number of signaling families (Skerker et al. 2008). However, it is not as simple as in prokaryotes to build a set of cognate pairs of those protein sequences in eukaryotes.

Hopf et al. (2014) developed a contact score, EVcomplex, for every inter-protein residue pair based on the overall inter-protein EC score distributions, evaluated its performance in blinded tests on 76 complexes of known 3D structure, predicted protein-protein contacts in 32 complexes of unknown structure, and then demonstrated how evolutionary direct couplings can be used to distinguish between interacting and non-interacting protein pairs in a large complex. In their analysis, protein sequence pairs that are encoded close on *E. coli* genome were employed to reduce incorrect protein pairings.

## 9.8 Discussion

Determination of protein structure is essential to understand protein function. However, despite significant effort to explore unknown folds in the protein structural space, protein structures determined by experiment are far less than known protein families. Only about 41–42% of the Pfam families (Finn et al. 2016) (Pfam-A release 31.0, 16712 families) include at least one member whose structure is known. The number and also the size of protein families will further grow as genome/metagenome sequencing projects proceed with next-generation sequencing technologies. Thus, accurate de novo prediction of three-dimensional structure is desirable to catch up with the high growing speed of protein families with unknown folds. Coevolutionary information can be used to predict not only proteins but also RNAs (Weinreb et al. 2016) and those complexes, together with experimental informations such as X-ray, NMR, SAS, FRET, crosslinking, Cryo-EM, and others.

Here, statistical methods for disentangling direct from indirect couplings between sites with respect to evolutionary variations/substitutions of amino acids in homologous proteins have been briefly reviewed. Dramatic improvements on contact prediction and successful 3D de novo predictions based on predicted contacts are described in details in the recent reports of CASP-11 (Moult et al. 2016) and CASP-12 meetings (CASP12 2017). Machine learning methods, particularly deep neural network (DNN) such as MetaPSICOV, iFold, and RaptorX, have shown to significantly augment contact prediction accuracy based on coevolutionary information. However, the present state-of-the-art DNN methods are, at least at the very moment, not powerful enough to extract coevolutionary information directly from homologous sequences. It was reported that without coevolutionary strength produced by CCMpred the top  $L/10$  long-range prediction accuracy of RaptorX might drop by 0.15 for soluble proteins and more for membrane proteins (Wang et al. 2017), indicating that the direct coupling analysis is still essential for contact prediction.

The primary requirement for the direct coupling analysis is a high quality deep alignment. However, genome/metagenome sequencing projects provide more genetic variations from which more accurate and more comprehensive information on evolutionary constraints can be extracted. One of problems is that species being sequenced may be strongly biased to prokaryotes, making it hard to analyze eukaryotic proteins based on coevolutionary substitutions. Experiments of *in vitro* evolution may be useful to provide sequence variations for eukaryotic proteins (Ovchinnikov et al. 2016).

For a large-scale of protein structure prediction, computationally intensive methods such as the ACE and Boltzmann machine (MCMC and mpDCA) can hardly be employed. The Gaussian approximation with a normal-inverse-Wishart prior, the Gaussian approximations with other priors (PSICOV) and mean field approximation (mfDCA) are fast enough but their performance of contact prediction tends to be compared unfavorably with the pseudo-likelihood approximation (plmDCA), indicating that they may be inappropriate for proteins with sparse couplings.

The accurate estimates of fields and couplings are very informative in evaluating the effects ( $\Delta H_{\text{Potts}}$ ) of mutations (Hopf et al. 2017), identifying protein family members and also studying folding mechanisms (Morcos et al. 2014; Jacquin et al. 2016) and protein evolution (Miyazawa 2017b). It should be also examined whether the distribution of dimensionless energies ( $H_{\text{Potts}}$ ) over homologous proteins can be well reproduced. Accuracy of estimates of fields and couplings and the distribution of dimensionless energies depends on regularization parameters or the ratio of pseudocount (Barton et al. 2016; Miyazawa 2017b), and therefore they should be optimized. It was also pointed out that group  $L_1$  regularization performs better than  $L_2$  for the maximum pseudolikelihood method (Ingraham and Marks 2016). The ACE algorithm, which can be applied only for systems of sparse couplings, may be more favorable with respect to computational load for the estimation of fields and couplings than Boltzmann learning with Monte Carlo simulation or with message passing. However, both the methods are computationally intensive. Recently, another approach consisting of two methods named persistent-vi and Fadeout, in which the posterior probability density with horseshoe prior is approximately estimated by using variational inference and noncentered parameterization for such a sparsity-inducing prior, has shown to perform better with twofold cpu time than the maximum pseudolikelihood method with  $L_2$  and group  $L_1$  regularizations (Ingraham and Marks 2016).

The remarkable advances of sequencing technologies and also statistical methods are likely to bring many targets within range of the present approach in the near future, and have a potential to transform the field (Moult et al. 2016).

## Appendix

An appendix described in full will be found in the article (Miyazawa 2017a) submitted to the arXiv.



## Inverse Potts Model

### A Gauge Employed for $h_i(a_k)$ and $J_{ij}(a_k, a_l)$

Unless specified, a following gauge is employed; we call it  $q$ -gauge, here.

$$h_i(a_q) = J_{ij}(a_k, a_q) = J_{ij}(a_q, a_l) = 0 \quad (9.16)$$

In this gauge, the amino acid  $a_q$  is the reference state for fields and couplings, and  $P_i(a_q)$ ,  $P_{ij}(a_k, a_q) = P_{ji}(a_q, a_k)$ , and  $P_{ij}(a_q, a_q)$  are regarded as dependent variables. Common choices for the reference state  $a_q$  are the most common (consensus) state at each site. Any gauge can be transformed to another by the following transformation.

$$J_{ij}^I(a_k, a_l) \equiv J_{ij}(a_k, a_l) - J_{ij}(\cdot, a_l) - J_{ij}(a_k, \cdot) + J_{ij}(\cdot, \cdot) \quad (9.17)$$

$$h_i^I(a_k) \equiv h_i(a_k) - h_i(\cdot) + \sum_{j \neq i} (J_{ij}(a_k, \cdot) - J_{ij}(\cdot, \cdot)) \quad (9.18)$$

where “ $\cdot$ ” denotes the reference state, which may be  $a_q$  for each site ( $q$ -gauge) or the average over all states (Ising gauge).

## Boltzmann Machine

Fields  $h_i(a_k)$  and couplings  $J_{ij}(a_k, a_l)$  are estimated by iterating the following 2-step procedures.

1. For a given set of  $h_i$  and  $J_{ij}(a_k, a_l)$ , marginal probabilities,  $P^{\text{MC}}(\sigma_i = a_k)$  and  $P^{\text{MC}}(\sigma_i = a_k, \sigma_i = a_l)$ , are estimated by a Markov chain Monte Carlo method (the Metropolis-Hastings algorithm (Metropolis et al. 1953)) or by any other method (for example, the message passing algorithm (Weigt et al. 2009)).
2. Then,  $h_i$  and  $J_{ij}(a_k, a_l)$  are updated according to the gradient of negative log-posterior-probability per instance,  $\partial S_0 / \partial h_i(a_k)$  or  $\partial S_0 / \partial J_{ij}(a_k, a_l)$ , multiplied by a parameter-specific weight factor (Barton et al. 2016),  $w_i(a_k)$  or  $w_{ij}(a_k, a_l)$ ; see Eqs. 9.8 and 9.12.

$$\Delta h_i(a_k) = -(P^{\text{MC}}(\sigma_i = a_k) + \frac{\partial R}{\partial h_i(a_k)} - P_i(a_k)) \cdot w_i(a_k) \quad (9.19)$$

$$\begin{aligned} \Delta J_{ij}(a_k, a_l) &= -(P^{\text{MC}}(\sigma_i = a_k, \sigma_i = a_l) + \frac{\partial R}{\partial J_{ij}(a_k, a_l)} \\ &\quad - P_{ij}(a_k, a_l)) \cdot w_{ij}(a_k, a_l) \end{aligned} \quad (9.20)$$

where weights are also updated as  $w_i(a_k) \leftarrow f(w_i(a_k))$  and  $w_{ij}(a_k, a_l) \leftarrow f(w_{ij}(a_k, a_l))$  according to the RPROP (Riedmiller and Braun 1993) algorithm; the function  $f(w)$  is defined as

$$f(w) \equiv \begin{cases} \max(w \cdot s_-, w_{\min}) & \text{if the gradient changes its sign,} \\ \min(w \cdot s_+, w_{\max}) & \text{otherwise} \end{cases} \quad (9.21)$$

$w_{\min} = 10^{-3}$ ,  $w_{\max} = 10$ ,  $s_- = 0.5$ , and  $s_+ = 1.9 < 1/s_-$  were employed (Barton et al. 2016). After updated,  $h_i(a_k)$  and  $J_{ij}(a_k, a_l)$  may be modified to satisfy a given gauge.

The Boltzmann machine has a merit that model correlations are calculated.

### Gaussian Approximation for $P(\sigma)$ with a Normal-Inverse-Wishart Prior

The normal-inverse-Wishart distribution (NIW) is the product of the multivariate normal distribution ( $\mathcal{N}$ ) and the inverse-Wishart distribution ( $\mathcal{W}^{-1}$ ), which are the conjugate priors for the mean vector and for the covariance matrix of a multivariate Gaussian distribution, respectively. The NIW is employed as a prior in GaussDCA (Baldassi et al. 2014), in which the sequence distribution  $P(\sigma)$  is approximated as a Gaussian distribution. In this approximation, the q-gauge is used, and  $P_i(a_q)$ ,  $P_{ij}(a_k, a_q) = P_{ji}(a_q, a_k)$ , and  $P_{ij}(a_q, a_q)$  are regarded as dependent variables; see section “A Gauge Employed for  $h_i(a_k)$  and  $J_{ij}(a_k, a_l)$ ”; in GaussDCA, deletion is excluded from independent variables.

The posterior distribution for the NIW is also a NIW. Thus, the cross entropy  $S_0$  can be represented as

$$S_0(\boldsymbol{\mu}, \Sigma | \{P_i\}, \{P_{ij}\}) = \frac{-1}{B} \log \left[ \prod_{\tau=1}^B \mathcal{N}(\{\delta_{\sigma_i^\tau a_k}\} | \boldsymbol{\mu}, \Sigma) \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}^0, \Sigma/\kappa) \mathcal{W}^{-1}(\Sigma | \Lambda, \nu) \right] \quad (9.22)$$

$$= \frac{-1}{B} \log [\mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}^0, \Sigma/\kappa) \mathcal{W}^{-1}(\Sigma | \Lambda^B, \nu^B)] \quad (9.23)$$

$$(\det(2\pi \Sigma))^{-B/2} \left(\frac{\kappa}{\kappa^B}\right)^{\dim \Sigma/2} \frac{(\det(\Lambda/2))^{\nu/2}}{(\det(\Lambda^B/2))^{\nu^B/2}} \frac{\Gamma_{\dim \Sigma}(\nu^B/2)}{\Gamma_{\dim \Sigma}(\nu/2)} (\det \Sigma)^{-(\nu-\nu^B)2} \quad (9.24)$$

where  $\Gamma_{\dim \Sigma}(\nu/2)$  is the multivariate  $\Gamma$  function,  $\boldsymbol{\mu}$  is the mean vector, and  $\dim \Sigma$  is the dimension of covariance matrix  $\Sigma$ ,  $\dim \Sigma = (q-1)L$  excluding deletion in GaussDCA. The normal and NIW distributions are defined as follows.

$$\mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}^0, \Sigma) \equiv (\det(2\pi \Sigma))^{-1/2} \exp\left(-\frac{(\boldsymbol{\mu} - \boldsymbol{\mu}^0)^T \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}^0)}{2}\right) \quad (9.25)$$

$$\mathcal{W}^{-1}(\Sigma|\Lambda, \nu) \equiv \frac{(\det(\Lambda/2))^{v/2}}{\Gamma_{\dim \Sigma}(\nu/2)} (\det \Sigma)^{-(v+\dim \Sigma+1)/2} \exp\left(-\frac{1}{2} \text{Tr} \Lambda \Sigma^{-1}\right) \quad (9.26)$$

Parameters  $\boldsymbol{\mu}^B$ ,  $\kappa^B$ ,  $\nu^B$ , and  $\Lambda^B$  satisfy

$$\mu_i^B(a_k) = (\kappa \mu_i^0(a_k) + B P_i(a_k)) / (\kappa + B), \quad \kappa^B = \kappa + B, \quad \nu^B = \nu + B \quad (9.27)$$

$$\begin{aligned} \Lambda_{ij}^B(a_k, a_l) &= \Lambda_{ij}(a_k, a_l) + B C_{ij}(a_k, a_l) \\ &+ \frac{\kappa B}{\kappa + B} [(P_i(a_k) - \mu_i^0(a_k))(P_j(a_l) - \mu_j^0(a_l))] \end{aligned} \quad (9.28)$$

where the  $\Lambda$  and  $\nu$  are the scale matrix and the degree of freedom, respectively, shaping the inverse-Wishart distribution, and  $C$  is the given covariance matrix;  $C_{ij}(a_k, a_l) \equiv P_{ij}(a_k, a_l) - P_i(a_k)P_j(a_l)$ . The mean values of  $\boldsymbol{\mu}$  and  $\Sigma$  under NW posterior are  $\boldsymbol{\mu}^B$  and  $\Lambda^B/(\nu^B - \dim \Sigma - 1)$ , and their mode values are  $\boldsymbol{\mu}^B$  and  $\Lambda^B/(\nu^B + \dim \Sigma + 1)$ , which minimize the cross entropy or maximize the posterior probability. The covariance matrix  $\Sigma$  can be estimated to be the exactly same value by adjusting the value of  $\nu$ , whichever the mean posterior or the maximum posterior is employed for the estimation of  $\Sigma$ . In GaussDCA, the mean posterior estimate was employed but here the maximum posterior estimate is employed according to the present formalism.

$$(\boldsymbol{\mu}, \Sigma) = \arg \min_{(\boldsymbol{\mu}, \Sigma)} S_0(\boldsymbol{\mu}, \Sigma | \{P_i\}, \{P_{ij}\}) = (\boldsymbol{\mu}^B, \Lambda^B/(\nu^B + \dim \Sigma + 1)) \quad (9.29)$$

According to GaussDCA,  $\nu$  is chosen in such a way that  $\sigma_{ij}(a_k, a_l)$  is nearly equal to the covariance matrix corrected by pseudocount;  $\nu = \kappa + \dim \Sigma + 1$  for the mean posterior estimate in GaussDCA, but  $\nu = \kappa - \dim \Sigma - 1$  for the maximum posterior estimate here.

From Eq. 9.15, the estimates of couplings and fields are calculated.

$$J_{ij}^{\text{NIW}}(a_k, a_l) = -\frac{\partial S_0(\{P_i\}, \{P_{ij}\})}{\partial P_{ij}(a_k, a_l)} = -\frac{(\kappa + B + 1)}{\kappa + B} (\Sigma^{-1})_{ij}(a_k, a_l) \quad (9.30)$$

Because the number of instances is far greater than 1 ( $B \gg 1$ ), these estimates of couplings are practically equal to the estimates ( $J^{\text{MF}} = -\Sigma^{-1}$ ) in the mean field approximation, which was employed in GaussDCA (Baldassi et al. 2014).

$$\begin{aligned} h_i^{\text{NIW}}(a_k) &= -\sum_{j \neq i} \sum_l J_{ij}^{\text{NIW}}(a_k, a_l) P_j(a_l) - \frac{(\kappa + B + 1)}{\kappa + B} \sum_j \sum_{l \neq q} (\Sigma^{-1})_{ij}(a_k, a_l) \\ &[\delta_{ij} \frac{\delta_{kl} - 2P_l(a_l)}{2} + \frac{\kappa B}{\kappa + B} (P_j(a_l) - \mu_j^0(a_l))] \end{aligned} \quad (9.31)$$

The  $(h_i^{\text{NIW}}(a_k) - h_i^{\text{NIW}}(a_q))$  does not converge to  $\log P_i(a_k)/P_i(a_q)$  as  $J^{\text{NIW}} \rightarrow 0$  but  $h_i^{\text{MF}}(a_k) - h_i^{\text{MF}}(a_q)$  does; in other words, the mean field approximation gives a better  $h$  for the limiting case of no couplings than the present approximation. Barton et al. (2016) reported that the Gaussian approximation generally gave a better generative model than the mean field approximation.

In GaussDCA (Baldassi et al. 2014),  $\mu^0$  and  $\Lambda/\kappa$  were chosen to be as uninformative as possible, i.e., mean and covariance for a uniform distribution.

$$\mu_i^0(a_k) = 1/q, \quad \frac{\Lambda_{ij}(a_k, a_l)}{\kappa} = \frac{\delta_{ij}}{q} (\delta_{kl} - \frac{1}{q}) \quad (9.32)$$

## Pseudo-likelihood Approximation

### Symmetric Pseudo-likelihood Maximization

The probability of an instance  $\sigma^\tau$  is approximated as follows by the product of conditional probabilities of observing  $\sigma_i^\tau$  under the given observations  $\sigma_{j \neq i}^\tau$  of all other sites.

$$P(\sigma^\tau) \approx \prod_i P(\sigma_i = \sigma_i^\tau | \{\sigma_{j \neq i} = \sigma_j^\tau\}) \quad (9.33)$$

Then, cross entropy is approximated as

$$S_0(h, J | \{P_i\}, \{P_{ij}\}) \approx S_0^{\text{PLM}}(h, J | \{P_i\}, \{P_{ij}\}) \equiv \sum_i S_{0,i}(h, J | \{P_i\}, \{P_{ij}\}) \quad (9.34)$$

$$S_{0,i}(h, J | \{P_i\}, \{P_{ij}\}) \equiv \frac{-1}{B} \sum_{\tau} \ell_i(\sigma_i = \sigma_i^\tau | \{\sigma_{j \neq i} = \sigma_j^\tau\}, h, J) + R_i(h, J) \quad (9.35)$$

where conditional log-likelihoods and  $\ell_2$  norm regularization terms employed in Ekeberg et al. (2013) are

$$\ell_i(\sigma_i = \sigma_i^\tau | \{\sigma_{j \neq i} = \sigma_j^\tau\}, h, J) = \log \left[ \frac{\exp(h_i(\sigma_i^\tau) + \sum_{j \neq i} J_{ij}(\sigma_i^\tau, \sigma_j^\tau))}{\sum_k \exp(h_i(a_k) + \sum_{j \neq i} J_{ij}(a_k, \sigma_j^\tau))} \right] \quad (9.36)$$

$$R_i(h, J) \equiv \gamma_h \sum_k h_i(a_k)^2 + \frac{\gamma_J}{2} \sum_k \sum_{j \neq i} \sum_l J_{ij}(a_k, a_l)^2 \quad (9.37)$$

The optimum fields and couplings in this approximation are estimated by minimizing the pseudo-cross-entropy,  $S_0^{\text{PLM}}$ .

$$(h^{\text{PLM}}, J^{\text{PLM}}) = \arg \min_{h, J} S_0^{\text{PLM}}(h, J | \{P_i\}, \{P_{ij}\}) \quad (9.38)$$

Equation 9.38 is not invariant under gauge transformation; the  $\ell_2$  norm regularization terms in Eq.9.38 favors only a specific gauge that corresponds to  $\gamma_J \sum_l J_{ij}(a_k, a_l) = \gamma_h h_i(a_k)$ ,  $\gamma_J \sum_k J_{ij}(a_k, a_l) = \gamma_h h_j(a_l)$ , and  $\sum_k h_i(a_k) = 0$  for all  $i, j (> i), k$  and  $l$  (Ekeberg et al. 2013).  $\gamma_J = \gamma_h = 0.01$  that is relatively a large value independent of  $B$  was employed in Ekeberg et al. (2013).  $\gamma_h = 0.01$  but  $\gamma_J = q(L-1)\gamma_h$  were employed in Hopf et al. (2017), in which gapped sites in each sequence were excluded in the calculation of the Hamiltonian  $H(\sigma)$ , and therefore  $q = 20$ .

GREMLIN (Kamisetty et al. 2013) employs Gaussian prior probabilities that depend on site pairs.

$$R_i(h, J) \equiv \gamma_h \sum_k h_i(a_k)^2 + \sum_k \sum_{j \neq i} \frac{\gamma_{ij}}{2} \sum_l J_{ij}(a_k, a_l)^2 \quad (9.39)$$

$$\gamma_{ij} \equiv \gamma_c (1 - \gamma_p \log(P_{ij}^0)) \quad (9.40)$$

where  $P_{ij}^0$  is the prior probability of site pair  $(i, j)$  being in contact.

### Asymmetric Pseudo-likelihood Maximization

To speed up the minimization of  $S_0$ , a further approximation, in which  $S_{0,i}$  is separately minimized, is employed (Ekeberg et al. 2014), and fields and couplings are estimated as follows.

$$J_{ij}^{\text{PLM}}(a_k, a_l) \simeq \frac{1}{2} (J_{ij}^*(a_k, a_l) + J_{ji}^*(a_l, a_k)) \quad (9.41)$$

$$(h_i^{\text{PLM}}, J_i^*) = \arg \min_{h_i, J_i} S_{0,i}(h, J | \{P_i\}, \{P_{ij}\}) \quad (9.42)$$

It is appropriate to transform  $h$  and  $J$  estimated above into a some specific gauge such as the Ising gauge.

### ACE (Adaptive Cluster Expansion) of Cross-Entropy for Sparse Markov Random Field

The cross entropy  $S_0(\{h_i, J_{ij}\} | \{P_i\}, \{P_{ij}\}, i, j \in \Gamma)$  of a cluster of sites  $\Gamma$ , which is defined as the negative log-likelihood per instance in Eq.9.14, is approximately minimized by taking account of sets  $L_k(t)$  of only significant clusters consisting of

$k$  sites, the incremental entropy (cluster cross entropy)  $\Delta S_\Gamma$  of which is significant ( $|\Delta S_\Gamma| > t$ ) (Cocco and Monasson 2011, 2012; Barton et al. 2016).

$$S_0(\{P_i, P_{ij}|i, j \in \Gamma\}) \simeq \sum_{l=1}^{|\Gamma|}, \sum_{\Gamma' \in L_l(t), \Gamma' \subset \Gamma} \Delta S_0(\{P_i, P_{ij}|i, j \in \Gamma'\}) \quad (9.43)$$

$$\Delta S_0(\{P_i, P_{ij}|i, j \in \Gamma\}) \equiv S_0(\{P_i, P_{ij}|i, j \in \Gamma\}) - \sum_{\Gamma' \subset \Gamma} \Delta S_0(\{P_i, P_{ij}|i, j \in \Gamma'\}) \quad (9.44)$$

$$= \sum_{\Gamma' \subset \Gamma} (-1)^{|\Gamma| - |\Gamma'|} S_0(\{P_i, P_{ij}|i, j \in \Gamma'\}) \quad (9.45)$$

$L_{k+1}(t)$  is constructed from  $L_k(t)$  by adding a cluster  $\Gamma$  consisting of  $(k+1)$  sites in a lax case provided that any pair of size  $k$  clusters  $\Gamma^1, \Gamma^2 \in L_k(t)$  and  $\Gamma^1 \cup \Gamma^2 = \Gamma$  or in a strict case if  $\Gamma' \in L_k(t)$  for  $\forall \Gamma'$  such that  $\Gamma' \subset \Gamma$  and  $|\Gamma'| = k$ . Thus, Eq. 9.43 yields sparse solutions. The cross entropies  $S_0(\{P_i, P_{ij}|i, j \in \Gamma'\})$  for the small size of clusters are estimated by minimizing  $S_0(\{h_i, J_{ij}\}|\{P_i, P_{ij}\}, i, j \in \Gamma')$  with respect to fields and couplings. Starting from a large value of the threshold  $t$  (typically  $t = 1$ ), the cross-entropy  $S_0(\{P_i, P_{ij}\}|i, j \in \{1, \dots, N\})$  is calculated by gradually decreasing  $t$  until its value converges. Convergence of the algorithm may also be more difficult for alignments of long proteins or those with very strong interactions. In such cases, strong regularization may be employed.

The following regularization terms of  $\ell_2$  norm are employed in ACE (Barton et al. 2016), and so Eq. 9.43 is not invariant under gauge transformation.

$$-\frac{1}{B} \log P_0(h, J|i, j \in \Gamma) = \gamma_h \sum_{i \in \Gamma} \sum_k h_i(a_k)^2 + \gamma_J \sum_{i \in \Gamma} \sum_k \sum_{J > i, j \in \Gamma} \sum_l J_{ij}(a_k, a_l)^2 \quad (9.46)$$

$\gamma_h = \gamma_J \propto 1/B$  was employed (Barton et al. 2016).

The compression of the number of Potts states,  $q_i \leq q$ , at each site can be taken into account. All infrequently observed states or states that insignificantly contribute to site entropy can be treated as the same state, and a complete model can be recovered (Barton et al. 2016) by setting  $h_i(a_k) = h_i(a_{k'}) + \log(P_i(a_k)/P_i(a_{k'}))$ , and  $J_{ij}(a_k, a_l) = J'_{ij}(a_{k'}, a_{l'})$ , where “ $r$ ” denotes a corresponding aggregated state and a potential.

Starting from the output set of the fields  $h_i(a_k)$  and couplings  $J_{ij}(a_k, a_l)$  obtained from the cluster expansion of the cross-entropy, a Boltzmann machine is trained with  $P_i(a_k)$  and  $P_{ij}(a_k)$  by the RPROP algorithm (Riedmiller and Braun 1993) to refine the parameter values of  $h_i$  and  $J_{ij}(a_k, a_l)$  (Barton et al. 2016); see section “Boltzmann Machine”. This post-processing is also useful because model correlations are calculated.

An appropriate value of the regularization parameter for trypsin inhibitor were much larger ( $\gamma = 1$ ) for contact prediction than those ( $\gamma = 2/B = 10^{-3}$ ) for

recovering true fields and couplings (Barton et al. 2016), probably because the task of contact prediction requires the relative ranking of interactions rather than their actual values.

## Scoring Methods for Contact Prediction

Corrected Frobenius Norm ( $L_{22}$  Matrix Norm),  $S_{ij}^{\text{CFN}}$

For scoring, plmDCA (Ekeberg et al. 2013, 2014) employs the corrected Frobenius norm of  $J_{ij}^1$  transformed in the Ising gauge, in which  $J_{ij}^1$  does not contain anything that could have been explained by fields  $h_i$  and  $h_j$ ;  $J_{ij}^1(a_k, a_l) \equiv J_{ij}(a_k, a_l) - J_{ij}(\cdot, a_l) - J_{ij}(a_k, \cdot) + J_{ij}(\cdot, \cdot)$  where  $J_{ij}(\cdot, a_l) = J_{ji}(a_l, \cdot) \equiv \sum_{k=1}^q J_{ij}(a_k, a_l)/q$ .

$$S_{ij}^{\text{CFN}} \equiv S_{ij}^{\text{FN}} - S_{\cdot j}^{\text{FN}} S_{i \cdot}^{\text{FN}} / S_{\cdot \cdot}^{\text{FN}}, \quad S_{ij}^{\text{FN}} \equiv \sqrt{\sum_{\kappa \neq \text{gap}} \sum_{l \neq \text{gap}} J_{ij}^1(a_k, a_l)^2} \quad (9.47)$$

where “ $\cdot$ ” denotes average over the indicated variable. This CFN score with the gap state excluded in Eq. 9.47 performs better (Ekeberg et al. 2014; Baldassi et al. 2014) than both scores of FN and DI/EC (Weigt et al. 2009; Morcos et al. 2011; Marks et al. 2011; Hopf et al. 2012).

## References

- Adhikari B, Bhattacharya D, Cao R, Cheng J (2015) CONFOLD: residue-residue contact-guided ab initio protein folding. *Proteins* 83:1436–1449. <https://doi.org/10.1002/prot.24829>
- Adhikari B, Nowotny J, Bhattacharya D, Hou J, Cheng J (2016) ConEVA: a toolbox for comprehensive assessment of protein contacts. *BMC Bioinf* 17:517. <https://doi.org/10.1186/s12859-016-1404-z>
- Altschuh D, Vernet T, Berti P, Moras D, Nagai K (1988) Coordinated amino acid changes in homologous protein families. *Protein Eng* 2:193–199
- Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D (2013) Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci USA* 114:9122–9127. <https://doi.org/10.1073/pnas.1702664114>
- Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol* 17:164–178
- Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. *Proteins* 79:1061–1078. <https://doi.org/10.1002/prot.22934>
- Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A (2014) Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS ONE* 9(3):e92721. <https://doi.org/10.1371/journal.pone.0092721>

- Barton JP, Leonardis ED, Coucke A, Cocco S (2016) ACE: adaptive cluster expansion for maximum entropy graphical model inference. *Bioinformatics* 32:3089–3097. <https://doi.org/10.1093/bioinformatics/btw328>
- Braun W, Go N (1985) Calculation of protein conformations by proton-proton distance constraints: a new efficient algorithm. *J Mol Biol* 186:611–626. [https://doi.org/10.1016/0022-2836\(85\)90134-2](https://doi.org/10.1016/0022-2836(85)90134-2)
- Brünger AT (2007) Version 1.2 of the crystallography and NMR system. *Nat Protoc* 2:2728–2733. <https://doi.org/10.1038/nprot.2007.406>
- Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* 4:165
- Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput Biol* 6(1):e1000633. <https://doi.org/10.1371/journal.pcbi.1000633>
- CASP12 (2017) 12th community wide experiment on the critical assessment of techniques of protein structure prediction. <http://predictioncenter.org/casp12/>
- Cocco S, Monasson R (2011) Adaptive cluster expansion for inferring Boltzmann machines with noisy data. *Phys Rev Lett* 106:090601. <https://doi.org/10.1103/PhysRevLett.106.090601>
- Cocco S, Monasson R (2012) Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. *J Stat Phys* 147:252–314. <https://doi.org/10.1007/s10955-012-0463-4>
- Cocco S, Feinauer C, Figliuzzi M, Monasson R, Weigt M (2017) Inverse statistical physics of protein sequences: a key issues review. arXiv:1703.01222 [q-bio.BM]
- Doron-Faigenboim A, Pupko T (2007) A combined empirical and mechanistic codon model. *Mol Biol Evol* 24:388–397
- Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24:333–340
- Dutheil J (2012) Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Brief Bioinform* 13:228–243
- Dutheil J, Galtier N (2007) Detecting groups of coevolving positions in a molecule: a clustering approach. *BMC Evol Biol* 7:242
- Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005) A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol* 22:1919–1928
- Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E* 87:012707–1–16. <https://doi.org/10.1103/PhysRevE.87.012707>
- Ekeberg M, Hartonen T, Aurell E (2014) Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J Comput Phys* 276:341–356
- Fares M, Travers S (2006) A novel method for detecting intramolecular coevolution. *Genetics* 173:9–23
- Fariselli P, Olmea O, Valencia A, Casadio R (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Eng* 14:835–843
- Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. *Nucl Acid Res* 44:D279–D285. <https://doi.org/10.1093/nar/gkv1344>
- Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet* 4:579–593
- Fleishman SJ, Yifrach O, Ben-Tal N (2004) An evolutionarily conserved network of amino acids mediates gating in voltage-dependent potassium channels. *J Mol Biol* 340:307–318
- Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignment. *Proteins* 56:211–221
- Giraud BG, Heumann JM, Lapedes AS (1999) Superadditive correlation. *Phys Rev E* 59:4973–4991



- Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18:309–317
- Gulyás-Kovács A (2012) Integrated analysis of residue coevolution and protein structure in ABC transporters. *PLoS ONE* 7(5):e36546. <https://doi.org/10.1371/journal.pone.0036546>
- Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. *Cell* 138:774–786
- Havel TF, Kuntz ID, Crippen GM (1983) The combinatorial distance geometry method for the calculation of molecular conformation. I. A new approach to an old problem. *J Theor Biol* 104:359–381
- Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149:1607–1621. <https://doi.org/10.1016/j.cell.2012.04.012>
- Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Bonvin, AMJJ, Sander C, Marks DS (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3:e03430. <https://doi.org/10.7554/eLife.03430>
- Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS (2017) Mutation effects predicted from sequence co-variation. *Nature Biotech* 35:128–135. <https://doi.org/10.1038/nbt.3769>
- Ingraham J, Marks D (2016) Variational inference for sparse and undirected models. *arXiv:1602.03807 [stat.ML]*
- Jacquin H, Gilson A, Shakhnovich E, Cocco S, Monasson R (2016) Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. *PLoS Comput Biol* 12:e1004889. <https://doi.org/10.1371/journal.pcbi.1004889>
- Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinf* 11:431
- Jones DT (2001) Predicting novel protein folds by using FRAGFOLD. *Proteins* 45(S5):127–132
- Jones DT, Bryson K, Coleman A, McGuffin LJ, Sadowski MI, Sodhi JS, Ward JJ (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins* 61(S7):143–151. <https://doi.org/10.1002/prot.20731>
- Jones DT, Buchan DWA, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28:184–190. <https://doi.org/10.1093/bioinformatics/btr638>
- Jones DT, Singh T, Kosciolk T, Tetcher S (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* 31:999–1006. <https://doi.org/10.1093/bioinformatics/btu791>
- Kaján L, Hopf TA, Kalaš M, Marks DS, Rost B (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinf* 15:85
- Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence-and structure-rich era. *Proc Natl Acad Sci USA* 110:15674–15679. <https://doi.org/10.1073/pnas.1314045110>
- Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Rosetta server. *Nucl Acid Res* 32:W526–W531
- Kim DE, Blum B, Bradley P, Baker D (2009) Sampling bottlenecks in *de novo* protein structure prediction. *J Mol Biol* 393:249–260
- Kosciolk T, Jones DT (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS ONE* 9:e92197. <https://doi.org/10.1371/journal.pone.0092197>
- Kosciolk T, Jones DT (2016) Accurate contact predictions using covariation techniques and machine learning. *Proteins* 84(S1):145–151. <https://doi.org/10.1002/prot.24863>
- Lapedes AS, Giraud BG, Liu LC, Stormo GD (1999) Correlated mutations in protein sequences: phylogenetic and structural effects. In: Seillier-Moiseiwitsch F (ed) *IMS lecture notes: statistics in molecular biology and genetics: selected proceedings of the joint AMS-IMS-SIAM summer conference on statistics in molecular biology, 22–26 June 1997*, pp 345–352. Institute of Mathematical Statistics

- Lapedes A, Giraud B, Jarzynsk C (2002) Using sequence alignments to predict protein structure and stability with high accuracy. LANL Science Magazine LA-UR-02-4481
- Lapedes A, Giraud B, Jarzynsk C (2012) Using sequence alignments to predict protein structure and stability with high accuracy. arXiv:1207.2484 [q-bio.QM]
- Maisnier-Patin S, Andersson DI (2004) Adaptation to the deleterious effect of antimicrobial drug resistance mutations by compensatory evolution. *Res Microbiol* 155:360–369
- Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6(12):e28766. <https://doi.org/10.1371/journal.pone.0028766>
- Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotech* 30:1072–1080. <https://doi.org/10.1038/nbt.2419>
- Martin LC, Gloor GB, Dunn SD, Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21:4116–4124
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087–1092
- Miyazawa S (2013) Prediction of contact residue pairs based on co-substitution between sites in protein structures. *PLoS ONE* 8(1):e54252. <https://doi.org/10.1371/journal.pone.0054252>
- Miyazawa S (2017a) Prediction of structures and interactions from genome information. arXiv:1709.08021 [q-bio.BM]
- Miyazawa S (2017b) Selection originating from protein stability/foldability: relationships between protein folding free energy, sequence ensemble, and fitness. *J Theor Biol* 433:21–38. <https://doi.org/10.1016/j.jtbi.2017.08.018>
- Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J Mol Biol* 256:623–644. <https://doi.org/10.1006/jmbi.1996.0114>
- Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108:E1293–E1301. <https://doi.org/10.1073/pnas.11114711108>
- Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG (2014) Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc Natl Acad Sci USA* 111:12408–12413. <https://doi.org/10.1073/pnas.1413575111>
- Moult J, Fidelis K, Krysztafowych A, Schwede T, Tramontano A (2016) Critical assessment of methods of protein structure prediction: progress and new directions in round XI. *Proteins* 84(S1):4–14. <https://doi.org/10.1002/prot.25064>
- Nugent T, Jones DT (2012) Accurate *de novo* structure prediction of large transmembrane protein domains using fragment assembly and correlated mutation analysis. *Proc Natl Acad Sci USA* 109:E1540–E1547. <https://doi.org/10.1073/pnas.1120036109>
- Ovchinnikov S, Kim DE, Wang RYR, Liu Y, DiMaio F, Baker D (2016) Improved *de novo* structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins* 84(S1):67–75. <https://doi.org/10.1002/prot.24974>
- Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. *J Mol Biol* 271:511–523
- Pollock DD, Taylor WR (1997) Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Protein Eng* 10:647–657
- Pollock DD, Taylor WR, Goldman N (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol* 287:187–198
- Poon AFY, Lewis FI, Frost SDW, Kosakovsky Pond SL (2008) Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. *Bioinformatics* 24:1949–1950
- Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 9:173–175
- Riedmiller M, Braun H (1993) A direct adaptive method for faster backpropagation learning: the RPROP algorithm. *IEEE Int Conf Neural Netw* 1993:586–591

- Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R (2005) Natural-like function in artificial WW domains. *Nature* 437:579–583
- Seemayer S, Gruber M, Söding J (2014) CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* 30:3128–3130. <https://doi.org/10.1093/bioinformatics/btu500>
- Sfriso P, Duran-Frigola M, Mosca R, Emperador A, Aloy P, Orozco M (2016) Residues coevolution guides the systematic identification of alternative functional conformations in proteins. *Structure* 24:116–126. <https://doi.org/10.1016/j.str.2015.10.025>
- Shendure J, Ji H (2017) EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction. *BMC Bioinf* 18:303. <https://doi.org/10.1186/s12859-017-1713-x>
- Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* 7:349–358
- Skerker JM, Perchuk BS, Siryapom A, Lubin EA, Ashenberg O, Goulian M, Laub MT (2008) Rewiring the specificity of two-component signal transduction systems. *Cell* 133:1043–1054
- Skwark MJ, Abdel-Rehim A, Elofsson A (2013) PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics* 29:1815–1816
- Skwark MJ, Raimondi D, Michel M, Elofsson A (2014) Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput Biol* 10:e1003889. <https://doi.org/10.1371/journal.pcbi.1003889>
- Skwark MJ, Michel M, Hurtado DM, Ekeberg M, Elofsson A (2016) Accurate contact predictions for thousands of protein families using PconsC3. *bioRxiv*. <https://doi.org/10.1101/079673>
- Sufkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci USA* 109:10340–10345. <https://doi.org/10.1073/pnas.1207864109>
- Sutto L, Marsili S, Valencia A, Gervasio FL (2015) From residue coevolution to protein conformational ensembles and functional dynamics. *Proc Natl Acad Sci USA* 112:13567–13572. <https://doi.org/10.1073/pnas.1508584112>
- Talavera D, Lovell SC, Whelan S (2015) Covariation is a poor measure of molecular coevolution. *Mol Biol Evol* 32:2456–2468. <https://doi.org/10.1093/molbev/msv109>
- Taylor WR, Sadowski MI (2011) Structural constraints on the covariance matrix derived from multiple aligned protein sequences. *PLoS ONE* 6(12):e28265. <https://doi.org/10.1371/journal.pone.0028265>
- Tokuriki N, Tawfik DS (2009) Protein dynamism and evolvability. *Science* 324:203–207
- Toth-Petroczy A, Palmado P, Ingraham J, Hopf TA, Berger B, Sander C, Marks DS (2016) Structured states of disordered proteins from genomic sequences. *Cell* 167:158–170. <https://doi.org/10.1016/j.cell.2016.09.010>
- Tufféry P, Darlu P (2000) Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Mol Biol Evol* 17:1753–1759
- Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate *de novo* prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 13:e1004324. <https://doi.org/10.1371/journal.pcbi.1005324>
- Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci USA* 106:67–72. <https://doi.org/10.1073/pnas.0805923106>
- Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell* 165:1–13. <https://doi.org/10.1016/j.cell.2016.03.030>
- Wuyun Q, Zheng W, Peng Z, Yang J (2016) A large-scale comparative assessment of methods for residue-residue contact prediction. *Brief Bioinform* 19:219–230. <https://doi.org/10.1093/bib/bbw106>
- Yanovsky C, Hom V, Thorpe D Protein structure relationships revealed by mutation analysis. *Science* 146:1593–1594 (1964)

# Chapter 10

## A Hybrid Approach for Protein Structure Determination Combining Sparse NMR with Evolutionary Coupling Sequence Data



Yuanpeng Janet Huang, Kelly P. Brock, Chris Sander, Debora S. Marks,  
and Gaetano T. Montelione

**Abstract** While 3D structure determination of small (<15 kDa) proteins by solution NMR is largely automated and routine, structural analysis of larger proteins is more challenging. An emerging hybrid strategy for modeling protein structures combines sparse NMR data that can be obtained for larger proteins with sequence co-variation data, called evolutionary couplings (ECs), obtained from multiple sequence alignments of protein families. This hybrid “EC-NMR” method can be used to accurately model larger (15–60 kDa) proteins, and more rapidly determine structures of smaller (5–15 kDa) proteins using only backbone NMR data. The resulting structures have accuracies relative to reference structures comparable to those obtained with full backbone and sidechain NMR resonance assignments. The requirement that evolutionary couplings (ECs) are consistent with NMR data recorded on a specific member of a protein family, under specific conditions, potentially also allows identification of ECs that reflect alternative allosteric or excited states of the protein structure.

---

Y. J. Huang · G. T. Montelione (✉)

Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers, The State University of New Jersey, Piscataway, NJ, USA  
e-mail: [gtm@rutgers.edu](mailto:gtm@rutgers.edu)

K. P. Brock

cBio Center, Dana-Farber Cancer Institute, Boston, MA, USA

C. Sander

Department of Cell Biology, Harvard Medical School, Boston, MA, USA

cBio Center, Dana-Farber Cancer Institute, Boston, MA, USA

D. S. Marks

Department of Systems Biology, Harvard Medical School, Boston, MA, USA

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,

Advances in Experimental Medicine and Biology 1105,

[https://doi.org/10.1007/978-981-13-2200-6\\_10](https://doi.org/10.1007/978-981-13-2200-6_10)

**Keywords** Hybrid methods · Protein NMR spectroscopy · Protein families · Multiple sequence alignment · Maximum entropy · Evolutionary couplings · Automated NMR data analysis · AutoStructure/ASDP

## 10.1 Introduction

Solution-state NMR can generally provide accurate three-dimensional (3D) structures of small (MW  $< \sim 15$  kDa) proteins (Mao et al. 2011, 2014). However, for larger proteins the efficient transverse spin relaxation of the  $^1\text{H}$ - $^1\text{H}$  network results in broad NMR line widths, preventing collection of sufficient data to allow structural analysis. Perdeuteration and selective reprotonation (i.e. replacement of most  $^1\text{H}$  atoms with  $^2\text{H}$ ) decreases transverse relaxation rates of the remaining  $^1\text{H}$ ,  $^{15}\text{N}$ , and  $^{13}\text{C}$  nuclei, increasing the sensitivity and feasibility of NMR for larger proteins (Gardner et al. 1997). However, perdeuteration also reduces the number of  $^1\text{H}$ 's providing  $^1\text{H}$ - $^1\text{H}$  NOEs, and generally excludes most sidechain protons, providing much fewer structural restraints. This incompleteness of NOE data can be compensated to some degree using conformational restraints based on chemical shift and orientation restraints from residual dipolar coupling (RDC) data. Although protein structure models based on such “sparse NMR data” can be improved using advanced knowledge-based molecular modeling methods (Raman et al. 2010; Lange et al. 2012; Sgourakis et al. 2014), the resulting structures are generally less accurate and precise than those obtained for smaller, fully-protonated proteins with complete sidechain resonance assignments.

It has long been a goal of bioinformatics research to use sequence co-variation to provide information about residue pair contacts, which could enable protein structure prediction and modeling (Gobel et al. 1994; Neher 1994; Taylor and Hatrick 1994; Shindyalov et al. 1994; Thomas et al. 1996). Historically, a key challenge was created by transitive correlations, or relay effects; i.e., to distinguish A-B covariation due to A->B interactions from A-C covariation due to relayed A->B->C interactions. Recently, methods have been developed using maximum entropy global statistical models and maximum likelihood parameter inference that distinguish direct evolutionary couplings from transitive correlations, allowing reliable analysis of evolutionary residue-residue couplings from multiple alignments of structurally related protein sequences (Lapedes et al. 2002; Morcos et al. 2011; Marks et al. 2011; Sulkowska et al. 2012; Kamisetty et al. 2013). Such evolutionary couplings (ECs), derived from evolutionary-correlated mutations, can provide accurate information about residue pair contacts in the 3D structures of proteins and protein complexes (Morcos et al. 2011; Marks et al. 2011, 2012; Sulkowska et al. 2012; Hopf et al. 2012, 2014; Kamisetty et al. 2013; Michel et al. 2014; Ovchinnikov et al. 2014, 2016, 2017; Anishchenko et al. 2017; Simkovic et al. 2017). Most often, the highest scoring evolutionary couplings are between residues that indeed contact one another in the 3D structure. These contacts can

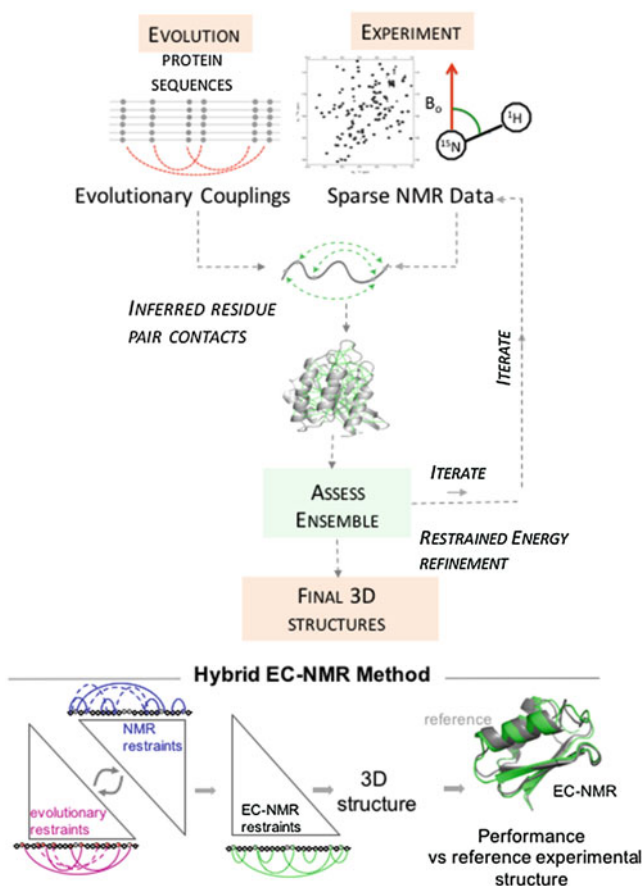
then be used, together with molecular dynamics, knowledge-based, and/or energy minimization methods to model the native structure of the protein, with often correct identification of the protein fold (Marks et al. 2011; Sulkowska et al. 2012; Hopf et al. 2012; Sheridan et al. 2015; Ovchinnikov et al. 2015, 2016, 2017). Importantly, high-confidence ECs may also reflect protein-protein interactions (Hopf et al. 2014; Cheng et al. 2014; Ovchinnikov et al. 2014; dos Santos et al. 2015; Toth-Petroczy et al. 2016), alternative conformational or allosteric states (Morcos et al. 2013; Toth-Petroczy et al. 2016), and/or more subtle features of the protein structure and dynamics.

While a breakthrough in the area of computational protein folding and protein structure prediction, the modeling of 3D structures from evolutionary couplings has a number of limitations. ECs provide information on residue-residue contacts present in many of the 3D structures of the proteins across the multiple sequence alignment (i.e., across the iso-structural protein subfamily or family), and may not accurately reflect the specific structural details of the particular protein under investigation. More specifically, there may be “structural drift” across the protein family, and sequence co-variation across distantly related members of the family may be inconsistent with the structure of the subject protein (Tang et al. 2015). In addition, even when there is extensive sequence information, residue-residue contacts indicated by high-ranked ECs may not be consistent with the native structure under investigation, but rather reflect important but confounding effects, such as conformational alternatives, allosteric networks, excited-state conformations, homo-oligomerization, and/or indirect residue interactions via substrates or binding partners. They may also result from simple false positives in the parameter inference computation, especially when insufficiently diverse sequences are available. As a result, EC-derived models of proteins may differ in detail from the predominant native structure.

Residue contact information derived from sparse NMR data or from evolutionary couplings can provide highly complementary information. This creates the opportunity to combine the two for more reliable structure determination than can be achieved using either data type alone (Tang et al. 2015). Sparse NMR contact information is incomplete and often ambiguous in its assignment to specific  $^1\text{H}$ - $^1\text{H}$  interactions. Nonetheless, all (or most) of the NOE, chemical shift, and RDC data should be consistent with the 3D structure model(s), across the ensemble at finite temperature. EC-based contacts can complement this spectroscopic information to provide more complete contact information, and more accurate models, but potentially include interactions that are not consistent with the predominant structure of the subject protein under the conditions that the NMR data is acquired. The requirement that the overall structure be consistent with all of the experimental NMR data, however, provides “hard” constraints on the interpretation of ECs, allowing identification and removal of proposed residue pair contacts that are inconsistent with the dominant structure present under the solution conditions under investigation (Tang et al. 2015).

## 10.2 The EC-NMR Algorithm

The general EC-NMR method, as described by Tang et al. (2015) is outlined in Fig. 10.1. The overall process can be divided into three sub processes. Step 1 provides a ranked list of direct evolutionary couplings (ECs) from multiple sequence alignments using either maximum entropy or pseudo likelihood models of the protein sequence, constrained by the statistics of the multiple sequence alignment, that have been developed to distinguish direct from transitive couplings (Morcos et al. 2011; Marks et al. 2011; Jones et al. 2012; Ekeberg et al. 2013; Kamisetty et al. 2013). In generating the multiple sequence alignment, it is important to carefully choose an appropriate range of evolutionary neighbors: not too many, so as to



**Fig. 10.1** 3D structure determination by the hybrid EC-NMR method. The hybrid EC-NMR strategy combines Evolutionary Coupling (EC) information from protein sequences with sparse experimental nuclear magnetic resonance (NMR) data

optimize specificity of structural constraints to the target of interest, and not too few, so as to retrieve as many sequences as possible at maximum sequence diversity and thus reduce sampling bias. In our published implementation of EC-NMR, the interaction parameters in the model, i.e., the evolutionary residue-residue couplings, were computed using pseudo-likelihood maximization in the computer program *plmc*, part of the *Evcouplings* software suite (Ekeberg 2013; <https://github.com/debbiemarkslab/plmc>).

In Step 2, sparse NMR data is collected using uniformly  $^{13}\text{C}$ ,  $^{15}\text{N}$ -enriched and/or  $^2\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ -enriched protein samples prepared with  $^1\text{H}$ - $^{13}\text{C}$  labeling of sidechain Leu, Val, and Ile( $\delta$ 1) methyl groups (Gardner et al. 1997; Rosen et al. 1996; Tugarinov et al. 2006), providing backbone  $^1\text{H}^{\text{N}}$ ,  $^{13}\text{C}$ , and  $^{15}\text{N}$ , as well as sidechain amide  $^1\text{H}^{\text{N}}$ - $^{15}\text{N}$  and some methyl  $^{13}\text{CH}_3$  resonance assignments. Backbone resonance assignments are determined, and backbone dihedral angle restraints are defined from  $^{13}\text{C}^{\alpha}$  and  $^{13}\text{C}^{\beta}$  chemical shift data using the program TALOS-N (Shen and Bax 2015). Unassigned NOESY peak lists are then generated from simultaneous 3D  $^{15}\text{N}$ ,  $^{13}\text{C}$ -NOESY spectra, and, in some cases,  $^{15}\text{N}$ - $^1\text{H}$  residual dipolar coupling (RDC) data are measured using one or more RDC alignment media. Such sparse NMR data can generally be obtained for perdeuterated proteins with molecular weights as large as 40–70 kDa (Hiller et al. 2008; Raman et al. 2010; Lange et al. 2012), and have been used to determine chain folds for proteins as large as 82 kDa (Tugarinov et al. 2005; Grishaev et al. 2008).

Step 3 identifies and iteratively refines distance restraints using both sources of information simultaneously, and determines a small set of accurate 3D structures. Chemical shift, NOESY peak list, EC, and RDC data are interpreted together to determine NOESY cross peak assignments, rule out ECs that are inconsistent with the NMR data, and to generate initial 3D models of the protein. This automated combined analysis of NMR and EC data is implemented in the NOESY assignment program *ASDP* (Huang et al. 2006). Intermediate 3D structures are generated from these combined NMR and evolutionary distance restraints using the program *CYANA* (Hermann et al. 2002). The resulting residue-pair contacts, derived by the combined analysis of EC and NMR data, are then deconvoluted into atom-specific distance restraints, which are used to refine the protein structure using restrained energy minimization. In the published implementation (Tang et al. 2015), the refinement step used a specific restrained energy minimization and knowledge-based modeling protocol with the program *Rosetta*, described by Mao et al. (2014), but alternative energy refinement protocols could also be used.

### 10.3 EC-NMR Results

Tang et al. (2015) tested the overall performance of the EC-NMR method using experimental chemical shift, NOESY peak list, and RDC data for 8 proteins ranging in size from 6 to 41 kDa. These data were obtained from the archives of the Northeast Structural Genomics Consortium ([www.nesg.org](http://www.nesg.org)) (Everett et al. 2016).



The resulting EC-NMR structures were compared with “reference structures”, which have been determined either by X-ray crystallography or by NMR using essentially complete backbone and sidechain resonance assignments. These EC-NMR structures were observed to have accurate backbone and all-heavy-atom positions; i.e.  $< 2 \text{ \AA}$  backbone atom positional root mean square deviations (RMSDs) and  $< 3 \text{ \AA}$  all-heavy atom RMSDs relative to the reference structure, in 6/8 proteins. The remaining two proteins studied, human p21 H-R as and maltose binding protein had no or limited RDC data, respectively, but were nevertheless reasonably accurate; both protein structures had backbone RMSDs  $< 2.8 \text{ \AA}$  and all-heavy-atom RMSDs  $< 3.6 \text{ \AA}$  relative to the corresponding X-ray crystal structures (Tang et al. 2015).

For this monograph, we re-determined five of the EC-NMR structures reported by Tang et al. (2015) using the same archived NMR data, but an updated database of protein sequences, downloaded in April 2017. These five proteins and the NMR data used for this study are summarized in Table 10.1. For the four smaller protein targets, with molecular weights of 6 to 15 kDa, the NMR data include only  $\text{H}^{\text{N}}\text{-H}^{\text{N}}$  NOE data, along with restraints on backbone dihedral angles computed from  $\text{C}^{\alpha}/\text{C}^{\beta}$  chemical shifts using Talos-N (Shen and Bax 2015). For two of these four proteins,  $^{15}\text{N}\text{-}^1\text{H}$  RDCs were measured using two different molecular alignment conditions, for a third  $^{15}\text{N}\text{-}^1\text{H}$  RDCs were measured using only one alignment condition, and for the fourth no RDC data are available. These four EC-NMR structures were compared with NMR structures determined with complete sidechain proton assignments and much more extensive NOESY data. The results of these EC-NMR calculations are shown in Fig. 10.2.

These four EC-NMR 3D structures were assessed based on (i) accuracy of atomic positions (Table 10.2) and (ii) accuracy of sidechain  $\chi_1$  rotamer states for well-defined (i.e. converged), buried (i.e., not on the protein surface) side chains (Table 10.3). In each case, the representative structure from the NMR ensemble (either the EC-NMR ensemble or the reference NMR structure ensemble) was selected as the medoid conformer of the ensembles, as described elsewhere (Montelione et al. 2013; Tejero et al. 2013). The backbone RMSD's between EC-NMR structures ranges from 1.5 to 1.8  $\text{\AA}$ , while the RMSD's for all C, N, O and S atoms (both backbone and sidechain) range from 2.4 to 2.9  $\text{\AA}$  (Table 10.2). The  $\chi_1$  values of well-defined buried sidechains (17–38 sidechains in the 4 structures), compared for all conformers in the EC-NMR ensemble with all conformers in the reference ensemble, also agree in 73–85% of pair-wise comparisons (Table 10.3). Similar results were observed for the corresponding earlier EC-NMR structures of these same proteins reported by Tang et al. (2015). In both studies, the EC-NMR structures are significantly more accurate than models generated using either the EC or sparse NMR data alone. Remarkably, these EC-NMR structures determined using only  $\text{H}^{\text{N}}\text{-H}^{\text{N}}$  NOE data together with ECs have accuracies that compare with high quality NMR structures determined with complete backbone and sidechain resonance assignments, suggesting that when good quality ECs are available for

**Table 10.1** Experimental data and benchmark reference structures

| Protein name and Uniprot ID   | N <sup>aa</sup> / MW <sup>a</sup> (kDa) | NOE Data <sup>b</sup>                                      | <sup>15</sup> N- <sup>1</sup> H RDC Data <sup>c</sup> | No. Sequences in MSA <sup>d</sup> | PDB ID of reference structure and method of structure determination |
|---|---|--|---|-----------------------------------|---|
| <b>Proteins &lt; ~15 kDa</b>  |   |  |   |                                   |   |
| <i>A. tumefaciens</i> protein of unknown function A9CJD6_AGRIT5   | 64 / 6.3                                | H <sup>N</sup> -H <sup>N</sup> only                        | None  | 28,265                            | 2K2P NMR  |
| <i>E. carotovora</i> cold-shock-like protein Q6D6V0_ERWCT   | 66 / 7.3                                | H <sup>N</sup> -H <sup>N</sup> only                        | 2 alignment tensors                                   | 7108                              | 2K5N NMR  |
| <i>A. thaliana</i> ubiquitin-like domain Q9ZV63_ARATH   | 84 / 9.7                                | H <sup>N</sup> -H <sup>N</sup> only                        | 2 alignment tensors                                   | 5396                              | 2KAN NMR  |
| <i>R. metallidurans</i> Rmet5065 QILD49_RALME   | 134 / 15.0                              | H <sup>N</sup> -H <sup>N</sup> only                        | 1 alignment tensor                                    | 31,674                            | 2LCG NMR  |
| <b>Proteins &gt; ~30 kDa</b>  |   |  |   |                                   |   |
| <i>E. coli</i> maltose binding protein MALE_ECOLI NTD (1-112; 259-329) CTD (113-258; 330-370) Full-length (1-370) | 370 / 40.7                              | H <sup>N</sup> -H <sup>N</sup> , Me-Me, H <sup>N</sup> -Me | 1 alignment tensor                                    | 43,759                            | IDMB Xray<br>IDMB Xray<br>IDMB Xray                                 |

<sup>a</sup>Number of residues (N) and molecular weight (MW) of the protein construct studied by NMR, excluding affinity purification tags

<sup>b</sup>H<sup>N</sup>-H<sup>N</sup> NOESY cross peak data include NOEs between backbone and sidechain amide H<sup>N</sup> resonances. For MALE\_ECOLI, additional H<sup>N</sup>-Me NOESY cross peak data obtained for uniformly <sup>15</sup>N, <sup>13</sup>C, <sup>2</sup>H-enriched samples with <sup>13</sup>CH<sub>3</sub> labeling of Ile(δ1), Leu, and Val methyls were also included

<sup>c</sup>All experimental <sup>15</sup>N-<sup>1</sup>H RDC data were measured in the laboratory of James Prestegard

<sup>d</sup>Number of non-redundant sequences in multiple sequence alignment used to generate ECs (N<sub>eff</sub>)

<sup>e</sup>Residue range for superimpositions and RMSD calculations: 2-63

<sup>f</sup>Residue range for superimpositions and RMSD calculations: 1-64

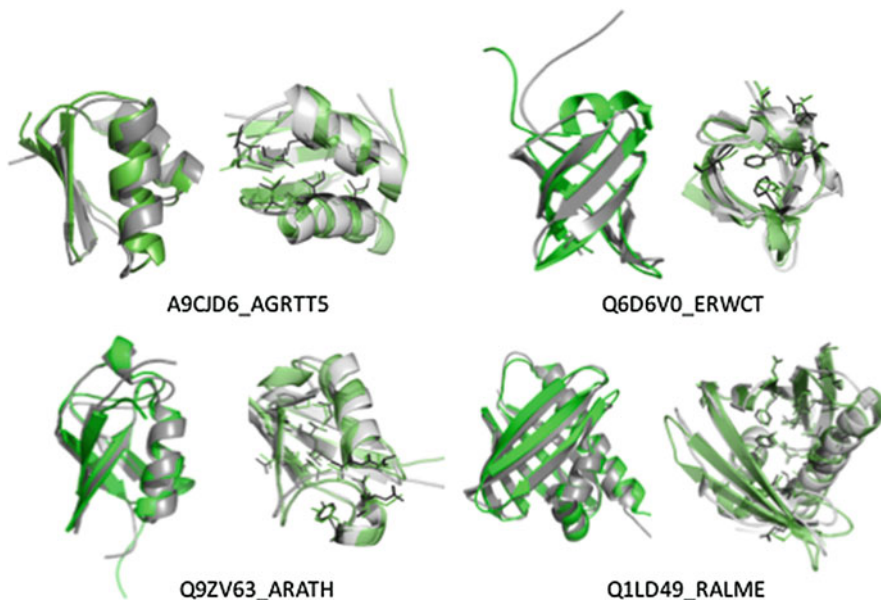
<sup>g</sup>Residue range for superimpositions and RMSD calculations: 7-78

<sup>h</sup>Residue ranges for superimpositions and RMSD calculations: 1-29, 36-58, 62-135

<sup>i</sup>Residue ranges for superimpositions and RMSD calculations: 2-12, 14-112, 259-329

<sup>j</sup>Residue ranges for superimpositions and RMSD calculations: 115-117, 125-142, 144-172, 175-218, 221-227, 247-258, 330-370. Interfacial residues 233-240 are exchange-broadened, precluding NMR assignments.

<sup>k</sup>Residue ranges for superimpositions and RMSD calculations: 2-12, 14-112, 259-329, 115-117, 125-142, 144-172, 175-218, 221-227, 247-258, 330-370. Interfacial residues 233-240 are exchange-broadened, precluding NMR assignments



**Fig. 10.2** EC-NMR structures determined using only  $H^N$ - $H^N$  NOESY data superimposed on reference conventional NMR structures. The representative structure from the ensemble of conformers generated by the EC-NMR method (green) is superimposed on a representative structure from reference NMR structure ensemble. For each protein, the left image is a superimposition of backbone atoms, and the right image a superimposition of backbone and well-defined core sidechain atoms

small (<15 kDa) proteins, it may only be necessary to complete the majority of backbone resonance assignments in order to determine a high-quality solution NMR structure.

As a fifth illustrative example, we also reanalyzed the EC-NMR structure of the 41 kDa *E. coli* maltose binding protein (MBP) bound to beta-cyclodextrin. The experimental NMR data for MBP include  $H^N$ - $H^N$  NOE data, as well as Ile( $\delta$ 1), Leu, and Val methyl proton assignments, providing also Me-Me and  $H^N$ -Me NOEs, along with restraints on backbone dihedral angles computed from  $C^\alpha/C^\beta$  chemical shifts using Talos-N (Shen and Bax 2015). These results (Fig. 10.3) demonstrate high-quality EC-NMR structures are produced, with backbone RMSD's to the corresponding X-ray crystal structure of 2.5 Å for backbone atoms, and 3.2 Å for all C, N, O and S atoms (both backbone and sidechain). MBP is a two-domain protein, and the relative orientation of domains depends on which sugars are bound; the “open form” being preferred when bound to beta-cyclodextrin (Evenas et al. 2001). Considered separately, the two individual domains of MBP in the EC-NMR structure of the two-domain protein are even more accurate when compared to the reference X-ray crystal structure (N-terminal domain/C-terminal domain backbone RMSD 1.8 Å / 1.7 Å, all-heavy-atom RMSD 2.7 Å / 2.6 Å; Table 10.2) than is apparent from rigid body superimposition for the entire protein.

**Table 10.2** Accuracy of EC-NMR structures

| Protein name and Uniprot ID   | Sequence database download (Month/Year) | No. sequences in MSA<br>$N_{\text{eff}}$ ( $N_{\text{eff}}/L$ ) | RMSD (Å) relative to reference: N, C $^{\alpha}$ , C $^{\beta}$ , O backbone / all C, N, O, S atoms |
|---|---|---|---|
| <i>A. tumefaciens</i> protein of unknown function<br>A9CJD6_AGRTT5<br>L = 63                            | Aug 2013                                | 10,964 (174)  | 1.5 ± 0.2 / 2.2 ± 0.2   |
| <i>E. carotovora</i> cold-shock-like protein<br>Q6D6V0_ERWCT<br>L = 63                                  | Apr 2017<br>Aug 2013                    | 28,265 (449)<br>4410 (70)                                       | 1.8 ± 0.2 / 2.4 ± 0.1<br>1.9 ± 0.3 / 2.9 ± 0.3  |
| <i>A. thaliana</i> ubiquitin-like domain Q9ZV63_ARATH<br>L = 73   | Apr 2017<br>Aug 2013                    | 7107 (113)<br>4964 (68)   | 1.7 ± 0.6 / 2.6 ± 0.4<br>1.4 ± 0.1 / 2.0 ± 0.1  |
| <i>R. metalldurans</i> Rmet5065 QILD49_RALME<br>L = 131   | Apr 2017<br>Aug 2013                    | 5396 (74)<br>2620 (20)  | 1.5 ± 0.1 / 2.4 ± 0.3<br>1.9 ± 0.3 / 3.0 ± 0.2  |
| <i>E. coli</i> maltose binding protein MALE_ECOLI<br>Full-length <sup>a</sup> (396 residues)<br>L = 388 | Apr 2017<br>Aug 2013                    | 31,674 (241)<br>12,416 (32)                                     | 1.7 ± 0.3 / 2.9 ± 0.2<br>2.9 ± 0.4 / 3.5 ± 0.4  |
| <i>E. coli</i> maltose binding protein MALE_ECOLI<br>N-terminal domain <sup>b</sup>                     | Apr 2017<br>Aug 2013                    | 43,759 (112)  | 2.5 ± 0.3 / 3.2 ± 0.3<br>1.6 ± 0.1 / 2.5 ± 0.2  |
| <i>E. coli</i> maltose binding protein MALE_ECOLI<br>C-terminal domain <sup>c</sup>                     | Apr 2017<br>Aug 2013                    |   | 1.8 ± 0.2 / 2.7 ± 0.2<br>1.9 ± 0.3 / 2.7 ± 0.2  |
|   | Apr 2017                                |   | 1.7 ± 0.2 / 2.6 ± 0.2   |

<sup>a</sup>Residues superimposed: 1–370<sup>b</sup>Residues superimposed: 1–112; 259–329<sup>c</sup>Residues superimposed: 113–258; 330–370

**Table 10.3** Assessment of the accuracy of well-defined, buried side chain  $\chi_1$  dihedral angles

| Protein NMR data set | Reference NMR structure | Number of buried, well-defined sidechains <sup>a</sup> | $\chi_1$ rotamer agreement (%) |
|----------------------|-------------------------|--|--------------------------------|
| A9CJD6_AGRT5         | 2K2P                    | 20   | 84                             |
| Q6D6V0_ERWCT         | 2K5N                    | 17   | 75                             |
| Q9ZV63_ARATH         | 2KAN                    | 21   | 73                             |
| Q1LD49_RALME         | 2LCG                    | 38   | 85                             |

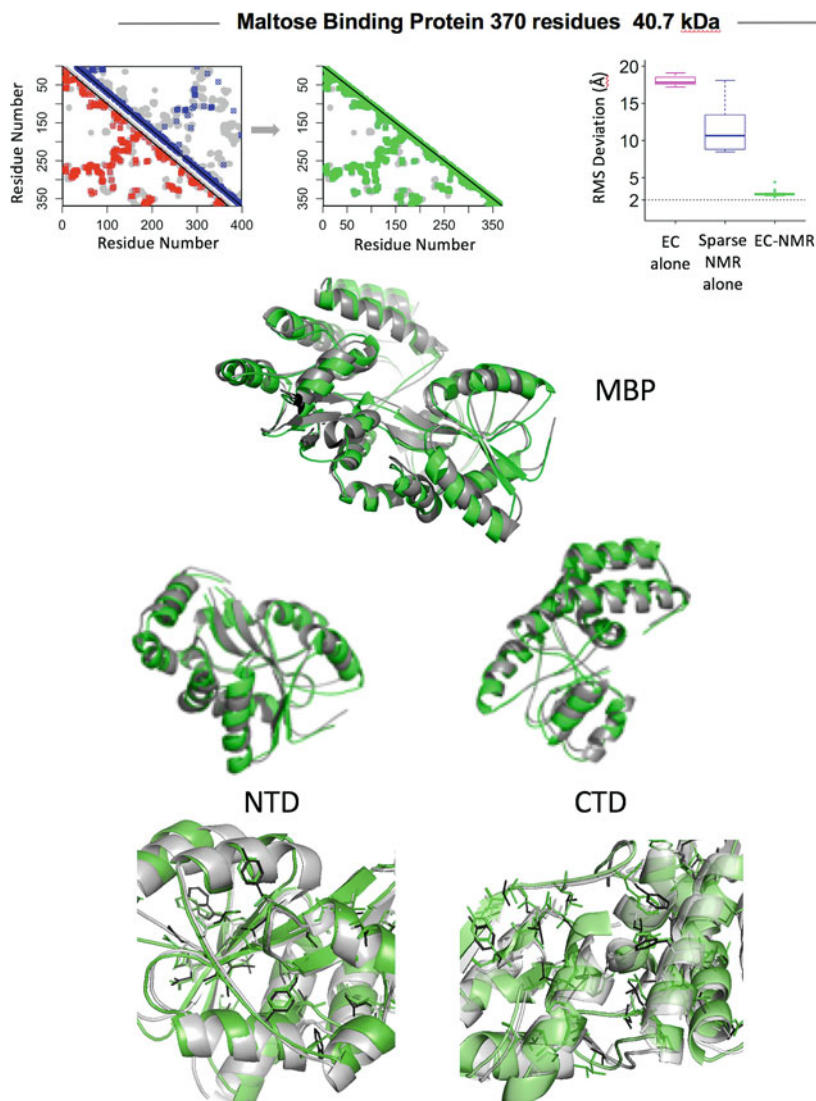
<sup>a</sup>Side chains that are buried (average SASA < 40 Å<sup>2</sup> in the NMR structures) and well-defined ( $\chi_1$  angle S.D. < 30 degrees in the NMR ensemble)

| Maltose binding protein |  |                                |   |                                |   |
|-------------------------|--|--------------------------------|---|--------------------------------|---|
| NMR Structure           | Number of buried, well-defined, side chains <sup>a</sup> | $\chi_1$ rotamer agreement (%) | Number of common buried, well-defined sidechains <sup>a</sup> | $\chi_1$ rotamer agreement (%) | RMSD to X-ray crystal structure <sup>b</sup><br>Full-length / NTD / CTD (Å) |
| 2D21                    | 105  | 76                             | 15  | 57                             | 5.4 / 1.6 / 1.5   |
| 1EZP                    | 33   | 26                             | 15  | 23                             | 3.3 / 2.8 / 2.6   |
| 2MV0                    | 80   | 75                             | 15  | 60                             | 4.7 / 2.0 / 3.7   |
| EC-NMR                  | 102  | 73                             | 15  | 57                             | 2.5 / 1.8 / 1.7   |

<sup>a</sup>Side chains that are buried (SASA < 40 Å<sup>2</sup> in the X-ray structure) and well-defined ( $\chi_1$  angle S.D. < 30 degrees in the NMR ensemble)

<sup>b</sup>The reference X-ray crystal structure is PDB ID 1DMB

We also compared the accuracy of the EC-NMR structure of MBP relative to previously published NMR structures determined with more extensive sidechain assignments (Table 10.3). The core sidechains of the EC-NMR structure are significantly more accurate than PDB ID 1EZP, determined using similar sparse NMR data together with 5 kinds of RDC data (Mueller et al. 2000). The core sidechain accuracy of the EC-NMR structure is similar to that of the solution NMR structure PDB ID 2D21, which was determined using extensive side chain resonance assignments provided by the sophisticated and expensive stereo-arrayed isotope labeling (SAIL) method (Kainosho et al. 2006). Based on RMSD relative to the X-ray crystal structure of beta-cyclodextrin-bound MBP, the overall structure of the EC-NMR models are more accurate than any previously published NMR structures. Similar results for MBP were also reported by Tang et al. (2015). Hence, we conclude that the EC-NMR method of Tang et al. (2015) can deliver structures with accurate backbone and core side chain atomic positions for larger (~40 kDa, or larger) proteins, with accuracy comparable or better than models obtained with sophisticated side chain labeling methods.



**Fig. 10.3** EC-NMR structure of *E. coli* maltose binding protein superimposed on the reference X-ray crystal structure. The top horizontal panels illustrate EC-NMR analysis process using sparse NMR data. Red contacts – initial EC residue-pair contacts. Blue contacts – contacts indicated by unambiguous NOESY peak assignments obtained by the *ASDP* program (Huang et al. 2006). Green contacts – final residue pair contacts resulting from simultaneous analysis of EC and NMR data. Grey contacts – contacts in the reference X-ray crystal structure. Box plots – RMSD to reference structures for backbone atoms of structures generated with EC data alone (red), sparse NMR data alone (blue), and the hybrid EC-NMR method (green). Superimposed backbone and core sidechain structures are for full length MBP, and for the individual N-terminal domain (NTD) and C-terminal domain (CTD) in the full-length EC-NMR structure. Green ribbon structures – final EC-NMR structure of MBP. Grey ribbons – reference X-ray crystal structure.

## 10.4 Sensitivity to Numbers of Sequence Homologs in Multiple Sequence Alignment

A prerequisite for the EC-NMR approach is extensive, diverse sequence data, required to obtain accurate co-evolutionary couplings between the residues (Marks et al. 2011; Hopf et al. 2012; Kamisetty et al. 2013). Recent experience suggests that more than  $2^*L$  non-redundant sequences ( $N_{\text{eff}}$ ) are generally required for confident predictions of overall protein fold from EC's alone, where  $L$  is the length of the target sequence (Marks et al. 2012; Michel et al. 2014; Ovchinnikov et al. 2014; Hopf et al. 2014; Kamisetty et al. 2013; Ovchinnikov et al. 2017). For a target protein that is 200 residues long, this typically requires on the order of 5000 sequences, before removal of redundancy, in an initial multiple sequence alignment of a family of structurally homologous proteins as inferred using standard sequence similarity methods with, if in doubt, a fairly conservative cutoff in sequence similarity, equivalent to typically not less than about 20–30% identical residues fairly evenly distributed over the entire length of the protein (Sander and Schneider 1991).

For EC-NMR, our goal is to obtain models with accuracies comparable to high-quality NMR structures; i.e. backbone positional root mean square deviations (RMSD's) relative to reference structures  $< 2.5 \text{ \AA}$  and accurate core sidechain packing. Tang et al. (2015) analyzed a series of multiple sequence alignments, testing the number of sequences from  $N_{\text{eff}}/L \sim 150$  down to  $N_{\text{eff}}/L < 0.1$ . In that analysis, using this implementation of the EC-NMR method and good quality NMR data for a perdeuterated, Ile( $\delta 1$ ), Leu and Val  $^{13}\text{CH}_3$  methyl labeled protein, the cutoff point for accurate modeling ( $< 2.5 \text{ \AA}$  backbone RMSD) was estimated to be  $N_{\text{eff}}/L \sim 5$ , with little improvement in structural accuracy for higher values of  $N_{\text{eff}}/L$  (see Fig. 4 of Tang et al. 2015).

For the five EC-NMR structures described above, the number of non-redundant sequences  $N_{\text{eff}}$  ranged from  $\sim 7100$  to  $\sim 44,000$  sequences (Table 10.2), with  $N_{\text{eff}}/L$  ranging from 113 to 241 sequences / residue. In order to assess the impact of the growth of the sequence databases over the last few years, we also compared these five EC-NMR structures, determined with protein sequence data available in April 2017, with the corresponding structures described by Tang et al. (2015), using protein sequence data downloaded in August 2013 (Table 10.2). Between these dates, the number of non-redundant sequences available for each of these five proteins increased significantly; by about 10% (for *A. thaliana* Ubiquitin-like domain Q9ZV63\_ARATH) to 12-fold (for *R. metallidurans* Rmet5065 Q1LD49\_RALME). This observation is consistent with our estimate that the size of the relevant sequence databases is doubling every 2–3 years (Tang et al. 2015), and that many proteins which cannot yet be reliably studied using the EC-NMR method will become amenable as the sequence data base grows. However, as these targets already had high  $N_{\text{eff}}/L$  using the 2013 sequence databases (ranging from 20 to 174 non-redundant sequences per residue), despite this increase in sequence data, with the available protein NMR data there was little or no improvements in structural

accuracy (Table 10.2). This is consistent with the conclusions of Tang et al. (2015), that good EC-NMR models can be produced with  $N_{\text{eff}}/L$  as low as 5 sequences / residue, with little improvement for higher values of  $N_{\text{eff}}/L$ . However, this cutoff depends also on the quality of sparse NMR data that is available.

## 10.5 Conclusions and Future Prospects

Evolutionary information and sparse NMR data, used together with knowledge-based modeling, are highly complementary for protein structure determination. The EC-NMR approach improves the accuracy of models generated by EC data alone, by requiring that EC-based contacts are consistent with experimental NMR data collected for one member of the protein family under specific conditions. This requirement eliminates important, but confounding, EC-derived contact restraints that may arise from structural drift across the protein family, and allosteric networks and/or excited states which may also be detected as evolutionary co-variation. More specifically, the experimentally reliable, but ambiguous, contact information of sparse NOESY peak list data, together with orientation restraints from RDC data and backbone dihedral restraints from chemical shift data, can rule out ECs that are not relevant to the structure of the specific target protein. Simultaneously, ECs complement the sparse NOESY and RDC data that can be obtained on largely perdeuterated protein samples, a requirement for studies of larger proteins and membrane proteins reconstituted in micelles or nano disks. In this way, complementarity EC and NMR data provide much more complete and accurate residue contact information than can be obtained from either method alone.

The EC-NMR method outlined in this monograph is **largely automated**, and provides high-quality 3D structures with accurate backbone and core sidechain conformations (Tang et al. 2015). **For small proteins** and domains up to 150 residues ( $< \sim 15$  kDa) with extensive sequence information, EC-NMR is a new, powerful, and efficient approach for protein structure determination using only backbone NMR data. **For larger proteins**, up to 400–500 residues (40–60 kDa, or larger), for which extensive side chain resonance assignment is challenging if not prohibitive, ECs can be combined with sparse NMR data obtained on perdeuterated protein samples to provide structures that are more accurate and complete than those obtained using such NMR data alone. In the method outlined here, ECs are combined with NMR data to determine both small and larger soluble protein structures, but the same approach should be applicable to membrane proteins (Hopf et al. 2012; Ovchinnikov et al. 2017), for solid-state NMR data, and for RNA structure determination (Weinreb et al. 2016). This advance significantly expands the range of biomolecules for which accurate structures can be determined using either evolutionary coupling analysis or NMR spectroscopy data alone.

The EC-NMR method requires large multiple sequence alignments, which are only currently available for a fraction of known proteins. However, as the sequence databases continue to grow, more proteins will be amenable to this



approach. Fortunately, combining ECs together with sparse NMR data reduces the requirements for the amount and diversity of sequence information.

In this work, we used a simple restrained energy minimization protocol of *Rosetta* in the final refinement step (Mao et al. 2014). This protocol improves both backbone and sidechain structure accuracy. It is advantageous because it is relatively fast, and can be implemented with limited computer resources. However, the resolution adapted recombination protocol (RASREC) developed by Lange and Baker (Raman et al. 2010; Lange and Baker 2012) has significant advantages for generating accurate structures of proteins from sparse NMR data (Raman et al. 2010; Lange et al. 2012). The RASREC protocol has also been used successfully for modeling protein structures for EC data (Braun et al. 2015). While it is much more computationally demanding, currently limiting its broad application, the RASREC protocol has the potential to provide more accurate EC-NMR structures with less complete and/or more noisy EC and sparse NMR data.

The EC-NMR method also allows identification of ECs which are not consistent with the NMR data collected for the target protein under specific conditions. While these are “false positives” relative to the modeling of this particular state of the protein, ECs with strong signals and high reliability that are not consistent with this particular state of the protein structure can provide information on alternative conformations accessible to the protein, excited states, and potentially provide information on allosteric networks. Further investigations of the combined use of ECs and NMR data to characterize the multiple conformational states of proteins and their energy landscapes is an exciting emerging area which can be explored using these powerful hybrid methods.

**Acknowledgements** This work was supported by National Institutes of Health grants 1R01-GM120574 (to G.T.M.) and 1R01-GM106303 (C.S. & D.M.). We thank all of the members of the Northeast Structural Genomics Consortium who generated and archived NMR data used in this work, particularly scientists in the laboratories of C. Arrowsmith, M. Kennedy, G.T. Montelione, T. Szyperski, and J. Prestegard.

## References

- Anishchenko I, Ovchinnikov S, Kamisetty H, Baker D (2017) Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci U S A* 114(34):9122–9127. <https://doi.org/10.1073/pnas.1702664114>
- Braun T, Koehler Leman J, Lange OF (2015) Combining evolutionary information and an iterative sampling strategy for accurate protein structure prediction. *PLoS Comput Biol* 11(12):e1004661. <https://doi.org/10.1371/journal.pcbi.1004661>
- Cheng RR, Morcos F, Levine H, Onuchic JN (2014) Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc Natl Acad Sci U S A* 111(5):E563–E571. <https://doi.org/10.1073/pnas.1323734111>
- dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN (2015) Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci Rep* 5:13652. <https://doi.org/10.1038/srep13652>

- Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlinear Soft Matter Phys* 359 87(1):012707
- Evenas J, Tugarinov V, Skrynnikov NR, Goto NK, Muhandiram R, Kay LE (2001) Ligand-induced structural changes to maltodextrin-binding protein as studied by solution NMR spectroscopy. *J Mol Biol* 309(4):961–974. <https://doi.org/10.1006/jmbi.2001.4695>
- Everett JK, Tejero R, Murthy SB, Acton TB, Aramini JM, Baran MC, Benach J, Cort JR, Eletsky A, Forouhar F, Guan R, Kuzin AP, Lee HW, Liu G, Mani R, Mao B, Mills JL, Montelione AF, Pederson K, Powers R, Ramelot T, Rossi P, Seetharaman J, Snyder D, Swapna GV, Vorobiev SM, Wu Y, Xiao R, Yang Y, Arrowsmith CH, Hunt JF, Kennedy MA, Prestegard JH, Szyperski T, Tong L, Montelione GT (2016) A community resource of experimental data for NMR / X-ray crystal structure pairs. *Protein Sci* 25(1):30–45. <https://doi.org/10.1002/pro.2774>
- Gardner KH, Rosen MK, Kay LE (1997) Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry* 36(6):1389–1401
- Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18(4):309–317. <https://doi.org/10.1002/prot.340180402>
- Grishaev A, Tugarinov V, Kay LE, Trewheella J, Bax A (2008) Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *J Biomol NMR* 40(2):95–106. <https://doi.org/10.1007/s10858-007-9211-5>
- Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* 319(1):209–227
- Hiller S, Garces RG, Malia TJ, Orekhov VY, Colombini M, Wagner G (2008) Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. *Science* 321(5893):1206–1210. <https://doi.org/10.1126/science.1161302>
- Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149(7):1607–1621. <https://doi.org/10.1016/j.cell.2012.04.012>
- Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Sander C, Bonvin AM, Marks DS (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3:e03430. <https://doi.org/10.7554/eLife.03430>
- Huang YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* 62(3):587–603. <https://doi.org/10.1002/prot.20820>
- Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28(2):184–190. <https://doi.org/10.1093/bioinformatics/btr638>
- Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Mei Ono A, Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. *Nature* 440(7080):52–57. <https://doi.org/10.1038/nature04525>
- Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* 110(39):15674–15679. <https://doi.org/10.1073/pnas.1314045110>
- Lange OF, Baker D (2012) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins* 80(3):884–895
- Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc Natl Acad Sci U S A* 109(27):10873–10878. <https://doi.org/10.1073/pnas.1203013109>
- Lapedes A, Giraud B, Jarzynski C (2002) Using sequence alignments to predict protein structure and stability with high accuracy. National Laboratory Report LA-UR-02-4481. <http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-02-4481> and arXiv:1207.2484 [q-bio.QM] (2012 copy)

- Mao B, Guan R, Montelione GT (2011) Improved technologies now routinely provide protein NMR structures useful for molecular replacement. *Structure* 19(6):757–766. <https://doi.org/10.1016/j.str.2011.04.005>
- Mao B, Tejero R, Baker D, Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. *J Am Chem Soc* 136(5):1893–1906. <https://doi.org/10.1021/ja409845w>
- Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6(12):e28766. <https://doi.org/10.1371/journal.pone.0028766>
- Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. *Nat Biotechnol* 30(11):1072–1080. <https://doi.org/10.1038/nbt.2419>
- Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A (2014) PconsFold: improved contact predictions improve protein models. *Bioinformatics* 30(17):i482–i488. <https://doi.org/10.1093/bioinformatics/btu458>
- Montelione GT, Nilges M, Bax A, Güntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS, Berman HM, Kleywegt GJ, Markley JL (2013) Recommendations of the wwPDB NMR validation task force. *Structure* 21(9):1563–1570. <https://doi.org/10.1016/j.str.2013.07.021>
- Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 108(49):E1293–E1301. <https://doi.org/10.1073/pnas.1111471108>
- Morcos F, Jana B, Hwa T, Onuchic JN (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci U S A* 110(51):20533–20538. <https://doi.org/10.1073/pnas.1315625110>
- Mueller GA, Choy WY, Yang D, Forman-Kay JD, Venters RA, Kay LE (2000) Global folds of proteins with low densities of NOEs using residual dipolar couplings: application to the 370-residue maltodextrin-binding protein. *J Mol Biol* 300(1):197–212. <https://doi.org/10.1006/jmbi.2000.3842>
- Neher E (1994) How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci U S A* 91(1):98–102
- Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife* 3:e02030. <https://doi.org/10.7554/eLife.02030>
- Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. *elife* 4:e09248. <https://doi.org/10.7554/eLife.09248>
- Ovchinnikov S, Kim DE, Wang RY, Liu Y, DiMaio F, Baker D (2016) Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins* 84(Suppl 1):67–75. <https://doi.org/10.1002/prot.24974>
- Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D (2017) Protein structure determination using metagenome sequence data. *Science* 355(6322):294–298. <https://doi.org/10.1126/science.aah4043>
- Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010) NMR structure determination for larger proteins using backbone-only data. *Science* 327(5968):1014–1018. <https://doi.org/10.1126/science.1183649>
- Rosen MK, Gardner KH, Willis RC, Parris WE, Pawson T, Kay LE (1996) Selective methyl group protonation of perdeuterated proteins. *J Mol Biol* 263(5):627–636
- Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9(1):56–68. <https://doi.org/10.1002/prot.340090107>

- Sgourakis NG, Natarajan K, Ying J, Vogeli B, Boyd LF, Margulies DH, Bax A (2014) The structure of mouse cytomegalovirus m04 protein obtained from sparse NMR data reveals a conserved fold of the m02-m06 viral immune modulator family. *Structure* 22(9):1263–1273. <https://doi.org/10.1016/j.str.2014.05.018>
- Shen Y, Bax A (2015) Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol Biol* 1260:17–32. [https://doi.org/10.1007/978-1-4939-2239-0\\_2](https://doi.org/10.1007/978-1-4939-2239-0_2)
- Sheridan R, Fieldhouse RJ, Hayat S, Sun Y, Antipin Y, Yang L, Hopf T, Marks DS, Sander C (2015) EVfold.org: evolutionary couplings and protein 3D structure prediction. *bioRxiv* 021022. <https://doi.org/10.1101/021022>
- Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* 7(3):349–358
- Simkovic F, Ovchinnikov S, Baker D, Rigden DJ (2017) Applications of contact predictions to structural biology. *IUCr* 4(Pt 3):291–300. <https://doi.org/10.1107/S2052252517005115>
- Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. *Proc Natl Acad Sci U S A* 109(26):10340–10345. <https://doi.org/10.1073/pnas.1207864109>
- Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat Methods* 12(8):751–754. <https://doi.org/10.1038/nmeth.3455>
- Taylor WR, Hatrick K (1994) Compensating changes in protein multiple sequence alignments. *Protein Eng* 7(3):341–348
- Tejero R, Snyder D, Mao B, Aramini JM, Montelione GT (2013) PDBStat: a universal restraint converter and restraint analysis software package for protein NMR. *J Biomol NMR* 56(4):337–351. <https://doi.org/10.1007/s10858-013-9753-7>
- Thomas DJ, Casari G, Sander C (1996) The prediction of protein contacts from multiple sequence alignments. *Protein Eng* 9(11):941–948
- Toth-Petroczy A, Palmedo P, Ingraham J, Hopf TA, Berger B, Sander C, Marks DS (2016) Structured states of disordered proteins from genomic sequences. *Cell* 167(1):158–170 e112. <https://doi.org/10.1016/j.cell.2016.09.010>
- Tugarinov V, Choy WY, Orekhov VY, Kay LE (2005) Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc Natl Acad Sci U S A* 102(3):622–627. <https://doi.org/10.1073/pnas.0407792102>
- Tugarinov V, Kanelis V, Kay LE (2006) Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. *Nat Protoc* 1(2):749–754. <https://doi.org/10.1038/nprot.2006.101>
- Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS (2016) 3D RNA and functional interactions from evolutionary couplings. *Cell* 165(4):963–975. <https://doi.org/10.1016/j.cell.2016.03.030>

# Chapter 11

## Harnessing the Combined Power of SAXS and NMR



A. M. Gronenborn

**Abstract** Single types of methodologies are no longer sufficient to adequately describe complex biological structures. As a result, integrated approaches that combine complementary data are being developed. This chapter describes the integration of nuclear magnetic resonance and small-angle scattering approaches to characterize solution structures of multi-domain proteins.

**Keywords** Integrated structural biology · Multi-domain proteins · NMR · SAXS · Molecular dynamics simulations

A major challenge for structural biology is providing a mechanistic understanding of the plethora of functions and associated conformational changes performed by macromolecular and supramolecular complexes that underlie cell biology. Obtaining structures of such assemblies is a necessary prerequisite, and the rich data that they provide will open up new opportunities in the biomedical, biotechnological, and pharmacological arenas.

In order to investigate and adequately describe multifaceted biological systems, single types of methodologies are no longer sufficient: researchers are turning more and more to integrated approaches, using complementary structural data. The complexity of biological phenomena, linked to the inherent partiality of any representation, requires the pursuit of multiple methods and models. As is universally appreciated, individual types of structural data are limited in scope, accuracy and generality, and any inherent shortcomings can be overcome or minimized using complementary information in an integrative fashion.

In addition to the traditional structural biology techniques of X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy (EM), additional

---

A. M. Gronenborn (✉)

Department of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

e-mail: [amg100@pitt.edu](mailto:amg100@pitt.edu)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,

Advances in Experimental Medicine and Biology 1105,

[https://doi.org/10.1007/978-981-13-2200-6\\_11](https://doi.org/10.1007/978-981-13-2200-6_11)

methods are increasingly used, alone and in combination, with traditional methods to generate structural information. These include mass spectrometry of crosslinked complexes (Cohen and Chait 2001) and native complexes (Mehmood et al. 2015), synchrotron radiation circular dichroism spectroscopy (Cowieson et al. 2008), electron paramagnetic resonance spectroscopy (EPR) combined with site-directed spin labelling (Hubbell et al. 2000), Small-Angle Scattering (SAXS) (Lipfert and Doniach 2007), and computational docking with sparse distance restraints (Schneidman-Duhovny et al. 2012).

Although the integration of all structural methodologies with cell biology, biochemistry and computational approaches has made major strides over the last few years, the current chapter focusses specifically on the integration of NMR and SAXS for structural biology, emphasizing their remarkable complementarity.

NMR has unique capabilities for studying structure and dynamics of biomolecules at the atomic level. Structural characterization of a protein or any other biological macromolecule by NMR in solution invariably describes a distribution of interconverting conformers, in contrast to most structural descriptions from X-ray crystallography, cryo EM or solid-state magic-angle spinning NMR. Solution NMR ensembles encompass conformational families that range from a narrow distribution for well-folded, globular proteins or domains to a wide distribution for unfolded or partially folded polypeptide ensembles.

In contrast to the atomic-level information available by NMR, SAXS affords low resolution information but furnishes important data on the global size and shape of a particle in solution, ideally complementing the NMR-derived data. Or, in other words, SAXS provides an overall picture of the 3D space occupied by all coexisting conformers, while high resolution NMR describes the details of the conformational landscape at the atomic level. Several excellent reviews describing the general use of SAXS for biomolecules in solution have been published, covering a number of different aspects of the technique (Guinier and Fournet 1955; Doniach 2001; Koch et al. 2003; Putnam et al. 2007; Svergun and Koch 2003; Doniach and Lipfert 2012). Furthermore, a focused review on the use of SAXS to derive global shape information of folded RNA molecules is also available (Bhandari et al. 2016).

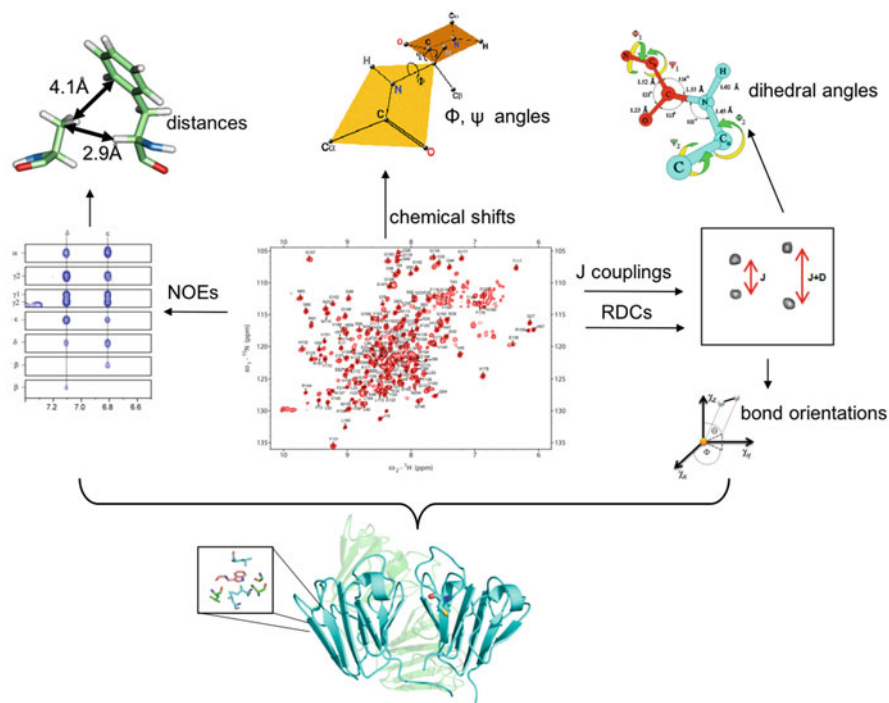
Like all structural techniques, NMR and SAXS each have advantages and disadvantages, as well as unique strengths and shortcomings. For example, SAXS is not limited by the molecular size of the particle under investigation (Graewert and Svergun 2013; Grant et al. 2011; Hura et al. 2009; Jeffries and Trehwella 2013; Martel et al. 2012) and can describe the contours of molecules with molecular masses of a few hundred kDa, a size too large for atomic level structure determination by solution NMR. Solution NMR, on the other hand, can provide detailed information about the atomic structure and dynamics of molecules, even for rare conformational sub-states (Sekhar and Kay 2013). However, both techniques are affected by potentially confounding factors to different degrees. While both methods ideally require monodispersity of the dissolved molecules, SAXS data quality is exquisitely sensitive to aggregation, and even a very small percentage (~1%) of aggregated species can compromise the data analysis. In contrast, such small amounts of aggregates would not be observed by solution NMR and the presence

of very large aggregates does not interfere with structural characterization of the smaller major component. For both SAXS and NMR, an additional complexity arises from conformational averaging on different timescales, reflecting the presence of local as well as global motions, which are important inherent properties of proteins (Henzler-Wildman and Kern 2007). Therefore, it is desirable to combine orthogonal techniques, which provide a more comprehensive description of the structure and dynamics than any individual method alone. In this regard, it is noteworthy that SAXS and NMR measurements can be performed on the same solution, ideally lending themselves to be used in an integrative fashion.

Given their complementarity, the integrated use of NMR and SAXS provides a powerful means to more completely describe the solution behavior of biological macromolecules, filling-in gaps or inherent imprecisions in the data extracted by either technique alone. Thus, when characterizing solution structures and architectures, it is desirable to obtain a SAXS shape envelope into which high resolution structures can be fitted, thus allowing the overall architecture of a multi-domain protein or multiprotein complex to be visualized.

NMR is an effective method for determining protein structure in solution at atomic resolution and has been routinely used for over 25 years (Fig. 11.1). However, for multi-domain proteins, even if a large number of distance-, angle- and chemical shift restraints are available, the relative orientations of individual domains are difficult to ascertain, given the predominantly local nature of the NMR-derived constraints. This limitation can be overcome, to some degree, by using extensive sets of residual dipolar couplings (RDCs). RDCs can be measured in solution NMR spectra, if molecules experience weak alignment in the magnetic field, either caused by the molecule's own magnetic susceptibility anisotropy or by employing very dilute liquid crystalline media (Tjandra and Bax 1997). These couplings contain information about the orientation of the associated inter-nuclear vector, relative to the molecular susceptibility anisotropy tensor and, therefore, provide angular restraints for structure calculations. Addition of RDC-derived restraints to conventional structure determination algorithms results in remarkable improvements, both locally as well as globally.

Algorithms for determining NMR structures aim to locate the global minimum of a target function containing terms for covalent geometry, non-bonded contacts, and the experimentally derived distance and angular restraints. The most important geometric information is provided by the nuclear Overhauser effect (NOE), which is translated into distances between proton pairs separated by  $<6 \text{ \AA}$ . Despite their short-range nature, these distances are highly conformationally restrictive, especially if they involve atoms that belong to units (amino acids or nucleotides) that are far apart in the linear sequence. Other experimental NMR restraints that provide short range structural information are three-bond coupling constants and secondary  $^1\text{H}$  and  $^{13}\text{C}$  chemical shifts. Three-bond coupling constants ( $^3J$ ) are related to torsion angles by the Karplus equation (Karplus 1963), with the  $^3J_{\text{HN}\alpha}$  coupling providing direct information about the phi backbone torsion angle. In a similar way, the empirical correlation between a protein's backbone conformation (phi/psi angles) and the difference in  $^{13}\text{C}\alpha$  and  $^{13}\text{C}\beta$  chemical shifts from random coil values

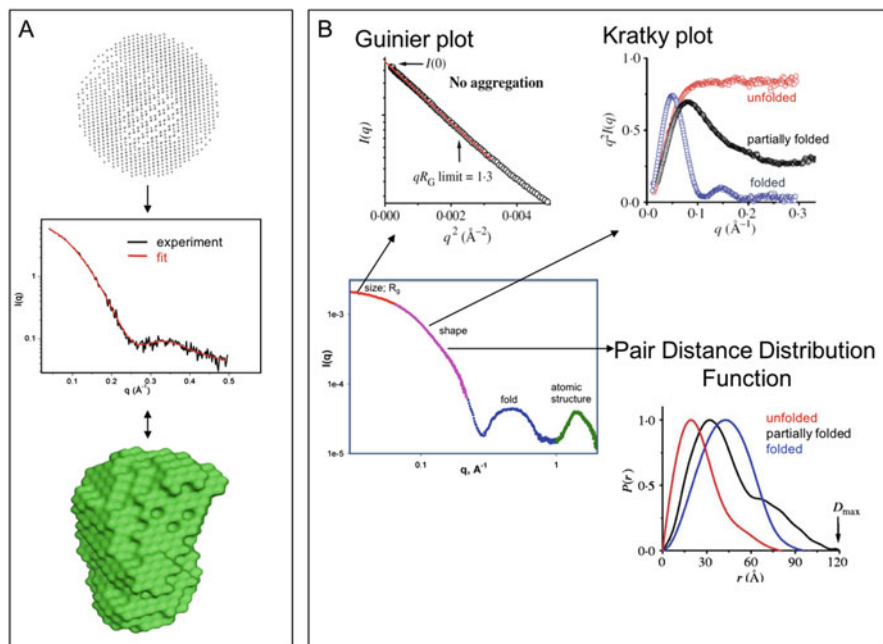


**Fig. 11.1** Schematic illustration of NMR-provided information. 2D spectrum (middle), NOESY data and distances (left), chemical shift-derived phi, psi angles (top), J coupling-derived dihedral angles and RDC-derived orientational restraints (right), are all combined to determine an atomic model (bottom)

are used in NMR structure determination.  $^1\text{H}$  chemical shifts are primarily used for refinement purposes, although recent advances in the *ab initio* calculation of proton shifts hold great promise for their routine use in NMR structure determination. In addition to these originally used parameters, paramagnetic relaxation enhancements (PREs) (Gillespie and Shortle 1997) and pseudocontact shifts (PCS) (Bertini and Luchinat 1999) augment the arsenal of geometric restraints that can be obtained by NMR.

SAXS data are measured as scattering signal intensity at a given value of  $q$ , where  $q = 4\pi\sin\theta/\lambda$ , with  $2\theta$  the scattering angle and  $\lambda$  the X-ray wavelength. Several program suites are available for processing SAXS data (e.g., PRIMUS, Scatter) (Rambo). The SAXS scattering profile (Fig. 11.2) at very small scattering angles (low  $q$  region) is frequently analyzed using the Guinier approximation, since the data for  $q$  close to zero vary linearly with  $q$  (Guinier and Fournet 1955). Thus, plotting the scattering intensity as  $\ln I(q)$  vs  $q^2$  results in a straight line with the slope equal to  $-R_g^2/3$  and the vertical intercept equal to the natural log of the zero-angle scattering intensity  $I(0)$ . In this manner, the radius of gyration,  $R_g$ , i.e. the





**Fig. 11.2** Schematic illustration of SAXS data and analysis. (a) Scattering pattern (top), an experimental scattering intensity profile with fit (middle), and a low-resolution dummy bead model (bottom). (b) A theoretical scattering intensity profile (middle) and the various basic methods for analysis of SAXS data

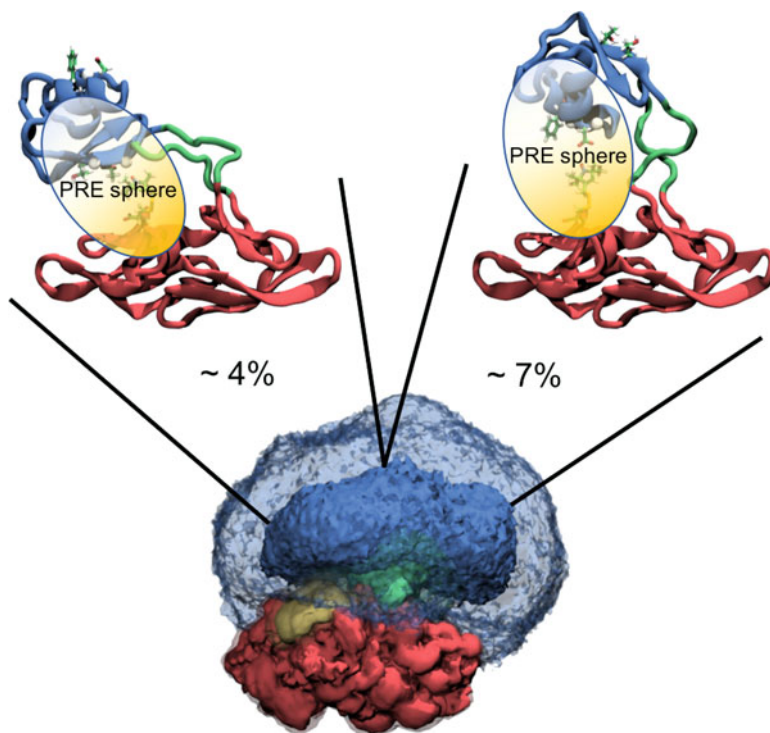
average root-mean-square distance from the center of density in the molecule can be extracted. Using the Guinier plots for the estimation of  $R_g$ , the maximum  $q$  that is acceptable to include in the fit is  $1.3/R_g$ . The extrapolated intensity at zero scattering angle,  $I(0)$ , is proportional to the electron density contrast between the scattering entity and the buffer and can be used to determine the molecular mass of the molecule (Fischer et al. 2010; Mylonas and Svergun 2007). Plotting  $I(0)$  vs concentration yields a straight line, unless large scale conformational averaging is present. Indeed, for highly flexible systems, the electron density contrast between the solute and the solvent is difficult to discern, rendering accurate determination of the volume and molecular weight values difficult.

Conformational flexibility or large amplitude motions in a molecule can be discerned from analysis of the scattering data using Kratky plots in which the scattering data is transformed as  $q^2 \cdot I(q)$  vs  $q$  (Fig. 11.2b) (Glatter and Kratky 1982). Kratky plots for well-ordered globular, disordered and highly flexible, as well as partially ordered entities exhibit characteristic features (Hammel 2012; Kikhney and Svergun 2015; Rambo and Tainer 2011) that can be used for an initial characterization of the system under investigation.

The most powerful means for analyzing SAXS data consists of Fourier transforming the scattering intensity  $I(q)$  into a pair-distance distribution function  $P(r)$  (Fig. 11.2b). This function represents a continuous  $r^2$ -weighted histogram of all electron-pair distances in the molecule (Glatter 1977). The  $P(r)$  function permits assessment of the overall quality of SAXS data analysis, since  $R_g$  and  $I(0)$  can be extracted directly from the  $P(r)$  function by integrating the function over all values of  $r$ . Calculating  $R_g$  and  $I(0)$  directly from  $P(r)$  uses all of the experimental data in real space, compared to solely using the linearly approximated points from the Guinier plot in the low- $q$  region.

SAXS data together with RDC data, initially, were used to successfully refine known solution NMR structures of single-chain proteins with simulated annealing (SA) protocols (Grishaev et al. 2005; Lee et al. 2007). The power of combining SAXS and NMR, however, is most evident for multi-domain proteins, in which individual domains are connected by flexible linkers (Hennig and Sattler 2014). For example, it is possible to determine global architectures of complexes, employing experimental SAXS and RDC data in conjunction with solution NMR-derived component structures, as shown by us and others (Wang et al. 2009; Ellis et al. 2009). A very instructive and comprehensive review on the integration of SAXS and NMR for the analysis of the structural dynamics of modular multi-domain proteins, using DNA replication proteins as examples, was published recently (Thompson et al. 2017). In addition, several methods for characterizing flexible systems in solution using SAXS data have been reported; these include ensemble optimization methods (Bernado et al. 2007; Schwieters and Clore 2007), a minimal ensemble search (Pelikan et al. 2009), a basis-set supported SAXS (Yang et al. 2010), an integrative modeling platform (Forster et al. 2008), a maximum-entropy refinement (Rozycki et al. 2011), and maximum occurrence method, MaxOcc (Bertini et al. 2012). These approaches entail the generation of a large number of structures to cover the accessible conformational space, from which a subset of conformers is selected that fit the experimental SAXS data. The methods differ in the way the starting conformational ensemble is generated and how the final ensemble is selected from the pool. Extending such ensemble refinement protocols to include NMR-derived distance and RDC restraints, in addition to SAXS data, in both, the pool generation and the optimal ensemble selection, have proven successful for two-domain proteins that possess significant inter-domain motions (Lemak et al. 2014).

An illustrative example of method integration, aimed at obtaining a more detailed picture of a macromolecule in solution is our recent study on the structure and dynamics of a domain-insertion protein (Fig. 11.3). In this case, we integrated crystallographic, NMR and SAXS data with microsecond-scale atomistic molecular dynamics to construct a structural model of the overall two-domain system. In particular, NMR relaxation and paramagnetic relaxation enhancement (PRE) experiments along with microsecond-scale MD simulations in explicit solvent were carried out. Using this comprehensive integrated approach, we established that the two domains in the protein have no fixed relative orientation, although certain orientations are preferred over others (Debiec et al. 2018). In summary,



Probability distributions of inter-domain orientations within the SAXS envelope

**Fig. 11.3** Integration of NMR- or X-ray-derived domain structure information, NMR relaxation data, SAXS data and long-time scale molecular dynamics simulations permits the characterization of a probabilistic ensemble of the overall solution structure. The LysM domain is shown in blue, the CVNH domain in red, the interdomain linkers in green, and the paramagnetic MTSL tag in yellow. Structures were best fit to the CVNH domain coordinates. Solid contours represent 1 Å<sup>3</sup> bins in the simulation that are occupied by a heavy atom in at least 1% of the ensemble, and transparent contours represent bins occupied in at least 0.1% of the ensemble

the integrated use of NMR and SAXS provides a powerful means to describe the solution behavior of biological macromolecules, as the combined data collected with each method permits one to derive a more complete picture of a multi-domain protein or multiprotein complex than can be provided by either technique alone. Thus, when characterizing solution structures of biological systems, one should consider obtaining a SAXS shape envelope into which high-resolution NMR structures can be fitted.

**Acknowledgment** I thank all former and present members of Gronenborn laboratory for their contributions to our studies referenced here and Teresa Brosenitsch for excellent editorial help. This work was supported by a National Institutes of Health Grant RO1GM080642 (A.M.G.).

## References

- Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* 129(17):5656–5664. <https://doi.org/10.1021/ja069124n>
- Bertini I, Luchinat C (1999) New applications of paramagnetic NMR in chemical biology. *Curr Opin Chem Biol* 3(2):145–151. [https://doi.org/10.1016/S1367-5931\(99\)80026-X](https://doi.org/10.1016/S1367-5931(99)80026-X)
- Bertini I, Ferella L, Luchinat C, Parigi G, Petoukhov MV, Ravera E, Rosato A, Svergun DI (2012) MaxOcc: a web portal for maximum occurrence analysis. *J Biomol NMR* 53(4):271–280. <https://doi.org/10.1007/s10858-012-9638-1>
- Bhandari YR, Jiang W, Stahlberg EA, Stagno JR, Wang YX (2016) Modeling RNA topological structures using small angle X-ray scattering. *Methods* 103:18–24. <https://doi.org/10.1016/j.ymeth.2016.04.015>
- Cohen SL, Chait BT (2001) Mass spectrometry as a tool for protein crystallography. *Annu Rev Biophys Biomol Struct* 30:67–85. <https://doi.org/10.1146/annurev.biophys.30.1.67>
- Cowieson NP, Miles AJ, Robin G, Forwood JK, Kobe B, Martin JL, Wallace BA (2008) Evaluating protein: protein complex formation using synchrotron radiation circular dichroism spectroscopy. *Proteins* 70(4):1142–1146. <https://doi.org/10.1002/prot.21631>
- Debiec KT, Whitley MJ, Koharudin LMI, Chong LT, Gronenborn AM (2018) Integrating NMR/SAXS experiments and atomistic simulations – structure and dynamics of a two-domain protein. *Biophys J* 114(4):839–855. <https://doi.org/10.1016/j.bpj.2018.01.001>
- Doniach S (2001) Changes in biomolecular conformation seen by small angle X-ray scattering. *Chem Rev* 101(6):1763–1778
- Doniach S, Lipfert J (2012) Small and wide angle X-ray scattering from biological macromolecules and their complexes in solution. In: *Comprehensive biophysics*. Elsevier, Amsterdam, pp 376–397. <https://doi.org/10.1016/B978-0-12-374920-8.00122-3>
- Ellis J, Gutierrez A, Barsukov IL, Huang WC, Grossmann JG, Roberts GC (2009) Domain motion in cytochrome P450 reductase: conformational equilibria revealed by NMR and small-angle x-ray scattering. *J Biol Chem* 284(52):36628–36637. <https://doi.org/10.1074/jbc.M109.054304>
- Fischer H, De Oliveira NM, Napolitano HB, Polikarpov I, Craievich A (2010) Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *J Appl Crystallogr* 43:101–109. <https://doi.org/10.1107/S0021889809043076>
- Forster F, Webb B, Krukenberg KA, Tsuruta H, Agard DA, Sali A (2008) Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. *J Mol Biol* 382(4):1089–1106. <https://doi.org/10.1016/j.jmb.2008.07.074>
- Gillespie JR, Shortle D (1997) Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. *J Mol Biol* 268(1):158–169. <https://doi.org/10.1006/jmbi.1997.0954>
- Glatter O (1977) A new method for the evaluation of small-angle scattering data. *J Appl Crystallogr* 10(5):415–421. <https://doi.org/10.1107/S0021889877013879>
- Glatter O, Kratky O (1982) *Small angle X-ray scattering in*. Academic, New York, p 515
- Graewert MA, Svergun DI (2013) Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS). *Curr Opin Struct Biol* 23(5):748–754. <https://doi.org/10.1016/j.sbi.2013.06.007>
- Grant TD, Luft JR, Wolfley JR, Tsuruta H, Martel A, Montelione GT, Snell EH (2011) Small angle X-ray scattering as a complementary tool for high-throughput structural studies. *Biopolymers* 95(8):517–530. <https://doi.org/10.1002/bip.21630>
- Grishaev A, Wu J, Trehwella J, Bax A (2005) Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. *J Am Chem Soc* 127(47):16621–16628. <https://doi.org/10.1021/ja054342m>
- Guinier A, Fournet G (1955) *Small-angle scattering of X-rays*. Wiley

- Hammel M (2012) Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS). *Eur Biophys J* 41(10):789–799. <https://doi.org/10.1007/s00249-012-0820-x>
- Hennig J, Sattler M (2014) The dynamic duo: combining NMR and small angle scattering in structural biology. *Protein Sci* 23(6):669–682. <https://doi.org/10.1002/pro.2467>
- Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. *Nature* 450(7172):964–972. <https://doi.org/10.1038/nature06522>
- Hubbell WL, Cafiso DS, Altenbach C (2000) Identifying conformational changes with site-directed spin labeling. *Nat Struct Biol* 7(9):735–739. <https://doi.org/10.1038/78956>
- Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL 2nd, Tsutakawa SE, Jenney FE Jr, Classen S, Frankel KA, Hopkins RC, Yang SJ, Scott JW, Dillard BD, Adams MW, Tainer JA (2009) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Methods* 6(8):606–612. <https://doi.org/10.1038/nmeth.1353>
- Jeffries CM, Trehwella J (2013) Small-angle scattering. In: Wall ME (ed) *Quantitative biology: from molecular to cellular systems*. CRC Press, Boca Raton, FL
- Karplus M (1963) Vicinal proton coupling in nuclear magnetic resonance. *J Am Chem Soc* 85:2870–2871
- Kikhney AG, Svergun DI (2015) A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* 589(19 Pt A):2570–2577. <https://doi.org/10.1016/j.febslet.2015.08.027>
- Koch MH, Vachette P, Svergun DI (2003) Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q Rev Biophys* 36(2):147–227
- Lee D, Walsh JD, Yu P, Markus MA, Choli-Papadopoulou T, Schwieters CD, Krueger S, Draper DE, Wang YX (2007) The structure of free L11 and functional dynamics of L11 in free, L11-rRNA(58 nt) binary and L11-rRNA(58 nt)-thiostrepton ternary complexes. *J Mol Biol* 367(4):1007–1022. <https://doi.org/10.1016/j.jmb.2007.01.013>
- Lemak A, Wu B, Yee A, Houliston S, Lee HW, Gutmanas A, Fang X, Garcia M, Semesi A, Wang YX, Prestegard JH, Arrowsmith CH (2014) Structural characterization of a flexible two-domain protein in solution using small angle X-ray scattering and NMR data. *Structure* 22(12):1862–1874. <https://doi.org/10.1016/j.str.2014.09.013>
- Lipfert J, Doniach S (2007) Small-angle X-ray scattering from RNA, proteins, and protein complexes. *Annu Rev Biophys Biomol Struct* 36:307–327. <https://doi.org/10.1146/annurev.biophys.36.040306.132655>
- Martel A, Liu P, Weiss TM, Niebuhr M, Tsuruta H (2012) An integrated high-throughput data acquisition system for biological solution X-ray scattering studies. *J Synchrotron Radiat* 19(Pt 3):431–434. <https://doi.org/10.1107/S0909049512008072>
- Mehmood S, Allison TM, Robinson CV (2015) Mass spectrometry of protein complexes: from origins to applications. *Annu Rev Phys Chem* 66:453–474. <https://doi.org/10.1146/annurev-physchem-040214-121732>
- Mylonas E, Svergun DI (2007) Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. *J Appl Crystallogr* 40 (s1):s245–s249. <https://doi.org/10.1107/S002188980700252X>
- Pelikan M, Hura GL, Hammel M (2009) Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen Physiol Biophys* 28(2):174–189
- Putnam CD, Hammel M, Hura GL, Tainer JA (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys* 40(3):191–285. <https://doi.org/10.1017/S0033583507004635>
- Rambo R BIOISIS ScÅtter. <http://www.bioisis.net/tutorial/9>. Accessed December 15, 2017
- Rambo RP, Tainer JA (2011) Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* 95(8):559–571. <https://doi.org/10.1002/bip.21638>
- Rozycki B, Kim YC, Hummer G (2011) SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* 19(1):109–116. <https://doi.org/10.1016/j.str.2010.10.006>

- Schneidman-Duhovny D, Rossi A, Avila-Sakar A, Kim SJ, Velazquez-Muriel J, Strop P, Liang H, Krukenberg KA, Liao M, Kim HM, Sobhanifar S, Dotsch V, Rajpal A, Pons J, Agard DA, Cheng Y, Sali A (2012) A method for integrative structure determination of protein-protein complexes. *Bioinformatics* 28(24):3282–3289. <https://doi.org/10.1093/bioinformatics/bts628>
- Schwieters CD, Clore GM (2007) A physical picture of atomic motions within the Dickerson DNA dodecamer in solution derived from joint ensemble refinement against NMR and large-angle X-ray scattering data. *Biochemistry* 46(5):1152–1166. <https://doi.org/10.1021/bi061943x>
- Sekhar A, Kay LE (2013) NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers. *Proc Natl Acad Sci U S A* 110(32):12867–12874. <https://doi.org/10.1073/pnas.1305688110>
- Svergun DI, Koch MHJ (2003) Small-angle scattering studies of biological macromolecules in solution. *Rep Prog Phys* 66(10):1735–1782
- Thompson MK, Ehlinger AC, Chazin WJ (2017) Analysis of functional dynamics of modular multidomain proteins by SAXS and NMR. *Methods Enzymol* 592:49–76. <https://doi.org/10.1016/bs.mie.2017.03.017>
- Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* 278(5340):1111–1114
- Wang J, Zuo X, Yu P, Byeon IJ, Jung J, Wang X, Dyba M, Seifert S, Schwieters CD, Qin J, Gronenborn AM, Wang YX (2009) Determination of multicomponent protein structures in solution using global orientation and shape restraints. *J Am Chem Soc* 131(30):10507–10515
- Yang S, Blachowicz L, Makowski L, Roux B (2010) Multidomain assembled states of Hck tyrosine kinase in solution. *Proc Natl Acad Sci U S A* 107(36):15757–15762. <https://doi.org/10.1073/pnas.1004569107>

# Chapter 12

## 2DHybrid Analysis



Atsushi Matsumoto and Kenji Iwasaki

**Abstract** We have developed an approach termed ‘2D hybrid analysis’ for building three-dimensional (3D) structures from electron microscopy (EM) images of biological molecules. The key advantage is that it is applicable to flexible molecules, which are difficult to analyze by the approach in which 3DEM maps are reconstructed. In the proposed approach, a large number of atomic models with different conformations are first built by computer simulation. Then, simulated EM images are produced from each atomic model. Finally, these images are compared with an experimental EM image to identify the best-fitting atomic model. Two kinds of models are used to simulate the EM images: the negative-stain model and the simple projection model. Although the former is more realistic, the latter permits faster computation. We applied this approach to the averaged EM images of integrin. Although many of these were reproduced well by the best-fitting atomic models, others did not closely resemble any of the simulated EM images. However, the latter group were well reproduced by averaging multiple simulated EM images originating from atomic models with rather different conformations or orientations. This indicated that our approach is capable of detecting mixtures of conformations in the averaged EM images, which should assist in their correct interpretation.

**Keywords** Simulated EM image · Negative stain · Averaging · Modeling · Protein structure

---

A. Matsumoto (✉)

Molecular Simulation and Modeling Group, National Institutes for Quantum and Radiological Science and Technology, Kizugawa, Kyoto, Japan  
e-mail: [matsumoto.atsushi@qst.go.jp](mailto:matsumoto.atsushi@qst.go.jp)

K. Iwasaki

Life Science Center for Survival Dynamics, Tsukuba Advanced Research Alliance (TARA), University of Tsukuba, Tsukuba, Ibaraki, Japan  
e-mail: [ikenji@tara.tsukuba.ac.jp](mailto:ikenji@tara.tsukuba.ac.jp)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*, Advances in Experimental Medicine and Biology 1105,  
[https://doi.org/10.1007/978-981-13-2200-6\\_12](https://doi.org/10.1007/978-981-13-2200-6_12)

181

## 12.1 Introduction

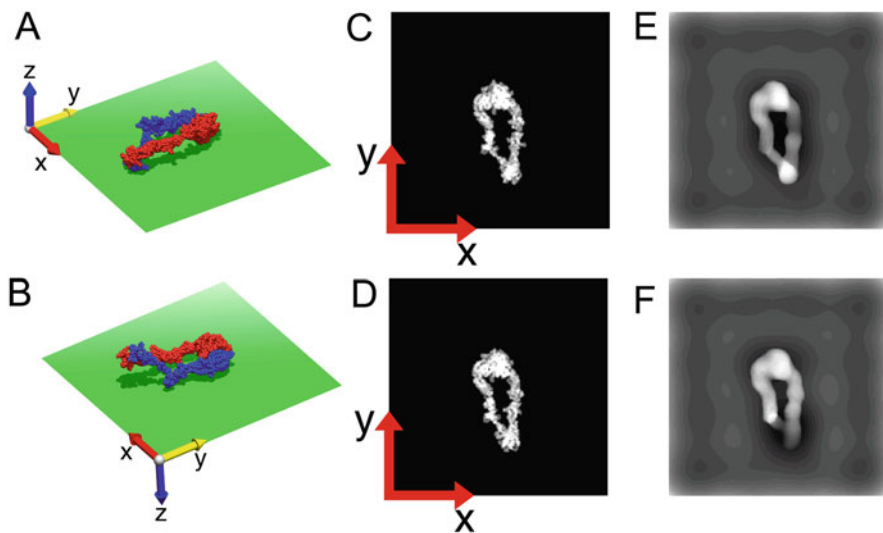
In this chapter, we describe a computational approach termed ‘2D hybrid analysis’, which we recently developed for building three-dimensional (3D) structural models of biological macromolecules by analyzing negative-stain EM images (Matsumoto et al. 2017). An application of this approach to the averaged EM images of integrin is also discussed.

A ‘3D hybrid approach’ involving cryo-electron microscopy and X-ray crystallography has been widely and successfully applied in revealing the complete structure of protein complexes that are difficult to crystallize (Schroder 2015), and in obtaining information about large-scale conformational changes in biological macromolecules (Villa and Lasker 2014). In this approach, single-particle analysis is used to reconstruct a three-dimensional Coulomb potential map (or 3DEM map). Despite the successful application of this hybrid approach, the reconstruction is not easy. In fact, it is often very difficult to reconstruct a 3DEM map of a flexible molecule. Additionally, multiple 3DEM maps are necessary for analyzing conformational changes in proteins, and consequently an enormous number of EM images and substantial computational resources for image analysis are necessary for reconstructing these multiple 3DEM maps. It would therefore be desirable to build 3D structures of macromolecules more easily and more swiftly, without having to reconstruct 3DEM maps. The 2D hybrid analysis approach was developed to satisfy these demands.

In 2D hybrid analysis, many atomic models with different conformations are first prepared. Then, simulated EM images are produced from each atomic model. Finally, these images are compared with experimental EM images to identify the best-fitting atomic model. At present, we use two different kinds of simulated EM image: the simple projection model and the negative-stain model. Previously, we used only the simple projection model, where each atom is projected as a point or a filled circle, and we analyzed the EM images of ‘giant’ cadherins to build 3D models successfully (Tsukasaka et al. 2014). However, when we analyzed the EM images of integrin in a similar way, we encountered problems when similar simulated EM images were obtained from atomic models that were rather different in terms of the conformations and orientations (Fig. 12.1). Cadherin typically exhibits a linear string-like topology, whereas integrin has a compact form. Possibly, then, this compactness required more-accurate simulated EM images. We therefore introduced the negative-stain model (Burgess et al. 1997). As shown in Fig. 12.1e, f, this model was clearly able to differentiate between the two atomic models.

The 2D hybrid analysis approach was developed to build an atomic model from each EM image, i.e., one atomic model from one image. However, during the application of this approach to averaged EM images, we often noticed that the EM image could not be reproduced well from a single atomic model. Instead, the EM image was reproduced well by combining multiple simulated EM images produced from atomic models with different conformations and orientations. This indicated that the conformations and orientations were intertwined in the averaging process





**Fig. 12.1** Example of a case in which atomic models with different conformations and orientations give similar simple projections. (a, b) Atomic models of integrins represented by sphere models contacting the supporting films, represented by the green rectangles. The arrows represent the axes of the coordinate system. These models are projected along the negative direction of the  $z$ -axis. (c, d) Simple projection models of (a) and (b), respectively. (e, f) Negative-stain models of (a) and (b), respectively. The stain thickness  $h$  was set to 30 Å in both cases. (Matsumoto et al. 2017)

(Marabini and Carazo 1994); that is, the molecules with different conformations and orientations generated similar raw images that were difficult to differentiate, and were therefore used for making an averaged EM image. Noise in the raw images would have exacerbated the difficulty in differentiating these images. The successful reproduction of such averaged EM images indicated that our approach is capable of detecting mixtures of conformations in the EM images, which should assist in the correct interpretation of EM images.

## 12.2 Methodological Overview

### 12.2.1 Overview of the Computation

In our computational approach, we first built many atomic models with different conformations by deforming the X-ray crystal structure or the modeled structure through a computational approach. For the integrin, we used the normal-mode analysis of the elastic network model (ENM) (Bahar et al. 1997; Tama and Brooks 2005; Tirion 1996). Then, each atomic model was projected in a variety of directions to produce the simulated EM images, which were compared with the experimental

EM images to select the best-fitting atomic model. Two kinds of models were used as the simulated EM images: the negative-stain model and the simple projection model. The former model is more realistic, but building it requires a longer computational time. Consequently, the latter model was used to narrow down the candidate atomic models in a shorter computational time.

## 12.2.2 Construction of the Elastic Network Model

The ENM is composed of points with masses that are connected by springs. Each amino acid residue is represented by a single point located at the position of the C $\alpha$  atom, and whose mass is the same as the total mass of the residue. The initial conformation of the ENM was built from the X-ray crystal structure [PDB ID: 3IJE for integrin (Xiong et al. 2001)]. We connected the representative points of two amino acid residues by a spring with the same spring constant when one of the following two conditions was satisfied (Matsumoto et al. 2008): (1) the minimum interatomic distance between the two amino acid residues is less than the threshold value  $d_c$ , which is set to 3.3 Å for integrin; and (2) the two amino acid residues are on the same chain, and the inter-residue distance is less than or equal to 3; that is, if the residue number of one of the amino acid residues is  $m$ , that of the other is  $m \pm 1$ ,  $m \pm 2$ , or  $m \pm 3$ .

## 12.2.3 Deformation of Atomic Models

We then built many different atomic models by deforming the X-ray crystal structure along the lowest-frequency normal modes. The atomic model  $\mathbf{r}^k$ , which is the  $3N$ -dimensional vector describing the positions of the  $N$  representative points, deformed along the  $k$ th lowest-frequency normal mode of the X-ray crystal structure  $\mathbf{r}^0$  is described as follows:

$$\mathbf{r}^k(a_k) = \mathbf{r}^0 + a_k \mathbf{u}_k,$$

where  $\mathbf{u}_k$  is the  $k$ th lowest-frequency normal-mode vector of the X-ray crystal structure and  $a_k$  is the magnitude of the deformation. However, in building models with large deformations, it is inappropriate to use this equation because linear movements of atoms often destroy the structure when  $a_k$  is large. Instead, we apply the normal-mode analysis and the small deformation in an iterative manner (Matsumoto and Ishida 2009; Matsumoto et al. 2008; Miyashita et al. 2003) to the X-ray crystal structure, as follows:

$$\mathbf{r}^k(n) = \mathbf{r}^0 + a_k^0 \mathbf{u}_k^0 + a_k^1 \mathbf{u}_k^1 + \cdots + a_k^{n-1} \mathbf{u}_k^{n-1},$$

where  $\mathbf{r}^k(n)$  is the atomic model deformed iteratively  $n$  times along the  $k$ th lowest-frequency normal mode and  $\mathbf{u}_k^n$  is the normal-mode vector for  $\mathbf{r}^k(n)$  ( $|\mathbf{u}_k^n| = 1$ ). In each iteration, the model is deformed so that the RMSD of  $\mathbf{r}^k(n)$  from  $\mathbf{r}^k(n-1)$  is 1 Å; i.e.,  $a_k^0 = a_k^1 = \dots = a_k^{n-1} = \sqrt{N}$ . In describing models deformed in the opposite direction, we use negative integers  $n$ . For example,  $\mathbf{r}^k(-1) (= \mathbf{r}^0 - a_k^0 \mathbf{u}_k^0)$  is the model deformed in the opposite direction with respect to  $\mathbf{r}^k(1) (= \mathbf{r}^0 + a_k^0 \mathbf{u}_k^0)$ .

By using this iterative approach, we constructed a library of deformed atomic models as follows. First, the X-ray crystal structure was deformed iteratively along the first-lowest-frequency normal mode, and a series of deformed atomic models  $\mathbf{r}^1(n_1)$  ( $n_1 = 0, \pm 1, \pm 2, \pm 3, \dots$ ) were built. Next, each atomic model  $\mathbf{r}^1(n_1)$  was deformed iteratively along the second-lowest-frequency normal modes, and series of atomic models  $\mathbf{r}^{12}(n_1, n_2)$  ( $n_2 = 0, \pm 1, \pm 2, \pm 3, \dots$ ) were built, where  $\mathbf{r}^{12}(n_1, 0) = \mathbf{r}^1(n_1)$ . By repeating this process for other normal modes, a library of deformed atomic models was built.

### 12.2.4 The Simple Projection Model

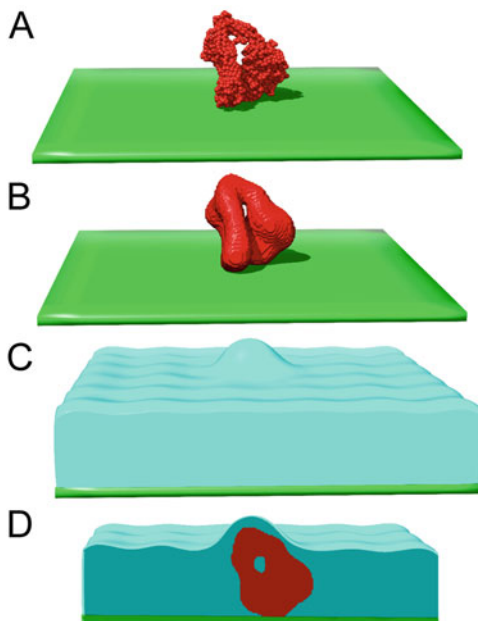
From the numerous deformed atomic models, we selected the model that best reproduced the EM image. To achieve this selection, we built simulated models of EM images from each atomic model. We built two kinds of model: a simple projection model and a negative-stain model. Although the latter was more realistic, building it required a much longer computational time. Therefore, the simple projection model was used to narrow down the number of candidates, and the negative-stain model was used to make the final selection.

We will now describe the simple projection model. We start from a deformed atomic model made of representative points. Each representative point is replaced by a sphere of uniform density and a radius of 3 Å to build the sphere model (Fig. 12.2a). The grid points within the spheres are projected onto the  $xy$  plane to produce a simple projection model. We define the simple projection model  $\rho_1(i, j)$  by the number of points projected into a pixel  $(i, j)$  ( $i = 1, 2, 3, \dots, i_{\max}$ ) ( $j = 1, 2, 3, \dots, j_{\max}$ ). Here, we assume that the pixel  $(i, j)$  corresponds to the square described by  $p(i-1) \leq x < pi$  and  $p(j-1) \leq y < pj$ , where  $p$  is the pixel size determined experimentally.

To compare the experimental EM image  $I(i, j)$  and the simple projection model  $\rho_1(i, j)$ , we first replace  $I(i, j)$  with  $I_1(i, j) (= I(i, j) - \langle I(i, j) \rangle)$ , where  $\langle \dots \rangle$  denotes the average, to remove the background intensity. If  $I(i, j)$  is less than  $\langle I(i, j) \rangle$ ,  $I_1(i, j)$  is set to zero. Then, to quantify the similarity between  $I_1(i, j)$  and  $\rho_1(i, j)$ , we define the score by using the normalized cross-correlation (NCC) as follows:

$$Sc_1 = \sum_{i,j} \rho_1(i, j) I_1(i, j) / \sqrt{\sum_{i,j} \rho_1(i, j)^2 \sum_{i,j} I_1(i, j)^2}.$$

**Fig. 12.2** Illustrations explaining how the simulated EM images are produced. (a) A sphere model on the supporting film, where each representative point is shown by a sphere. (b) An excluded-volume model. (c) The simulated negative stains cover the excluded-volume model in (b). (d) Cross-section of (c). The excluded-volume model in (b) is drawn within the volume of the simulated negative stains



Maximizing this score is equivalent to minimizing the difference between the two images,  $\sum (I_1(i,j) - c\rho_1(i,j))^2$ , where  $c$  is a constant.

To maximize the score, we apply rotational and translational manipulations to each atomic model. By the manipulations, each representative point  $\mathbf{r}_a$  ( $a = 1, 2, 3, \dots, N$ ) is moved to a new position  $\mathbf{r}'_a (= {}^t\mathbf{R}\mathbf{r}_a + \mathbf{s})$ , where  $\mathbf{R}$  is the rotation matrix and  $\mathbf{s}$  is the translational vector. We assume  $\mathbf{s} = {}^t(pk_x, pk_y, 0)$  ( $k_x, k_y = 0, \pm 1, \pm 2, \pm 3, \dots$ ) for faster computations.

To sample the entire range of orientations of the atomic model as evenly as possible, we prepared more than 230,000 rotation matrices in advance, as follows. The rotation matrix  $\mathbf{R}$  is described as  $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ , where  $\mathbf{e}_1, \mathbf{e}_2$ , and  $\mathbf{e}_3$  are unit column vectors that satisfy the equation  $\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3$ . We first selected 2562 different directions for  $\mathbf{e}_3$ . These directions were obtained as position vectors of the apexes of the icosahedron-based geodesic sphere (Sadourny et al. 1968), whose center is at the origin. The angle between neighboring vectors is about  $4^\circ$ . Then, vectors  $\mathbf{e}_1$  orthogonal to each  $\mathbf{e}_3$  were computed at  $4^\circ$  intervals. Finally,  $\mathbf{e}_2$  was obtained as  $\mathbf{e}_3 \times \mathbf{e}_1$ .

### 12.2.5 Contacts with Support Film

The EM images analyzed here were obtained by the negative-staining method, and the molecules were assumed to contact the supporting film in a stable manner.

We utilized this assumption in reducing the number of computations. To measure the stability of the contact between the molecules and the film, we define the contact area as follows. We assume that the supporting film is on the  $xy$  plane and that the top (the representative point with the maximum  $z$ -coordinate) or the bottom (the representative point with the minimum  $z$ -coordinate) of the atomic model is on the film. We regard representative points within  $10 \text{ \AA}$  of the  $xy$  plane as being in contact with the plane. We define the contact area  $S$  as the area of the minimum convex polygon that include all the contacting points projected onto the  $xy$  plane. The contact area  $S$  is dependent on the orientation, and the largest one is defined as  $S_{\max}$  for each atomic model. The ratio  $S/S_{\max}$  is then used as the measure of the stability of the contact.

### 12.2.6 The Negative-Stain Model

In some cases, several atomic models with quite different conformations and orientations give rise to similar simple projection models (Fig. 12.1). To differentiate between these atomic models, we used a more-realistic projection model, i.e., the negative-stain model. To produce the negative-stain model, we followed the approach proposed by Burgess et al. (1997). First, low-pass filtering (with a cut-off frequency  $\nu_1$ ) and thresholding are applied to the volume occupied by the sphere model (Fig. 12.2a) that is used to produce the simple projection model, in order to build an excluded-volume model (Fig. 12.2b). Then, the volume within  $h \text{ \AA}$  of the support film is added to this excluded volume. Note that the atomic model contacts the support film. Again, low-pass filtering (with a cut-off frequency  $\nu_2$ ) and thresholding are applied to this volume to obtain a new volume (Fig. 12.2c, d), from which the excluded volume of the atomic model is removed to acquire the volume of the simulated negative stain. The grid points within the volume are projected onto the  $xy$  plane to produce the negative-stain model. The number of points projected into a pixel  $(i, j)$  is counted as  $\rho_N(i, j)$ . We assume that the intensity of the incident electron beam decays exponentially with an increase in the thickness of the negative stain. Thus, the negative-stain model  $\rho_2(i, j)$  is described as  $\exp(-c_d \rho_N(i, j))$ , where  $c_d$  is a coefficient ( $>0$ ). Because  $c_d \rho_N \ll 1$  is expected,  $\rho_2(i, j)$  is approximately equal to  $1 - c_d \rho_N$ .

Note that the cut-off frequencies  $\nu_1$  and  $\nu_2$  are kept constant during the entire series of computations, but these might be dependent on the kind of negative stain that is used (uranyl acetate was used for integrin). Consequently, they must be optimized before performing the search. In the case of integrin, we optimized them so that the EM images of integrin in  $\text{Ca}^{2+}$  solution were reproduced well on average by the X-ray crystal structure. On the other hand, the thickness  $h$  is optimized for each EM image.

To quantify the similarity between the experimental EM image  $I(i, j)$  and the negative-stain model  $\rho_2(i, j)$ , we define a score by using zero-means normal cross-correlation(ZNCC) as follows:

$$Sc_2 = \frac{\sum_{i,j} (\rho_2(i,j) - \langle \rho_2 \rangle) (\mathbf{I}(i,j) - \langle \mathbf{I} \rangle)}{\sqrt{\sum_{i,j} (\rho_2(i,j) - \langle \rho_2 \rangle)^2 \sum_{i,j} (\mathbf{I}(i,j) - \langle \mathbf{I} \rangle)^2}}$$

Because ZNCC remains unaffected by the addition of a constant and multiplication with a positive constant,  $\rho_2(i,j)$  in the above equation can be replaced by  $-\rho_N(i,j)$ .

### 12.2.7 Strategy for Selecting the Best-Fitting Atomic Model

In the 2D hybrid analysis, two kinds of simulated models of EM images—the simple projection model and the negative-stain model—are built from each atomic model to select the best-fitting atomic model. We define the best-fitting atomic model as the one that produces the negative-stain model that is most similar to an experimental EM image, because the negative-stain model is more realistic than the simple projection model. However, because it is time consuming to build the negative-stain model, the simple projection model is also used to achieve faster computations in the following way.

1. By using the simple projection model, all the orientations of each atomic model are searched to compute  $Sc_1$  values.
2. The orientations with the local maxima of  $Sc_1$  are identified.
3. Negative-stain models are built for the orientations near the local and global maxima of  $Sc_1$ , apart from those with a small contact area.
4. The highest value of  $Sc_2$  is identified and used for comparison with other atomic models with different conformations to identify the best-fitting atomic model.

This strategy was developed by comparing the simple projection model and the negative-stain model in detail. For comparison, we built both of these models from the X-ray crystal structure of integrin in all possible orientations and we calculated the scores for the experimental EM images of clasped integrins in  $\text{Ca}^{2+}$  solution (Takagi et al. 2002) that had conformations similar to the X-ray crystal structure. By means of this comparison, we found that the global maxima of the two simulated models were not always observed in the same orientation. However, even when the two maxima were not in the same orientation, the global maximum of  $Sc_2$  was always observed near one of the local maxima of  $Sc_1$ . Therefore, we can find the global maximum of  $Sc_2$  by searching the orientations around the local and global maxima of  $Sc_1$ . In this way, we can reduce the number of computations for  $Sc_2$ , each of which requires a much longer time than the corresponding computation for  $Sc_1$ .

In addition, we computed the contact areas of the X-ray crystal structure in all possible orientations, and we observed that the atomic models had relatively

large contact areas when they had maximum values of  $S_{C2}$ . On the basis of this observation, we assume that the molecules contact the supporting film with relatively large contact areas. This also helps to reduce the number of computations by limiting the number of orientations of the atomic models.

### ***12.2.8 Expression and Purification of Integrins***

Soluble integrin heterodimers were constructed by using a previously described strategy (Takagi et al. 2001). Briefly, expression constructs for the  $\alpha$ -subunits contained the extracellular portion of the  $\alpha$ -chain (residues 1–960 for  $\alpha V$ ) followed by a 30-residue ACID-Cys peptide. Constructs for the  $\beta$ -subunits contained the extracellular portion of each  $\beta$ -chain (residues 1–691 for  $\beta 3$ ) followed by a tobacco etch virus (TEV) protease-recognition sequence, a 30-residue BASE-Cys peptide, and a hexahistidine tag. When combined, the C-terminal ACID-Cys and BASE-Cys segments formed an intersubunit disulfide-bridged  $\alpha$ -helical coiled coil (called a ‘clasp’), which could be released by treatment with TEV protease (Takagi et al. 2002). Combinations of the  $\alpha$  and  $\beta$  constructs were co-transfected into CHO Lec 3.2.8.1 cells to establish stable cell lines. Recombinant integrins were purified from the culture supernatants by immunoaffinity chromatography using anti-coiled-coil antibody 2H11 (Chang et al. 1994), followed by gel filtration on a Superdex 200 HR column (1.6  $\times$  60 cm, Pharmacia) equilibrated with 20 mM Tris, 150 mM NaCl, pH 7.5 (TBS) containing 1 mM  $\text{CaCl}_2$  and 1 mM  $\text{MgCl}_2$ . The peak fraction was concentrated to 1 mg/ml and stored at  $-80^\circ\text{C}$  until used.

### ***12.2.9 Electron Microscopy and Image Processing***

Approximately 10  $\mu\text{g}$  of each purified integrin was subjected to an additional gel-filtration process on a Superdex 200 HR column equilibrated with 50 mM Tris, 150 mM NaCl, pH 7.5, containing 5 mM  $\text{CaCl}_2$  or 1 mM  $\text{MnCl}_2$ . After gel filtration, the samples were immediately absorbed onto glow-discharged carbon-coated copper grids. Samples were negatively stained with 2.5% (w/v) uranyl acetate and examined under an electron microscope (H9500SD, Hitachi, Japan) operated at 200 kV with a nominal magnification of  $\times 80,000$ . Images were recorded on a 2048  $\times$  2048 CCD camera (TVIPS, Gauting, Germany). Single-particle EM analysis, including particle selection and 2D classification and averaging, was performed by using the EMAN suite (Ludtke et al. 1999) and IMAGIC program (van Heel et al. 1996). Particles were selected from individual frames (with an effective pixel size of 0.21 nm) by using the Boxer program in the EMAN suite. The particle images were

rotationally and translationally aligned by a multi reference alignment procedure, and subjected to multivariate statistical analysis by specifying 20 classes using the IMAGIC program.

## 12.3 Application of the 2D Hybrid Analysis

### 12.3.1 Electron Microscopy images of Integrins in $\text{Ca}^{2+}$ Solution

We applied the 2D hybrid analysis to 20 EM images of integrin in  $\text{Ca}^{2+}$  solution and we obtained the best-fitting atomic model for each EM image. As shown in Table 12.1, the scores for the best-fitting atomic models ( $Sc_2^{\text{max}}$ ) were generally high, suggesting that the models reproduced the EM images well. Actually, the X-ray crystal structure without deformation fitted well to many of the EM images, as indicated by the scores ( $Sc_2^0$ ). In such cases, the best-fitting models were not

**Table 12.1** Summary of the analysis of EM images for integrins in  $\text{Ca}^{2+}$  solution by using the X-ray crystal structure and the best-fitting atomic models

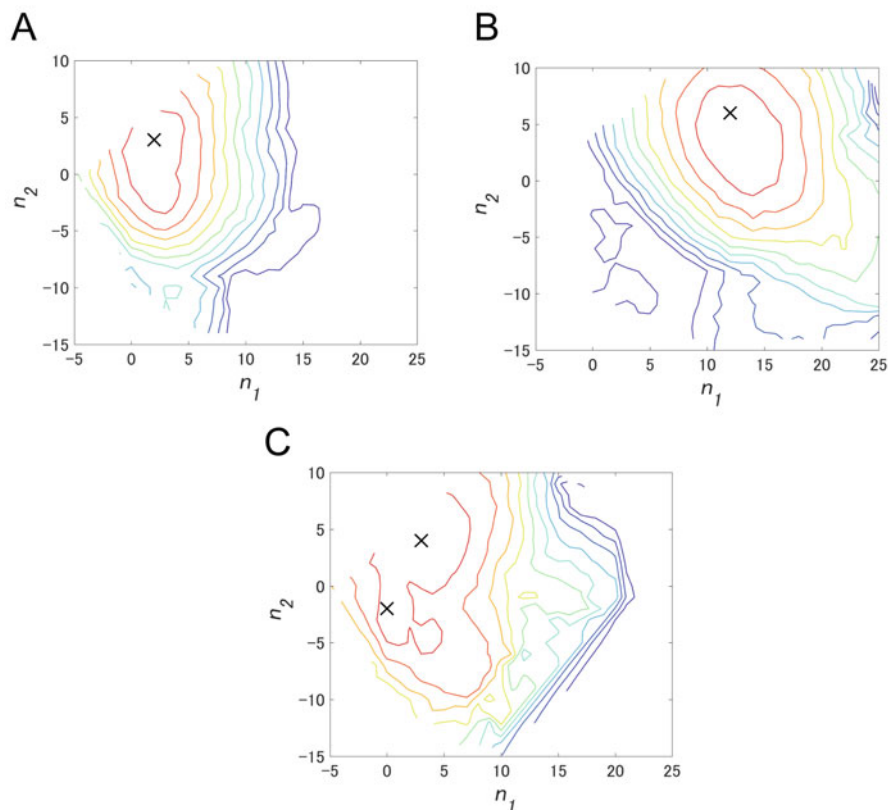
| Name   | $Sc_2^0$ | $Sc_2^{\text{max}}$ | $\Delta Sc_2^0(\%)^a$ | RMSD( $\text{\AA}$ ) <sup>b</sup> |
|--------|----------|---------------------|-----------------------|-----------------------------------|
| Ca-001 | 0.823    | 0.840               | 2.0                   | 19.8                              |
| Ca-002 | 0.861    | 0.868               | 0.9                   | 14.7                              |
| Ca-003 | 0.888    | 0.925               | 4.2                   | 6.1                               |
| Ca-004 | 0.884    | 0.901               | 1.9                   | 9.4                               |
| Ca-005 | 0.867    | 0.881               | 1.6                   | 7.8                               |
| Ca-006 | 0.923    | 0.933               | 1.2                   | 4.3                               |
| Ca-007 | 0.865    | 0.870               | 0.6                   | 2.3                               |
| Ca-008 | 0.802    | 0.887               | <b>10.6</b>           | 9.2                               |
| Ca-009 | 0.866    | 0.892               | 2.9                   | 5.0                               |
| Ca-010 | 0.909    | 0.931               | 2.5                   | 6.4                               |
| Ca-011 | 0.833    | 0.902               | 8.3                   | 8.2                               |
| Ca-012 | 0.863    | 0.881               | 2.1                   | 11.3                              |
| Ca-013 | 0.833    | 0.835               | 0.3                   | 2.0                               |
| Ca-014 | 0.797    | 0.908               | <b>13.8</b>           | 15.4                              |
| Ca-015 | 0.917    | 0.924               | 0.7                   | 4.1                               |
| Ca-016 | 0.883    | 0.903               | 2.3                   | 7.4                               |
| Ca-017 | 0.871    | 0.878               | 0.7                   | 3.0                               |
| Ca-018 | 0.824    | 0.859               | 4.2                   | 4.7                               |
| Ca-019 | 0.854    | 0.899               | 5.3                   | 8.8                               |
| Ca-020 | 0.812    | 0.912               | <b>12.4</b>           | 13.4                              |

Matsumoto et al. (2017)

<sup>a</sup> $(Sc_2^{\text{max}} - Sc_2^0)/Sc_2^0$ . Values larger than 10 appear in boldface

<sup>b</sup>RMSD of best-fitting atomic model from X-ray crystal structure





**Fig. 12.3** Contour maps of  $Sc_2$  scores for three EM images, (a) for Ca-006, (b) for Ca-020, and (c) for Ca-002 in Fig. 12.5 plotted as a function of the index numbers  $n_1$  and  $n_2$  for deformed atomic models  $\mathbf{r}^{12}(n_1, n_2)$ . The origin (0,0) corresponds to the X-ray crystal structure. The contour lines are drawn at intervals of 0.01, starting from the maximum scores. The peaks are indicated by crosses. (Matsumoto et al. 2017)

too different from the X-ray crystal structure, as indicated by the small root-mean-square deviation (RMSD). Corresponding to small values of the RMSD, the increments in the scores from those of the X-ray crystal structure ( $Sc_2^0$ ) were generally not very large (the average increments were 4%). However, there were cases in which the increments were more than 10% (written as bold numerals in Table 12.1). In such cases, the RMSDs were relatively large, and the X-ray crystal structure often incorrectly fitted the EM images (data not shown), indicating that the fitting was sensitive to conformational changes in the atomic model.

To examine how fitting was dependent on the conformation, we computed the  $Sc_2$  scores for a range of atomic models  $\mathbf{r}^{12}(n_1, n_2)$ , built by deforming the X-ray crystal structure along the two lowest-frequency normal modes. These are shown as contour maps in Fig. 12.3.

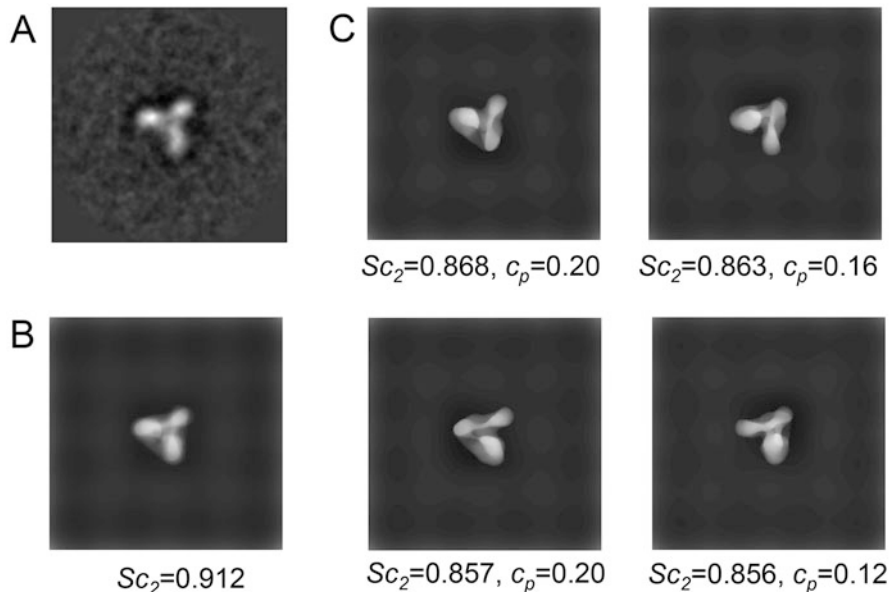
For about half of the EM images, we obtained contour maps with a single peak surrounded by crowded contour lines (see Fig. 12.3a, b), suggesting that the score decreased rapidly as the conformation deviated from the peak. Figure 12.3a shows the contour map for an EM image that was reproduced quite well by the X-ray crystal structure, whereas Fig. 12.3b shows the contour map for an image that was reproduced well only by atomic models that differed markedly from the crystal structure. Clearly, the peak was closer to the origin in Fig. 12.3a than in Fig. 12.3b, where the origin corresponded to the X-ray crystal structure. This result therefore shows that it is important to use an appropriate atomic model to achieve a good fit. In other words, this result shows that it is possible to identify a unique atomic model by the proposed 2D hybrid analysis approach.

### 12.3.2 Improperly Averaged Electron Microscopy Images

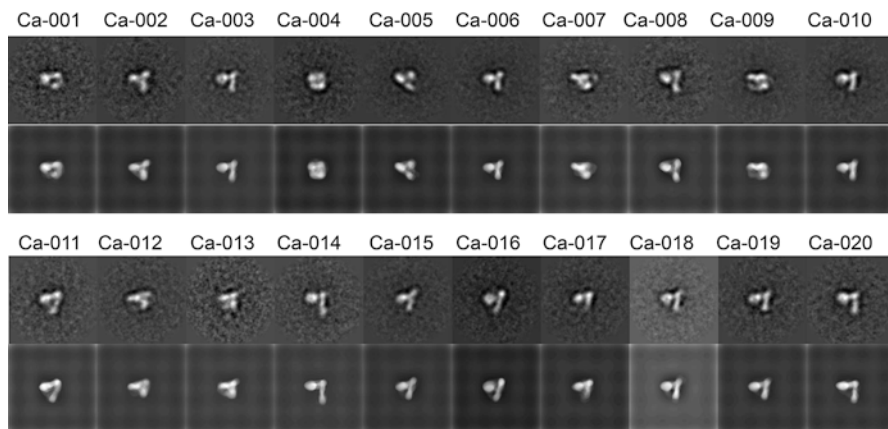
For the remaining EM images, multiple peaks appeared to be present in the contour maps (Fig. 12.3c), indicating that many conformations fitted well. The EM images studied here were averaged images and, in principle, the averaging should have been performed by using raw images of molecules with the same conformation and orientation. However, this is actually a difficult task, as described in the introduction.

This contour map suggests that raw images of molecules with relatively large differences in conformations or orientations were averaged. Indeed, Fig. 12.4 demonstrates how the averaging of the various negative-stain models reproduced the EM image quite well. The contour map in Fig. 12.3c is for the EM image shown in Fig. 12.4a. The negative-stain model built from the best-fitting model to this image is shown in the upper left-hand corner of Fig. 12.4c, and is not very similar to the EM image. Other models in Fig. 12.4c were built from the conformations that corresponded to peaks in the contour map. Actually, the peaks were selected not only from  $\mathbf{r}^{12}(n_1, n_2)$ , but also from the entire range of conformations. By combining these negative-stain models with the different weights ( $c_p$ ), we were able to obtain the combined (averaged) negative-stain model (Fig. 12.4b), which appeared more similar to the EM image than did any negative-stain model of the peak conformations.

We performed the same analysis on other EM images of integrins in  $\text{Ca}^{2+}$  solution; the results are summarized in Table 12.2, and the combined negative-stain models are shown in Fig. 12.5. There were several cases in which relatively large increments of the score ( $\Delta S_{c2}$ ) resulting from combinations of the negative-stain models were observed (written as bold numerals in Table 12.2). In such cases, a number of peak conformations were observed, although many of them made only a small contribution (small  $c_p$  values), as indicated by the numerals in parentheses. Actually, each averaged EM image was reproduced relatively well by a much smaller number of negative-stain models. In Table 12.2, the minimum number of negative-stain models required to achieve 99% of  $S_{c2}^{multi}$  is listed as  $n_c^{99}$  for each EM image. This number correlated well with  $\Delta S_{c2}$ .



**Fig. 12.4** Demonstration of how a combination of negative-stain models closely reproduced an EM image. (a) EM image of integrin in  $Ca^{2+}$  solution (Ca-002 in Fig. 12.5). (b) Combined negative-stain model. (c) Negative-stain models of peak conformations used to build the model in (b). Only those models with weighting factor  $c_p > 0.1$  are shown. Values of  $Sc_2$  and  $c_p$  are given beneath each negative-stain model in (b) and (c). (Matsumoto et al. 2017)



**Fig. 12.5** Combined negative-stain models for reproducing experimental EM images of integrin in  $Ca^{2+}$  solution. The experimental EM image is shown above each model for comparison. A label for each EM image is also given. (Matsumoto et al. 2017)

**Table 12.2** Summary of combinatorial analyses of EM images for integrins in  $\text{Ca}^{2+}$  solution

| Name   | $Sc_2^{\text{multi}}$ | $\Delta Sc_2(\%)^{\text{a}}$ | Number of peaks <sup>b</sup> |      | $n_c^{99}$ |
|--------|-----------------------|------------------------------|------------------------------|------|------------|
| Ca-001 | 0.894                 | <b>6.4</b>                   | 42                           | (9)  | 4          |
| Ca-002 | 0.912                 | <b>5.0</b>                   | 50                           | (7)  | 2          |
| Ca-003 | 0.931                 | 0.6                          | 8                            | (8)  | 1          |
| Ca-004 | 0.937                 | <b>4.0</b>                   | 116                          | (6)  | 4          |
| Ca-005 | 0.932                 | <b>5.8</b>                   | 45                           | (6)  | 3          |
| Ca-006 | 0.947                 | 1.5                          | 15                           | (10) | 2          |
| Ca-007 | 0.912                 | <b>4.8</b>                   | 59                           | (6)  | 3          |
| Ca-008 | 0.910                 | 2.6                          | 5                            | (4)  | 2          |
| Ca-009 | 0.930                 | <b>4.3</b>                   | 45                           | (4)  | 3          |
| Ca-010 | 0.944                 | 1.4                          | 15                           | (10) | 2          |
| Ca-011 | 0.925                 | 2.5                          | 13                           | (12) | 3          |
| Ca-012 | 0.917                 | <b>4.0</b>                   | 46                           | (6)  | 3          |
| Ca-013 | 0.864                 | 3.5                          | 77                           | (5)  | 3          |
| Ca-014 | 0.912                 | 0.4                          | 4                            | (3)  | 1          |
| Ca-015 | 0.934                 | 1.1                          | 19                           | (14) | 2          |
| Ca-016 | 0.933                 | 3.3                          | 18                           | (10) | 3          |
| Ca-017 | 0.916                 | <b>4.4</b>                   | 47                           | (4)  | 4          |
| Ca-018 | 0.883                 | 2.8                          | 7                            | (7)  | 3          |
| Ca-019 | 0.913                 | 1.6                          | 15                           | (13) | 2          |
| Ca-020 | 0.921                 | 1.0                          | 10                           | (8)  | 2          |

Matsumoto et al. (2017)

<sup>a</sup> $(Sc_2^{\text{multi}} - Sc_2^{\text{max}})/Sc_2^{\text{max}} \times 100$ , where  $Sc_2^{\text{max}}$  is listed in Table 12.1. Values of 4 or more appear in boldface

<sup>b</sup>The number of peaks whose coefficients ( $c_p$ ) were larger than 0.01 is given in parentheses

Note that our result differed from that of the so-called ‘Einstein-from-noise’ (Henderson 2013; van Heel 2013), which describes how any image can be reproduced by averaging many noise images. This phenomenon occurs because noise images are uncorrelated to each other. Thus, the more noise images we use, the better the averaged images we get. On the other hand, the negative-stain models of the peak conformations were strongly correlated to each other, because they were similar to the targeted EM image. Furthermore, we needed to combine only a few images at most to reproduce the averaged EM images well, and further increments in the number of images produced little improvement (data not shown).

## 12.4 Concluding Remarks

We have developed an approach for building atomic models that reproduce the EM images of proteins. In this approach, many atomic models with different conformations are initially prepared. These are obtained by performing a computer simulation using the X-ray crystal structure or the modeled structure as the initial

model. For integrin, we performed a normal-mode analysis of the elastic network model. However, other computational methods can also be employed. The use of finer simulations, such as all-atom molecular dynamics simulations, should increase the reliability of the results. Simulated EM images are then produced from each atomic model and these are compared with the experimental EM images to select the best-fitting atomic model. We use two kinds of models as the simulated EM images: the negative-stain model and the simple projection model. The former model is more realistic, but building it requires a longer computational time. Therefore, the latter model is used to produce a series of candidate atomic models in a shorter computational time.

The use of the negative-stain model enables us to analyze the averaged EM images in detail. Originally, we intended to use the 2D hybrid analysis to find the best-fitting atomic model for each EM image, i.e., one atomic model for one image. However, we often encountered cases where an averaged EM image could not be reproduced by a single atomic model. Instead, it was reproduced well when we combined multiple negative-stain models produced from atomic models with rather different conformations. This indicates that great care must be taken in interpreting an averaged EM image, because two or more different conformations might be mixed in the image. Also, it indicates that our proposed approach can detect such mixtures.

**Acknowledgments** Molecular graphics were performed using the UCSF Chimera package (Pettersen et al. 2004). This work was supported by the Platform Project for Supporting Drug Discovery and Life Science Research(Platform for Drug Discovery, Informatics, and Structural Life Science) from the Japan Agency for Medical Research and Development (AMED).

## References

- 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
- Burgess SA, Walker ML, White HD, Trinick J (1997) Flexibility within myosin heads revealed by negative stain and single-particle analysis. *J Cell Biol* 139(3):675–681
- Chang HC, Bao Z, Yao Y, Tse AG, Goyarts EC, Madsen M, Kawasaki E, Brauer PP, Sacchettini JC, Nathenson SG, Reinherz EL (1994) A general method for facilitating heterodimeric pairing between two proteins: application to expression of  $\alpha$  and  $\beta$  T-cell receptor extracellular segments. *Proc Natl Acad Sci U S A* 91(24):11408–11412
- Henderson R (2013) Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proc Natl Acad Sci U S A* 110(45):18037–18041
- Ludtke SJ, Baldwin PR, Chiu W (1999) EMAN: semiautomated software for high-resolution single-particle reconstructions. *J Struct Biol* 128(1):82–97
- Marabini R, Carazo JM (1994) Pattern recognition and classification of images of biological macromolecules using artificial neural networks. *Biophys J* 66(6):1804–1814
- Matsumoto A, Ishida H (2009) Global conformational changes of ribosome observed by normal mode fitting for 3D Cryo-EM structures. *Structure* 17(12):1605–1613
- Matsumoto A, Kamata T, Takagi J, Iwasaki K, Yura K (2008) Key interactions in integrin ectodomain responsible for global conformational change detected by elastic network normal-mode analysis. *Biophys J* 95(6):2895–2908

- Matsumoto A, Miyazaki N, Takagi J, Iwasaki K (2017) 2D hybrid analysis: approach for building three-dimensional atomic model by electron microscopy image matching. *SciRep* 7(1):377
- Miyashita O, Onuchic JN, Wolynes PG (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci U S A* 100(22):12570–12575
- Pettersen 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
- Sadourny R, Arakawa A, Mintz Y (1968) Integration of the nondivergent barotropic vorticity equation with an icosahedral-hexagonal grid for the sphere. *Mon Weather Rev* 96(6):351–356
- Schroder GF (2015) Hybrid methods for macromolecular structure determination: experiment with expectations. *Curr Opin Struct Biol* 31:20–27
- Takagi J, Erickson HP, Springer TA (2001) C-terminal opening mimics ‘inside-out’ activation of integrin  $\alpha 5\beta 1$ . *Nat Struct Mol Biol* 8(5):412–416
- Takagi J, Petre BM, Walz T, Springer TA (2002) Global conformational rearrangements in integrin extracellular domains in outside-in and inside-out signaling. *Cell* 110(5):599–511
- Tama F, Brooks CL 3rd (2005) Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis. *J Mol Biol* 345(2):299–314
- Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* 77(9):1905–1908
- Tsukasaki Y, Miyazaki N, Matsumoto A, Nagae S, Yonemura S, Tanoue T, Iwasaki K, Takeichi M (2014) Giant cadherins fat and Dachsous self-bend to organize properly spaced intercellular junctions. *Proc Natl Acad Sci U S A* 111(45):16011–16016
- van Heel M (2013) Finding trimeric HIV-1 envelope glycoproteins in random noise. *Proc Natl Acad Sci U S A* 110(45):E4175–E4177
- van Heel M, Harauz G, Orlova EV, Schmidt R, Schatz M (1996) A new generation of the IMAGIC image processing system. *J Struct Biol* 116(1):17–24
- Villa E, Lasker K (2014) Finding the right fit: chiseling structures out of cryo-electron microscopy maps. *Curr Opin Struct Biol* 25:118–125
- Xiong J-P, Stehle T, Diefenbach B, Zhang R, Dunker R, Scott DL, Joachimiak A, Goodman SL, Arnaout MA (2001) Crystal structure of the extracellular segment of integrin  $\alpha V\beta 3$ . *Science* 294(5541):339–345

**Part III**  
**New Computational Tools Enabling**  
**Hybrid Methods**

# Chapter 13

## Hybrid Methods for Macromolecular Modeling by Molecular Mechanics Simulations with Experimental Data



Osamu Miyashita and Florence Tama

**Abstract** Hybrid approaches for the modeling of macromolecular complexes that combine computational molecular mechanics simulations with experimental data are discussed. Experimental data for biological molecular structures are often *low-resolution*, and thus, do not contain enough information to determine the atomic positions of molecules. This is especially true when the dynamics of large macromolecules are the focus of the study. However, computational modeling can complement missing information. Significant increase in computational power, as well as the development of new modeling algorithms allow us to model structures of biological macromolecules reliably, using experimental data as references. We review the basics of molecular mechanics approaches, such as atomic model force field, and coarse-grained models, molecular dynamics simulation and normal mode analysis and describe how they could be used for *flexible fitting* hybrid modeling with experimental data, especially from cryo-EM and SAXS.

**Keywords** Cryo-EM · SAXS · Normal mode analysis · Molecular dynamics simulations · Coarse-grained models · Fitting · Modeling

### 13.1 Hybrid Approach for Structure Modeling from Low Resolution, Low Information Experimental Data

Three-dimensional structures of biomolecules provide critical information to elucidate their mechanism. X-ray crystallography has been the major approach that provides the detailed atomic resolution structural information of the molecules

---

O. Miyashita  
RIKEN R-CCS, Kobe, Hyōgo, Japan

F. Tama (✉)  
RIKEN R-CCS, Kobe, Hyōgo, Japan

Department of Physics and ITbM, Nagoya University, Nagoya, Japan  
e-mail: [florence.tama@nagoya-u.jp](mailto:florence.tama@nagoya-u.jp)

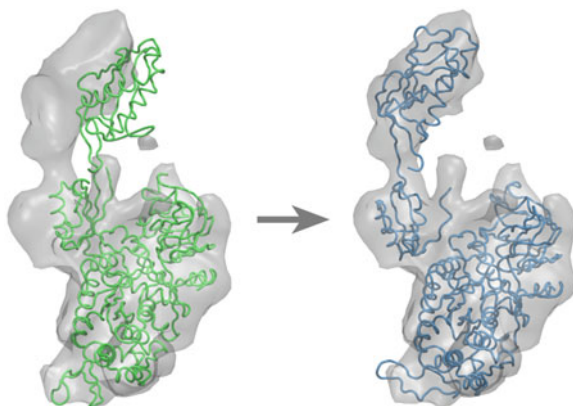
© Springer Nature Singapore Pte Ltd. 2018  
H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,  
Advances in Experimental Medicine and Biology 1105,  
[https://doi.org/10.1007/978-981-13-2200-6\\_13](https://doi.org/10.1007/978-981-13-2200-6_13)



(Garman 2014). However, the fundamental requirement for this approach – crystallization of the molecular complexes – often makes its application to large dynamic systems a significant challenge. Thus, complementary information from other experiments is important (Lander et al. 2012). However, such information is typically low-resolution, i.e., atomic details cannot be directly obtained from the data itself. For example, spectroscopy experiments do provide temporal information, but often only for specific parts of the system. Small angle X-ray scattering (SAXS) provides information of structures in solvent condition near native environment, but only as a one-dimensional profile related to the distribution of atom-pair distances (see another Chapter). Cryo-electron microscopy (EM) has been garnering attention due to the technological advances, which allow the 3D model reconstruction of large macromolecules at near atomic level resolution (Frank 2017). Still, the resolution is typically not high enough for *ab initio* structure modeling. There is great potential for EM technology to improve and raise the resolution limit (see another Chapter). However, since an advantage of the EM method is to capture functional states and dynamics of biomolecules, certain fractions of the resulting data will continue to be low-resolution due to heterogeneity. Moreover, the experimental data from cryo-EM is a collection of 2D snapshots of single particles, which can potentially provide a wealth of information about the structural heterogeneity in conformational ensembles related to function. Lastly, X-ray free electron laser (XFEL) is an exciting new development. Using its extremely bright X-ray pulse, biomolecular complexes can be observed without the need for the samples to be crystallized (Gallagher-Jones et al. 2016; Barty 2016; Miyashita and Joti 2017; Miao et al. 2015). While XFEL is still in the development phase in terms of both experimental and computational techniques, the field is advancing and more results are being reported.

Hybrid approaches aim to combine multiple experimental data at a variety of resolutions and details to obtain a comprehensive picture of biological molecules' structure and dynamics in order to reveal the mechanistic details of their functions (Lander et al. 2012). Computation is an essential part of this process. Biological molecules have complex structures and it is difficult to predict their expected conformational transitions by mere visual inspection. Computational models that define the mechanical property of the structures based on chemistry, physics and numerical algorithms to predict natural motions are essential for accurate modeling. Mathematical descriptions need to be established to integrate multiple experimental data into the modeling procedures. In this chapter, we will review the approaches for *flexible fitting*, where a known crystal or modeled structure is used as the starting point and to create new model structures that are consistent with experimental data utilizing molecular mechanics simulations of conformational transitions. We will focus on cases where the structural components are already assembled. Methods to assemble the biomolecular complexes from subcomponents based on experimental data are discussed in other chapters.

**Fig. 13.1** An example of MD based flexible fitting. An X-ray structure of elongation factor 2 (EF2) is not in agreement with EM map. Using MD based flexible fitting, the conformational change of EF2 was simulated to construct a structural model that agrees with the EM map (Miyashita et al. 2017). (Image created by VMD Humphrey et al. 1996)



## 13.2 Why Flexible Fittings?

Cryo-EM is becoming an important tool for structural biology. It does not require samples to be crystallized. The resolution of the results from cryo-EM used to be relatively low ( $\sim 10$  Å), but it was still very informative for studying large macromolecules that are beyond the reach of X-ray crystallography. Recent technological advances as well as development of software that sort out the noise and conformational heterogeneity in the data extended the limit of cryo-EM, and now the method provides atomic level resolutions (discussed in another Chapter in this book).

Yet, achieving atomic resolutions is not the only goal of cryo-EM studies. When functionally important dynamics of macromolecular complexes are studied, even low-resolution maps are valuable, since it may provide crucial information to construct hypotheses of mechanisms related to function. Indeed, Cryo-EM experiments can often capture more functional conformations than X-ray crystallography (for example (Unverdorben et al. 2014)). The flexible fitting approach is most useful in such a context. In cases where detailed atomic structure of one conformation is available from X-ray crystallography or modeling, while cryo-EM data provides information on other functional, details of the new conformational states can be modeled using flexible fitting approaches (Fig. 13.1).

## 13.3 Strategy Used for Flexible Fitting

The conceptually simplest form of the flexible fitting approach would be – (1) to generate a large number of models that are mechanistically and biochemically sound, and then (2) among these, to select the structures that are in agreement with available experimental data. Here, efficiency to generate candidate structures, or sampling efficiency, is a critical part of the algorithm, since the targets of hybrid

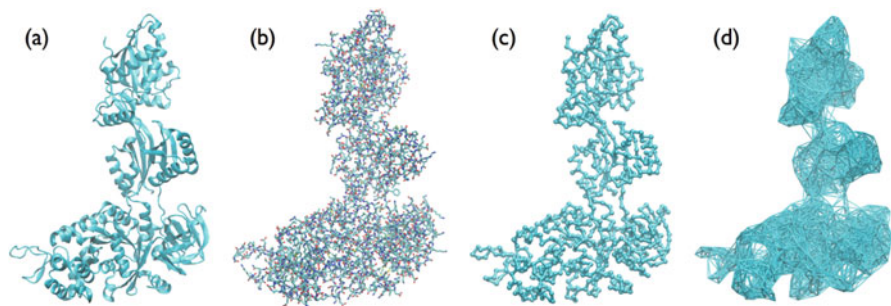
modeling tend to be large macromolecular complexes. Therefore, the commonly used algorithms employ (3) various techniques to focus the sampling to conformations that agree well with experimental data set. We discuss such algorithms in the following section.

### 13.4 Algorithms to Generate Candidate Structures

In this section, we will describe algorithms to generate candidate structures that are chemically sound and biochemically meaningful. A widely known approach for such conformational sampling is molecular dynamics (MD) simulation (Perilla et al. 2015). In MD, a set of equations and parameters to describe the energetics of molecular structures (usually via classical mechanics) are defined based on theoretical considerations and calibrations to reproduce experimental data. The major part of such potential energy function consists of the energetic terms, such as covalent bond energy,  $U_{\text{bond}}$ , angle energy to keep correct angles between two covalent bonds,  $U_{\text{angle}}$ , dihedral term to reproduce commonly observed dihedral angles,  $U_{\text{dihedral}}$ . In addition, non-bonded interactions such as electrostatic interaction between atomic charges,  $U_{\text{elec}}$ , and repulsive and weakly attractive vdw interactions,  $U_{\text{vdw}}$ , are also considered (Fig. 13.2b). Thus, the total energy is defined as:

$$U_{\text{mol}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{elec}} + U_{\text{vdw}}$$

Note that these are all functions of atomic coordinates,  $\mathbf{x}$ , as  $U_{\text{mol}}(\mathbf{x})$ . Several sets of equations and parameters have been historically developed and each parameter set (also called “force field”) uses slightly different equations and other modifications to the potential energy functions (Case et al. 2005; Huang et al. 2017). With this function, forces on each atom resulting from interactions are calculated as,



**Fig. 13.2** EF2 shown in different models (a) Ribbon model. (b) All-atom model; hydrogen atoms are not shown. For MD simulations, solvent molecules are added. (c)  $C\alpha$  model. In this model, only  $C\alpha$  atoms are considered and pseud-bonds and angles are used to approximately simulate the dynamics of protein molecules. (d) Elastic network model. All pairwise atomic interactions are approximated as springs

$$F = -dU_{\text{mol}}(\mathbf{x})/d\mathbf{x}$$

where the motions of atoms are estimated using simple Newton equation. However, this can be done only in an incremental manner (stepwise numerical integration), and one step typically advances the motions of the atoms by just 1–2 femto-second, requiring extensive computational time to obtain large scale conformational changes. Yet, it is probably the most reliable method that can generate the structures that are physicochemically valid.

Here, sampling inefficiency poses as a critical issue. Even though each structure is accurate, if the MD fails to sample the structures that are represented by the experimental data, it does not serve the purpose of molecular modeling. A variety of techniques have been proposed to enhance the sampling efficiency and those are also incorporated into the applications for modeling.

One approach to speed up conformational sampling is MD simulation with coarse-grained models (Saunders and Voth 2013; Takada et al. 2015). In these models, not all the atoms in the system are explicitly considered in the simulation; some groups of atoms are combined and represented as one pseudo-atom; for example, one residue can be represented by one pseudo-atom (Fig. 13.2c). In addition, solvent molecules are not usually considered. Standard all-atom MD simulation requires water molecules to be explicitly included, since the current force fields are not designed to run simulations in vacuum. Simply not considering water molecules in coarse-grained models significantly reduces the number of force calculations and speeds up the MD simulations. Obviously, this approach cannot produce atomic level detailed structural models as the end results; however, it is often useful since experimental data itself may not have enough information to support the atomic level details.

Even further coarse-graining has been employed for *flexible fitting*. One such approach is through elastic network models (Tirion 1996). In elastic network models, atomic details are completely discarded – the molecular structure is represented by a group of pseudo-atoms, each representing a few to several atoms, and interactions between these pseudo-atoms are modeled by simple harmonic potential. Although some variations exist, in its original form, all pseudo-atoms are treated equally and all harmonic potentials are defined so that its energy minimum conformation is its original structure.

$$U = \sum_{i,j,r_{ij}<R} \frac{1}{2}k(r_{ij} - r_{ij}^0)^2$$

where  $R$  is a cutoff parameter and interactions are defined only for the atomic pairs that are within the cutoff. In other words, the molecular complex is considered as an “elastic” object that could deform. Despite its simplicity, it has been shown that such potential function is sufficient to simulate large scale conformational transitions (Tama and Sanjoud 2001).

Dynamics of the molecule is then simulated typically using normal mode analysis with this potential (Tirion 1996; Tama and Sanejouand 2001; Mahajan and Sanejouand 2015). In normal mode analysis, the potential energy surface around the original structure is examined, and dynamics is represented as a combination of normal mode coordinate,  $\mathbf{q}$ , associated with normal mode vectors,  $\mathbf{a}_l$ .

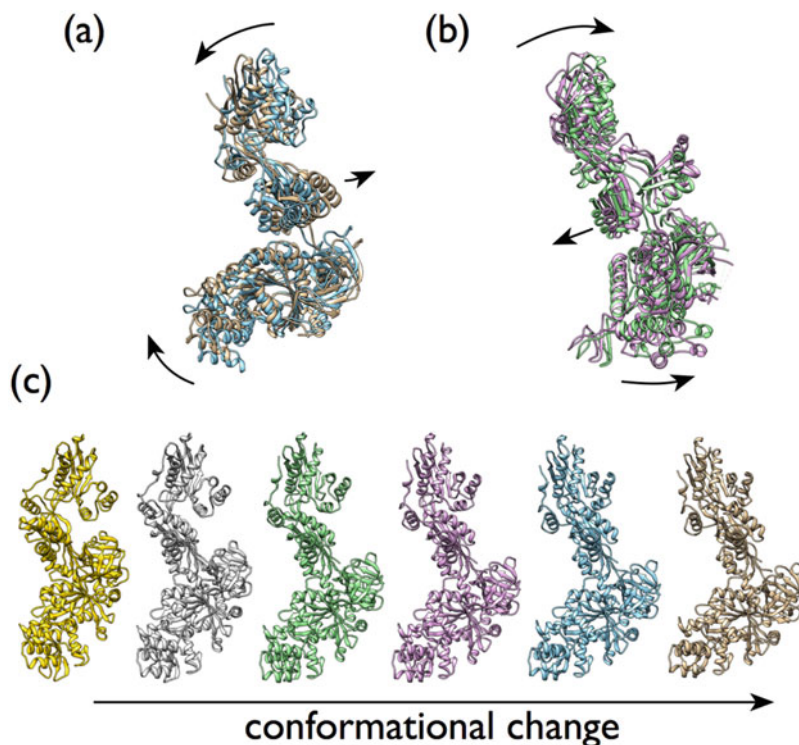
$$x^n(\mathbf{q}) = \sum_{l=1}^M a_l^n q_l + x_0^n$$

where  $\mathbf{a}_l = \{a_l^n\}$ ,  $\mathbf{q} = \{q_l\}$  and  $x_0^n$  represents the original coordinate of atom  $n$ . With this equation, the motions of the atoms are represented as a set of collective motions represented as normal modes. Normal modes,  $\mathbf{a}_l$ , are usually sorted in the way that  $l = 1$  corresponds to the lowest frequency mode, i.e., most flexible conformational changes, and then higher  $l$  for higher frequency modes.  $M$  is a parameter to choose how many normal modes are used to represent the motions and typically 10 lowest frequency modes are sufficient to describe important motions of proteins. In contrast to MD simulations, this is computationally efficient, and thus it was used in several important large systems for flexible fitting (Tama et al. 2003). Computations could further be accelerated by segmenting the structure into rigid body blocks of such as residues or domains (Tama et al. 2000) (Fig. 13.3).

With recent increase in computational power, the advantage of such simplicity is lessened, but it is still valuable because it can be defined for almost any system quickly, regardless of chain connectivity or missing residues. Often the original structures have missing residues and structural components, due to the large size of the system often studied in cryo-EM, and the preparation of all atom force fields are often not a simple task for such large systems.

Furthermore, this model can be applied to molecular systems with no atomic model. For example, it can be applied to the 3D volume map from EM reconstructions (Tama et al. 2002; Jin et al. 2014). Inside the continuous 3D map, a set of pseudo-atoms can be placed so that they represent the density of the volume as closely as possible. Then normal mode analysis with elastic network model can be applied to simulate expected dynamics purely based on the shape of the system. This approach has been used to analyze the conformational variations from single EM dataset, by generating variations of 3D models from a tentative, averaged, map (see following Section).

There are also other approaches to generate structural models using simplified potentials. DireX uses an elastic network model with iteratively updated distance restraint and random walk displacements to generate fitted structures (Schröder et al. 2007). YUP.SCX also uses a pair-wise distance potential, which are atom independent but more complex than the elastic network model. Structures are then deformed to fit the experimental data using simulated annealing optimization (Tan et al. 2008).



**Fig. 13.3** Examples of normal mode analysis. **(a)** Two structures represent the conformational change represented by the lowest frequency (softest) mode of EF2. **(b)** The second lowest frequency mode. The molecule is rotated to show the motions. Calculations using ElNemo (Suhre and Sanejouand 2004). **(c)** An example of conformational change of EF2 simulated using iterative NMA (Miyashita et al. 2003). (Images by Chimera Pettersen et al. 2004)

### 13.5 Quantification of Structure-Data Agreement

The goal of flexible fitting approach based on dynamics simulations is to identify structures that agree with experimental data and propose them as possible models for further investigation. In other chapters, a variety of experimental techniques to study biological structures are discussed.

For the modeling purpose, one essential requirement is that experimental data can be computationally simulated from a given model, at least approximately. This is not always trivial. Experimental data that describes some distance information between some atom groups, such as FRET or cross-linking, have often been used in MD simulations. Generally, it is rather straightforward to apply a constraint to

an MD simulation so that the distances between defined atom groups stay within a certain range. However, exact distances are not obtainable from experimental data and the data may represent the average distance between various conformations. Such uncertainty needs to be taken into account for the modeling.

In another example, SAXS profile contains information regarding the atomic-pair distances, but reported profiles are the difference between the one of proteins in solution and the one of pure solvent (Kikhney and Svergun 2015). Thus, the algorithms to simulate SAXS profile from protein structures need to model the scattering from solvent atoms implicitly (Svergun et al. 1995; Nguyen et al. 2014). Alternatively, large scale MD simulations with solvent molecules need to be performed (Merzel and Smith 2002; Oroguchi and Ikeguchi 2011). As such, some experimental data are not easy to incorporate into the modeling (H/D exchange data as another example).

In addition, some experimental data are not easy to be implemented into MD based flexible fitting approaches. Randomly generating a large number of structures and then finding ones that agree with experimental data is possible but not an efficient approach. Especially for large macromolecular complexes where hybrid approaches are often employed, the sampling could be a serious issue. Thus, the more efficient approach is to guide the conformational changes toward the structures that agree with experimental data. A common approach is the use of *biasing potential*,  $U_{\text{bias}}$ , which is included in MD simulation as an additional virtual potential energy:

$$U(\mathbf{x}) = U_{\text{mol}}(\mathbf{x}) + U_{\text{bias}}(\mathbf{x})$$

By adjusting such virtual forces, conformations that agree with experimental data could be generated. To implement this approach to MD simulation, we need to calculate derivatives of the scoring function as function of atomic coordinates or collective coordinates, such as normal modes. Not only for MD, but also for other optimization techniques, derivatives can make the computation significantly faster. However, such derivatives are difficult to calculate for the scoring functions with some experimental data.

There are also various approaches that employ Monte-Carlo type algorithms instead of MD. These algorithms focus on the generation of energetically accessible conformations rather than dynamics simulations. Choices of trial moves are not straightforward, but several algorithms have been successfully applied to generate fitted conformations efficiently. Such techniques are described in other chapters.

We will focus on MD based flexible fitting using 3D maps from cryo-EM. We will also briefly discuss the flexible fitting algorithms against SAXS data. In addition, multiple experimental data can be simultaneously used for modeling of complex systems (Fritz et al. 2013).

### 13.6 Flexible Fitting Against EM Data Using Elastic Network Normal Mode Analysis

In an early work, normal mode analysis with elastic network models was used to perform flexible fitting for large molecular complexes (Tama et al. 2004). Elastic network models could be applied to the systems at virtually any size even with limited computational resources, by adjusting the level of coarse-graining, i.e., the size of atom groups each pseudo-atom represents. The simple form of the potential energy function allows computations quickly performed using analytical equations. The similarity score between the model and experimental map,  $\rho_{\text{exp}}$ , is defined as a conventional correlation coefficient,

$$\text{CC} = \frac{\sqrt{\sum_i^N \rho_{\text{sim}}(i)\rho_{\text{exp}}(i)}}{\sqrt{\sum_i^N \rho_{\text{sim}}(i)^2}\sqrt{\sum_i^N \rho_{\text{exp}}(i)^2}} \quad (13.1)$$

where,  $\rho(i)$  is the density value of voxel  $i$ , and  $N$  is the number of voxels in the map. CC approaches 1 when the simulated map and experimental data are in agreement. Electron density map from the atomic model,  $\rho_{\text{sim}}$ , is generated using Gaussian kernels,  $g$ , placed on atom positions:

$$\rho_{\text{sim}}(i) = \sum_{n=1}^N \int_{V(i)} g(x, y, z; x^n, y^n, z^n) dx dy dz$$

$$g(x, y, z; x^n, y^n, z^n) = \exp\left[-\frac{3}{2\sigma^2} \left\{ (x - x^n)^2 + (y - y^n)^2 + (z - z^n)^2 \right\}\right]$$

where  $\mathbf{x}^n = (x^n, y^n, z^n)$  is the position of  $n$ th (pseudo-) atom.  $\sigma$  is a parameter that adjusts the width of Gaussians and is selected based on the resolution of the target experimental map. With these definitions, correlation coefficient is an analytical function of atomic coordinates and thus its derivatives can be calculated using analytical equations, allowing optimization with efficient algorithms:

$$\text{CC}(q) = \sum_{l=1}^M F_l q_l + \text{CC}(0)$$

where  $F_l = \partial\text{CC}/\partial q_l$  is the derivative of CC by normal mode coordinate  $q_l$ . This equation can be used to estimate the increase in correlation coefficient that can be expected by deforming the structure following a given normal mode vector. Typically, 10 normal modes are sufficient to represent expected large scale conformational transitions. Using the derivative values, a structure can be deformed so that the correlation coefficient is maximally increased by a small amount of



conformational deformations. Here, structure optimization is performed iteratively; normal mode analysis describes the conformational changes as “linear vectors” that represent the motion of atoms, but actual motions in biological molecules are quite nonlinear, including hinge bending motions, rotations and twisting motions. By performing normal mode analysis iteratively, such nonlinear motions could be simulated with reasonable accuracy (Miyashita et al. 2003). Typically, less than a hundred iteration steps were sufficient to reach the convergence. The program to perform such flexible fitting can be obtained as the original source code, NMFF (<https://mmts.org/software/nmff.html>). In a package, NORMA, normal mode based flexible fitting was implemented using an optimization routine (Suhre et al. 2006). Another new package, iMODFit, uses normal mode analysis with internal coordinates and the Monte-Carlo procedure to sample conformations (López-Blanco and Chacón 2013).

In a study describing early adaptation of flexible fitting, NMFF was used to perform flexible fitting of a homology-based atomic model of SecYEG dimer structure into the *E. coli* protein conducting channel (PCC) electron microscopy densities (Mitra et al. 2005). Prior to this study, two possible arrangements of the dimer, namely “front-to-front” and “back-to-back”, were being discussed. A model with higher correlation coefficient was obtained by NMFF when the front-to-front initial structure was used than when back-to-back structure was used, suggesting that a front-to-front arrangement of two SecYEG complexes in the PCC is more favorable, and supports channel formation by the opening of two linked SecY halves during polypeptide translocation.

In a more recent study, cryo-EM single particle analysis was used to obtain the structure of a macromolecular complex of transcription factor IID with IIA and core promoter DNA. The map was at sub-nanometer resolution and multiple levels of computational modeling were performed to construct atomic models. The system consists of a large number of subunits, and available crystal structures as well as homology models were first fitted into the map as rigid bodies. When the rigid body fitting was found to be poor, indicating some conformational changes, iMODfit was used for flexible fitting (Louder et al. 2016).

## 13.7 Flexible Fitting with Molecular Dynamics

Although, there are several advantages in the fitting algorithm with elastic network models, it suffers from a limitation that it cannot fully describe nonlinear conformational dynamics and motion such as domain association and dissociation due to its simple energy function. For this, molecular dynamics method with more atomistic detail becomes necessary.

As described above, in MD, conformations of a system are explored following the molecular mechanics force field, which could be all-atom models or coarse-grained models, defined as a potential energy function,  $U_{\text{mol}}$ . The conformations of the molecules are guided towards structures that have better agreement with the

electron density maps using “biasing potentials”. For cryo-EM data analysis, a 3D volume map is used to define such a biasing potential,  $U_{EM}$ , and include in the potential energy as

$$U = U_{mol} + U_{EM}$$

Although, the definition of the biasing potential is different among different algorithms, in general, it has lower values when the fitness of the model to the experimental data is higher (better), so that during the course of an MD simulation, the conformation is naturally guided into conformations that agree better with experiments. To be used within the algorithm of MD simulation, the biasing potential,  $U_{EM}$ , needs to be derivable by the atomic coordinate, which limits the possible functional form. One approach is to use the correlation coefficient described in the previous Section, Eq. 13.1, as:

$$U_{EM} = -kCC$$

Where  $k$  is a *force constant parameter*, which controls the strength of biasing force, making the high correlation translate to lower energy potential. This type of biasing potential has been implemented using Amber (Orzechowski and Tama 2008) and later Gromacs (Whitford et al. 2011).

In another implementation, the biasing potential is defined as a potential field, in which all atoms are pulled toward the regions where electron density is high in the 3D map (Trabuco et al. 2008).

$$U_{EM} = \sum_{n=1}^N w_n V_{EM}(\mathbf{r}_n) \quad (13.2)$$

$$V_{EM} = \begin{cases} \xi \left[ 1 - \frac{\Phi(\mathbf{r}_n) - \Phi_{thr}}{\Phi_{max} - \Phi_{thr}} \right] & \text{if } \Phi(\mathbf{r}_n) \geq \Phi_{thr} \\ \xi & \text{if } \Phi(\mathbf{r}_n) \leq \Phi_{thr} \end{cases}$$

where  $\Phi(\mathbf{r})$  is the potential converted from EM map.  $\Phi_{max}$  is the maximum value in the map, and  $\Phi_{thr}$  is the threshold value to remove background.  $w_n$  is the weighting factor, which is set to the atomic mass.  $\xi$  is the scaling factor that controls the biasing strength. With this biasing potential, the authors also incorporated restrain potentials,  $U_{SS}$ , to conserve the secondary structure of the molecules. This appears to be a requirement to prevent over-fitting in such a potential based approach, although it is not the case for the correlation based biasing potential. This approach is implemented in NAMD (Phillips et al. 2005). In both the approaches, the potential gradient can be calculated analytically.

Choice of the force constant, i.e., biasing strength, is not straightforward, and is system dependent. It needs to be sufficiently large to guide the conformation to well-fitted models with sufficient efficiency. On the other hand, a too strong bias

forcefully deforms the structures and leads to the models with unrealistic distortions. Such issues are especially serious for the fitting against the experimental data. In the experimental data, due to noise and unavoidable errors in reconstruction, the final electron density map is not exactly the electron density that is expected from atomic positions of single structure. In other words, correlation coefficient can never be 1 (maximum) against experimental data, and furthermore, there may be conformations that can have high correlation to the experimental data, but which are structurally unreasonable. In addition, the fitting procedures with MD are not deterministic. Except for the fittings with very simple conformational transitions, the resulting model can vary from one fitting run to the next, and multiple fitting trials need to be performed. It has been shown that “consensus” fitting can increase reliability of the fitting, i.e., flexible fittings are performed using many different types of fitting approaches and agreement (consensus) between the resulting models can be used as an indicator for the reliability of the models (Ahmed and Tama 2013). In an approach using replica exchange algorithms, different combinations of adjustable force constant parameters are used to run multiple (replica) simulations to increase the reliability of the fitting procedure (Fig. 13.1) (Miyashita et al. 2017).

Another issue for MD based flexible fitting is sampling efficiency. This is more critical for recent higher resolution EM maps, because high-resolution maps create a rugged energy surface for the fitting processes. A variety of approaches/algorithms have been developed to increase the sampling efficiency of MD simulations, and these can also be employed in flexible fitting. Temperature accelerated MD has been shown to increase the speed of fitting (Vashisth et al. 2012). To overcome the issues of conformational search for high-resolution maps, map resolution can be adjusted during MD simulation (Singharoy et al. 2016). Langevin dynamics method that guides the system toward the fitted model was also proposed (Wu et al. 2013).

MD based flexible fitting approaches have been applied to a large number of systems. Among those, the ribosome complex is a particularly important and challenging system to study, with a complex structure that undergoes large conformational transitions. Atomic structural models of the *E. coli* ribosome were constructed with MDFF using EM maps of two functional states at 9 and 6.7 Å resolutions (Trabuco et al. 2008). Dynamics was simulated using the CHARMM27 force field and the potential field, Eq. 13.2, was used for flexible fitting. To construct a ternary complex with several ligands, multiple steps of rigid-body fitting and flexible fittings were performed. MDfit was also used to study the transfer RNA movement through the ribosome and a model of the head-swivel transition was constructed (Whitford et al. 2011). In this study, an all-atom structure based model was employed to simulate the dynamics of the ribosome complex (Whitford et al. 2009). Models for intermediate states were proposed using available X-ray structure and cryo-EM maps at  $\sim 7$  Å resolution. Recent reviews cover other applications of MD based flexible fitting (McGreevy et al. 2016; Xu et al. 2015; Kim and Sanbonmatsu 2017).

## 13.8 Dynamics Extraction from 2D Image Set

One aim of hybrid approaches in structural biology is to obtain information beyond structure, on dynamics. In this regard, in many approaches, obtained structural information represents the “averaged” structure of an ensemble and not those of single particles within. The raw data from electron microscopy are 2D images capturing snapshots of single particles. During the 3D reconstruction, the data are averaged and assembled into 3D model. During this procedure, information regarding variations of the conformations may be lost. New approaches can identify “classes” of images each representing a different conformation. This requires that the number of classes to be set before the analysis and then the conformational states are divided into distinctive conformations. However, there is a risk of introduction of artifact in this procedure. Intuitively, conformational transitions are continuous process and not jumps between distinct conformations. It is not obvious what happens to the images that represent such intermediate structures between the defined conformational states in such classification procedures. Therefore, if 2D images are directly analyzed, additional information on conformational dynamics could be obtained.

New fitting approaches have been explored to analyze 2D images directly for obtaining information regarding conformational ensemble represented in the sample. In one such approach, 3D density map is first generated using conventional technique, i.e., using all data to construct one (average) conformation. Then possible conformational variations are predicted from the 3D map using above mentioned elastic network model with normal mode analysis. The structure can then be optimized to fit to a 2D image. For each image in the dataset, such analysis is performed to estimate a possible conformation that each 2D image may represents. The results represent the ensemble of conformations represented in the dataset, from which conformational dynamics of the system can be studied (Jin et al. 2014).

In another study, 2D images were analyzed using a dimensionality reduction technique to obtain the information on how 2D image set represent conformational ensemble space (Dashti et al. 2014). This approach relies on a simple concept that if two conformations are similar, the projection images of these should also be similar. Thus, by measuring similarities between 2D images and identifying the connectivity of similar image pairs, 2D images can be mapped on to a *manifold* (a multidimensional surface defined by reaction coordinates). This technique was used to analyze data from a ribosome sample that contained a mixture of conformational states in order to obtain conformational ensemble free energy surface instead of a few discrete conformations.

### 13.9 Flexible Fitting with SAXS Data

Flexible fitting is applicable to various experimental data. Particularly, information from small-angle X-ray scattering (SAXS) has similarity to the ones from EM experiments. The data from SAXS provides information about the atomic positions, but not their individual coordinates. In solution, target molecules are in random orientation and the resulting scattering data is spherically averaged. Thus, it cannot provide information on individual coordinates of the atoms, but it provides information about the relative positions of all the pairs of atoms. It is still sufficient to provide information on overall shapes, such as radius of gyration, and approximate shapes could be proposed through computational modeling (Liu et al. 2012; Putnam et al. 2007).

Several approaches for flexible fitting against SAXS experimental data have been developed. These approaches are similar to the flexible fitting to EM density maps in concept; however, a significant challenge of the SAXS data analysis is the low amount of information available for atomic modeling. The SAXS experimentals provide a one-dimensional curve as a function of  $q$  value, which is a function of scattering angle, and the number of independent data points is limited to 10–30 (Hub 2017; Rambo and Tainer 2013). This is significantly less than the information in 3D EM maps, and poses challenges to flexible fitting, leading to the over interpretation of the data. Yet, SAXS enables studies on the structure and dynamics of biological molecules in solution near native state and, by combining SAXS profiles with flexible fitting, it could provide a wealth of information (see a previous chapter).

To avoid over-fitting, one can limit the allowed flexibility during conformational modeling. Normal mode analysis would be an ideal approach in this aspect, since only a small number of modes need to be considered for structure optimization (Gorba et al. 2008). However, it cannot describe complex conformational changes, and for this MD based approaches are required. Significant flexibility of the molecules could lead to over-fitting problems, where a large number of conformations could equally fit the experimental data. In an approach, domains were treated as rigid bodies and only domain connections were allowed to move (Pelikan et al. 2009). In many approaches, MD is used to generate a large number of conformations and each snapshot is compared against the data. In some studies, a few snapshots are identified and proposed as the models to explain the experimental data. However, SAXS data reflects solution ensemble in principle, and thus ensembles of conformations are often discussed to annotate such data (Tria et al. 2015). Recently, an approach to perform biasing MD simulations using SAXS profile has been proposed for more efficient sampling (Chen and Hub 2015).

Assessment of the agreement between a model and SAXS profile is not a simple task. SAXS profile contains information about the distances between all the pairs of the atoms in the sample, which include solute as well as the solvent atoms. The difference from the SAXS profile of pure solvent (contrast) is used for analysis. Here, the distribution of solvent molecules that are bound to solute molecule is different from that of bulk solvent. This needs to be considered for the simulation of

SAXS profile from the molecule. Such issue would also exist in the flexible fitting to EM maps, but SAXS is often used to study smaller molecules and these correction becomes more important. Many algorithms use some models to implicitly simulate such effects from the structure of the solute. Recently, MD simulations with explicit solvent is also used to simulate the distributions of the bound water molecules, which are then used to calculate theoretical SAXS profiles (for detailed discussions, please see (Hub 2017)).

Here we mention a couple of recent studies that combined MD and SAXS data as examples of such approaches. In a study by Anami et al., solution conformations of Vitamin D receptor ligand-binding domain were proposed using a hybrid method combining SAXS and MD (Anami et al. 2016). Experimental SAXS profiles of the apo and antagonist bound states were not consistent with the profile simulated from a crystal structure. Therefore, they performed a series of MD simulations from which SAXS profiles were calculated from the snapshots using CRY SOL (Svergun et al. 1995). Then the conformations that were consistent with the experimental SAXS profiles were identified as the models of apo and antagonist bound solution structures.

In another study, Holdbrook et al. revealed the molecular mechanisms of Skp chaperone using microsecond time-scale MD with SAXS and NMR data (Holdbrook et al. 2017). Skp chaperone can adapt to differently sized clients, but its molecular mechanisms were not known. The X-ray crystal structure was again not consistent with SAXS data. MD simulation revealed significant flexibility in the molecule and thus the SAXS profiles calculated from individual conformations in the trajectories showed variation and no single structure could describe the experimental SAXS data. Good agreement with the experimental data was obtained by considering an ensemble of conformations that include “extreme open” and “extreme close” conformations, which lead to the conclusion that the remarkably flexible conformations allow the Skp chaperone to accommodate client molecules of different sizes.

## 13.10 Summary and Conclusions

In this chapter, we discussed *flexible fitting* approaches based on MD simulations. Structural information is critical for revealing the molecular mechanism of functions. However, for many large and flexible macromolecules, X-ray crystallography is quite challenging and other experimental techniques are employed. Experimental data from such techniques are often low in resolution, unable to provide atomic models. Flexible fitting techniques are particularly useful when a detailed structure, from crystallography or homology modeling, exists, but other experimental data suggests functionally relevant alternative conformations. MD simulation can be used to reveal the intrinsic dynamics of the molecules and to find the conformational transitions that can elucidate observations from low-resolution experimental data. We reviewed the details of molecular mechanics simulations at different

coarse-grained levels and the algorithms to explore the conformations that are consistent with experimental data, particularly from EM and SAXS. Importance of hybrid/integrative approaches for structural biology will continue in order to increase to study more complex macromolecular functions. Thus, computational algorithms that incorporate efficient sampling and multiple experimental data as well as the protocols for reliability and quality assessment need to be further developed.

**Acknowledgements** We thank Sandhya P. Tiwari and Ashutosh Srivastava for carefully reading the manuscript and providing comments. This work was supported by FOCUS for Establishing Supercomputing Center of Excellence, JSPS KAKENHI Grant Number 17K07305, 16K07286, 26119006, 15K21711 and RIKEN Dynamic Structural Biology Project.

## References

- Ahmed A, Tama F (2013) Consensus among multiple approaches as a reliability measure for flexible fitting into cryo-EM data. *J Struct Biol* 182:67–77
- Anami Y, Shimizu N, Ekimoto T, Egawa D, Itoh T, Ikeguchi M, Yamamoto K (2016) Apo- and antagonist-binding structures of vitamin D receptor ligand-binding domain revealed by hybrid approach combining small-angle X-ray scattering and molecular dynamics. *J Med Chem* 59:7888–7900
- Barty A (2016) Single molecule imaging using X-ray free electron lasers. *Curr Opin Struct Biol* 40:186–194
- 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* 26:1668–1688
- Chen PC, Hub JS (2015) Interpretation of solution x-ray scattering by explicit-solvent molecular dynamics. *Biophys J* 108:2573–2584
- Dashti A, Schwander P, Langlois R, Fung R, Li W, Hosseinizadeh A, Liao HY, Pallesen J, Sharma G, Stupina VA, Simon AE, Dinman JD, Frank J, Ourmazd A (2014) Trajectories of the ribosome as a Brownian nanomachine. *Proc Natl Acad Sci U S A* 111:17492–17497
- Frank J (2017) Advances in the field of single-particle cryo-electron microscopy over the last decade. *Nat Protoc* 12:209–212
- Fritz BG, Roberts SA, Ahmed A, Breci L, Li W, Weichsel A, Brailey JL, Wysocki VH, Tama F, Montfort WR (2013) Molecular model of a soluble guanylyl cyclase fragment determined by small-angle X-ray scattering and chemical cross-linking. *Biochemistry* 52:1568–1582
- Gallagher-Jones M, Rodriguez JA, Miao J (2016) Frontier methods in coherent X-ray diffraction for high-resolution structure determination. *Q Rev Biophys* 49
- Garman EF (2014) Developments in x-ray crystallographic structure determination of biological macromolecules. *Science* 343:1102–1108
- Garba C, Miyashita O, Tama F (2008) Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data. *Biophys J* 94:1589–1599
- Holdbrook DA, Burmann BM, Huber RG, Petoukhov MV, Svergun DI, Hiller S, Bond PJ (2017) A spring-loaded mechanism governs the clamp-like dynamics of the Skp chaperone. *structure* 25:1079–1088.e3
- Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmüller H, MacKerell AD (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* 14:71–73

- Hub JS (2017) Interpreting solution X-ray scattering data using molecular simulations. *Curr Opin Struct Biol* 49:18–26
- Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph Model* 14:33–38
- Jin Q, Sorzano COS, de la Rosa-Trevín JM, Bilbao-Castro JR, Núñez-Ramírez R, Llorca O, Tama F, Jonić S (2014) Iterative elastic 3D-to-2D alignment method using normal modes for studying structural dynamics of large macromolecular complexes. *Structure* 22:496–506
- Kikhney AG, Svergun DI (2015) A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* 589:2570–2577
- Kim DN, Sanbonmatsu KY (2017) Tools for the cryo-EM gold rush: going from the cryo-EM map to the atomistic model. *Biosci Rep* 37
- Lander GC, Saibil HR, Nogales E (2012) Go hybrid: EM, crystallography, and beyond. *Curr Opin Struct Biol* 22:627–635
- Liu H, Hexemer A, Zwart PH (2012) The Small Angle Scattering ToolBox(SASTBX): an open-source software for biomolecular small-angle scattering. *J Appl Crystallogr* 45:587–593
- Lopéz-Blanco JR, Chacón P (2013) iMODFIT: efficient and robust flexible fitting based on vibrational analysis in internal coordinates. *J Struct Biol* 184:261–270
- Louder RK, He Y, López-Blanco JR, Fang J, Chacón P, Nogales E (2016) Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature* 531:604–609
- Mahajan S, Sanejouand YH (2015) On the relationship between low-frequency normal modes and the large-scale conformational changes of proteins. *Arch Biochem Biophys* 567:59–65
- McGreavy R, Teo I, Singharoy A, Schulten K (2016) Advances in the molecular dynamics flexible fitting method for cryo-EM modeling. *Methods* 100:50–60
- Merzel F, Smith JC (2002) SASSIM: a method for calculating small-angle X-ray and neutron scattering and the associated molecular envelope from explicit-atom models of solvated proteins. *Acta Crystallogr D Biol Crystallogr* 58:242–249
- Miao J, Ishikawa T, Robinson IK, Murnane MM (2015) Beyond crystallography: diffractive imaging using coherent x-ray light sources. *Science* 348:530–535
- Mitra K, Schaffitzel C, Shaikh T, Tama F, Jenni S, Brooks CL, Ban N, Frank J (2005) Structure of the E. coli protein-conducting channel bound to a translating ribosome. *Nature* 438:318–324
- Miyashita O, Joti Y (2017) X-ray free electron laser single-particle analysis for biological systems. *Curr Opin Struct Biol* 43:163–169
- Miyashita O, Onuchic JN, Wolynes PG (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci U S A* 100:12570–12575
- Miyashita O, Kobayashi C, Mori T, Sugita Y, Tama F (2017) Flexible fitting to cryo-EM density map using ensemble molecular dynamics simulations. *J Comput Chem* 38:1447–1461
- Nguyen HT, Pabit SA, Meisburger SP, Pollack L, Case DA (2014) Accurate small and wide angle x-ray scattering profiles from atomic models of proteins and nucleic acids. *J Chem Phys* 141:22D508
- Oroguchi T, Ikeguchi M (2011) Effects of ionic strength on SAXS data for proteins revealed by molecular dynamics simulations. *J Chem Phys* 134:025102
- Orzechowski M, Tama F (2008) Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys J* 95:5692–5705
- Pelikan M, Hura G, Hammel M (2009) Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen Physiol Biophys* 28:174–189
- Perilla JR, Goh BC, Cassidy CK, Liu B, Bernardi RC, Rudack T, Yu H, Wu Z, Schulten K (2015) Molecular dynamics simulations of large macromolecular complexes. *Curr Opin Struct Biol* 31:64–74
- Pettersen 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* 25:1605–1612
- 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* 26:1781–1802



- Putnam CD, Hammel M, Hura GL, Tainer JA (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys* 40:191–285
- Rambo RP, Tainer JA (2013) Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* 496:477–481
- Saunders MG, Voth GA (2013) Coarse-graining methods for computational biology. *Annu Rev Biophys* 42:73–93
- Schröder GF, Brunger AT, Levitt M (2007) Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* 15:1630–1641
- Singharoy A, Teo I, McGreevy R, Stone JE, Zhao J, Schulten K (2016) Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *Elife* 5:e16105
- Suhre K, Sanejouand Y-H (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res* 32:W610–W614
- Suhre K, Navaza J, Sanejouand YH (2006) NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta Crystallogr D Biol Crystallogr* 62:1098–1100
- Svergun D, Barberato C, Koch MHJ (1995) CRY SOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J Appl Crystallogr* 28:768–773
- Takada S, Kanada R, Tan C, Terakawa T, Li W, Kenzaki H (2015) Modeling Structural Dynamics of Biomolecular Complexes by Coarse-Grained Molecular Simulations. *Acc Chem Res* 48:3026–3035
- Tama F, Sanejouand YH (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng* 14:1–6
- Tama F, Gadea FX, Marques O, Sanejouand YH (2000) Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* 41:1–7
- Tama F, Wriggers W, Brooks CL III (2002) Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *J Mol Biol* 321:297–305
- Tama F, Valle M, Frank J, Brooks CL III (2003) Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc Natl Acad Sci USA* 100:9319–9323
- Tama F, Miyashita O, Brooks CL (2004) Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J Mol Biol* 337:985–999
- Tan RK-Z, Devkota B, Harvey SC (2008) YUP.SCX: coaxing atomic models into medium resolution electron density maps. *J Struct Biol* 163:163–174
- Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* 77:1905–1908
- Trabuco LG, Villa E, Mitra K, Frank J, Schulten K (2008) Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16:673–683
- Tria F, Mertens HD, Kachala M, Svergun DI (2015) Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* 2:207–217
- Unverdorben P, Beck F, Śledź P, Schweitzer A, Pfeifer G, Plitzko JM, Baumeister W, Förster F (2014) Deep classification of a large cryo-EM dataset defines the conformational landscape of the 26S proteasome. *Proc Natl Acad Sci U S A* 111:5544–5549
- Vashisth H, Skiniotis G, Brooks CL (2012) Using enhanced sampling and structural restraints to refine atomic structures into low-resolution electron microscopy maps. *Structure* 20:1453–1462
- Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, Onuchic JN (2009) An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins* 75:430–441

- Whitford PC, Ahmed A, Yu Y, Hennelly SP, Tama F, Spahn CMT, Onuchic JN, Sanbonmatsu KY (2011) Excited states of ribosome translocation revealed through integrative molecular modeling. *Proc Natl Acad Sci U S A* 108:18943–18948
- Wu X, Subramaniam S, Case DA, Wu KW, Brooks BR (2013) Targeted conformational search with map-restrained self-guided Langevin dynamics: application to flexible fitting into electron microscopic density maps. *J Struct Biol* 183:429–440
- Xu X, Yan C, Wohlhueter R, Ivanov I (2015) Integrative Modeling of Macromolecular Assemblies from Low to Near-Atomic Resolution. *Comput Struct Biotechnol J* 13:492–503

# Chapter 14

## Rigid-Body Fitting of Atomic Models on 3D Density Maps of Electron Microscopy



**Takeshi Kawabata**

**Abstract** Cryo electron microscopy has revolutionarily evolved for the determination of the 3D structure of macromolecular complexes. The modeling procedures on the 3D density maps of electron microscopy are roughly classified into three categories: fitting, *de novo* modeling and refinement. The registered atomic models from the maps have mostly been *hand-built* and *auto-refined*. Several programs aiming at automatic modeling have also been developed using various kinds of molecular representations. Among these three classes of the modeling procedures, the rigid body fitting is reviewed here, because it is the most basic modeling process applied before the other steps. The fitting problems are classified as the fittings of single subunit or multiple subunits, and the fittings on global or local parts of maps. A higher resolution map enables more local fitting. Various molecular representations have been employed in the fitting programs. A point and digital image models are generally used to represent molecules, but new representations, such as the Gaussian mixture model, have been applied recently.

**Keywords** Electron microscopy · Gaussian mixture model · EM algorithm

### 14.1 Introduction

The structures of very large macromolecular machines are being determined by combining observations from complementary experimental methods, including X-ray crystallography, NMR spectroscopy, 3D electron microscopy, small-angle scattering, cross-linking, and many others. Among them, cryo electron microscopy

---

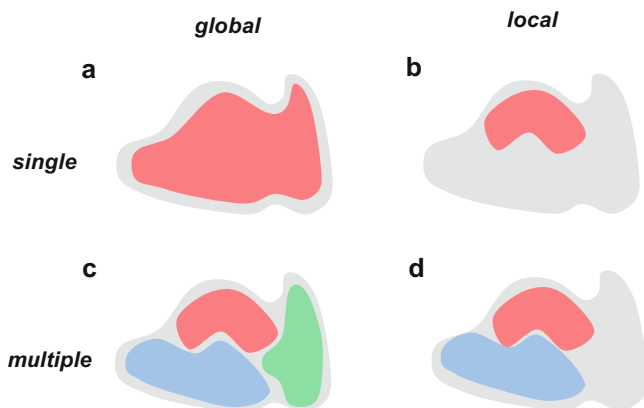
T. Kawabata (✉)

Institute for Protein Research, Osaka University, Suita, Osaka, Japan  
e-mail: [kawabata@protein.osaka-u.ac.jp](mailto:kawabata@protein.osaka-u.ac.jp)

© Springer Nature Singapore Pte Ltd. 2018

H. Nakamura et al. (eds.), *Integrative Structural Biology with Hybrid Methods*,  
Advances in Experimental Medicine and Biology 1105,  
[https://doi.org/10.1007/978-981-13-2200-6\\_14](https://doi.org/10.1007/978-981-13-2200-6_14)

219



**Fig. 14.1** Schematic views of various types of fitting problems. Gray shapes are density maps of the assembly of the subunits. Red, green and blue shapes are subunits. **(a)** A *single* and *global* problem. **(b)** A *single* and *local* problem. **(c)** A *multiple* and *global* problem. **(d)** A *multiple* and *local* problem

has rapidly evolved recently, and its resolution has been remarkably improved (Bai et al. 2015). Low resolution 3D maps ( $>10$  Å) require other structural information to build 3D atomic models, such as atomic models of subunits determined by other methods, while the high resolution maps (better than about 3.5 Å) can enable us to build an *de novo* atomic model, at least partially (Dimaio and Chiu 2016).

The modeling procedures on an EM density map are roughly classified into three categories. (1) *Fitting*: fit the atomic model of the subunit obtained from other experimental methods (X-ray or NMR) or computational prediction methods (homology modeling or *ab initio* modeling). The fitting is further classified as either rigid-body fitting or flexible fitting. (2) *De novo modeling*: model an atomic structure on the given map without using any pre-determined atomic models. (3) *Refinement*: small modifications of the conformation of the atomic model given by the fitting or *de novo* modeling.

Among the three modeling procedures, this review mainly focuses on the rigid-body fitting procedure, because it is the first procedure preceding any other processes. For example, flexible fitting requires an initial model often obtained by rigid-body fitting, *de novo* modeling becomes easy if a reference structure is available and rigidly fitted on the map. This chapter is organized as follows. First, the modeling software is surveyed from the statistics of the EMDB database and the journal *Nature*. Second, the fitting calculations are characterized by the types of -problems (*global -local, single-multiple*). Third, several molecular representations are summarized for the fitting. Finally, the representative fitting programs, including the methods using Gaussian mixture model, are described.

## 14.2 Statistics of Modeling Software

The statistics of the modeling software are described as follows. The EMDB database contains more than five thousand 3D density maps of electron microscopy, and each of the maps has an annotation of the software used for fitting the atomic models into the maps (Lawson et al. 2016). The frequently used software programs are summarized in Table 14.1. The statistics for the four classes of map resolutions are summarized in Table 14.2: “high” (less than or equal to 3.5 Å), “medium-high” (from 3.5 Å to 5.0 Å), “medium” (from 5.0 Å to 10 Å), and “low” (more than 10 Å). The fitting software annotations often include refinement programs, such as PHENIX and REFMAC. Note that the annotations have not been mandatory, and thus thousands of maps with atomic models lack descriptions of the fitting (“N.A.” in Table 14.2). Furthermore, all of the entries deposited in 2017 lack descriptions of the fitting software, because the EMDB may have decided not to include them. To compensate for the absence of software information for the latest EMDB entries, all of the cryo electron microscopy article published in Nature in 2017 were inspected (Table 14.3). More than half of the articles are classified as “medium high” resolution (from 3.5 Å to 5.0 Å), because biologically important complexes published in Nature are often unstable and too flexible for high resolution. Most of the *de novo* modeled structures consist of trans-membrane helices.

These statistics provide us useful information about the trends of the modeling methods. These tables can be summarized as follows. For maps with low and medium resolution ( $>5$  Å), the fitting calculations are primarily done with UCSF Chimera, Situs or COOT, and refined mainly by MDFF. In contrast, for the “high” or “medium-high” resolution maps ( $\leq 5.0$  Å), the *de novo* modeling plays an important role. The program COOT is primarily used for the *de novo* building, and the refinement is mainly accomplished by PHENIX in real space, REFMAC and Rosetta. The UCSF Chimera program is also frequently used for the fitting, even for “high” and “medium high” resolution maps. For high and medium-high resolution EM maps, X-ray crystallography tools, such as COOT, O, RefMac, Phenix and CNS, have been used for modeling and refinement.

Considering the fact that UCSF Chimera and Coot are interactive graphical software supporting manual fitting and modeling, the main trend for modeling on the cryo-EM map is *hand-built* and *auto-refined*. For lower resolution, manual fitting is performed with UCSF Chimera, and refined by MDFF. For higher resolution, manual *de novo* modeling is accomplished with Coot, and refined mainly by PHENIX.

However, more automatic tools are necessary for objective and efficient modeling. Especially, hybrid modeling with various experimental techniques often requires automatic and objective modeling. The previous summaries in Tables 14.1, 14.2 and 14.3 are only for the models submitted to the PDB, based on the

**Table 14.1** Frequently used fitting software for the EMDB database. The entries deposited up through 2016 are considered

| Software     | Number of entries | Comments   | References                |
|--------------|-------------------|--|---------------------------|
| UCSF CHIMERA | 710               | The molecular graphics program with various functions to manipulate 3D density map and atomic models.  | Pettersen et al. (2004)   |
| SITUS        | 116               | The program package with various fitting programs. The program “colores” is the most popular.  | Wriggers (2012)           |
| MDFE         | 106               | Molecular dynamics flexible fitting based on the NAMD program  | Trabuco et al. (2008)     |
| COOT         | 74                | The program to display and manipulate atomic models of macromolecules  | Emsley et al. (2010)      |
| URO          | 52                | The fitting program in the reciprocal space.   | Siebert and Navaza (2009) |
| EMfit        | 44                | The fitting program in the real space written in FORTRAN   | Rossmann et al. (2001)    |
| Flex-EM      | 34                | Flexible fitting program based on domain-based rigid-body fitting.   | Topf et al. (2008)        |
| O            | 33                | The program to visualize and build 3D atomic model on the density map  | Jones (2004)              |
| PHENIX       | 24                | The software suite for the determination of molecular structures using X-ray crystallography and other methods. The real space refinement program is often used  | Adams et al. (2010)       |
| REFMAC       | 18                | Crystallographic refinement program, distributed as part of the CCP4 suite.  | Murshudov et al. (2011)   |
| ROSETTA      | 17                | Low-resolution refinement tools are often used for electron microscopy. The software suite for modeling protein structures, using fragment-assembly and other various procedures. The refinement procedure for electron microscopy has been developed. | Nichollas et al. (2012)   |
| CNS          | 17                | The library for structural biology (crystallography and NMR system)  | Brunger (2007)            |

**Table 14.2** Frequently used fitting software for the EMDB database, summarized for three different resolution ranges. The entries deposited up through 2016 are considered

| High<br>(reso $\leq$ 3.5 Å) |     | Medium high<br>(3.5 Å < reso $\leq$ 5 Å) |     | Medium<br>(5 Å < reso $\leq$ 10 Å) |     | Low<br>(10 Å < reso) |     |
|-----------------------------|-----|--|-----|------------------------------------|-----|----------------------|-----|
| N.A. <sup>a</sup>           | 163 | N.A. <sup>a</sup>                        | 323 | N.A. <sup>a</sup>                  | 248 | UCSF CHIMERA         | 383 |
| UCSF CHIMERA                | 16  | UCSF CHIMERA                             | 78  | UCSF CHIMERA                       | 233 | N.A. <sup>a</sup>    | 148 |
| COOT                        | 15  | COOT                                     | 34  | MDFP                               | 59  | SITUS                | 83  |
| ROSETTA                     | 6   | REFMAC                                   | 14  | SITUS                              | 30  | URO                  | 36  |
| PHENIX                      | 6   | MDFP                                     | 11  | FLEX-EM                            | 24  | MDFP                 | 35  |
| EMFIT                       | 2   | PHENIX                                   | 10  | COOT                               | 23  | EMFIT                | 32  |
| URO                         | 1   | ROSETTA                                  | 6   | URO                                | 15  | O                    | 23  |
| O                           | 1   | SPDBV                                    | 6   | DIREX                              | 11  | GAP                  | 11  |
| MDFP                        | 1   | CNS                                      | 3   | CNS                                | 9   | FLEX-EM              | 10  |
| REFMAC                      | 1   | EMFIT                                    | 3   | PHENIX                             | 8   | MOLREP               | 9   |
|                             |     | SITUS                                    | 3   | O                                  | 8   | VEDA                 | 8   |

<sup>a</sup>Number of EMDB entries with fitted PDB ID, but no fitting software is described

**Table 14.3** Frequently used modeling software in 34Cryo-EM articles published in Nature from 2017/01/05 to 2017/12/21

| Rigid-body fitting |    | <i>De novo</i> modeling |    | Refinement          |    |
|--------------------|----|-------------------------|----|---------------------|----|
| UCSF CHIMERA       | 18 | COOT                    | 17 | PHENIX (real_space) | 24 |
| COOT               | 3  | Gorgon                  | 1  | RefMAC              | 10 |
| ROSETTA            | 1  | O                       | 1  | COOT                | 6  |
| SITUS              | 1  | Rosetta                 | 1  | MDFP                | 4  |

The 34articles were classified by the minimum resolution value, as follows; “high”: 11 articles, “medium high”:20 articles, “medium”:2 articles, and “low”:1 articles

EM density map. In contrast, the models generated by a combination of several experimental techniques are not registered in the PDB. 13 of them are registered in PDB-Dev database (Burley et al. 2017) at this time (June 19, 2018). Among the 13, seven models are built on the 3D EM density map of low resolution, with the help of chemical cross-linking data. Six models were constructed by the Integrative Modeling Platform (IMP) program package (Russel et al. 2012), one model was built by the HADDOCK-EM (van Zundert et al. 2015).

The programs for the rigid-body fitting are mainly reviewed in this chapter, because the rigid-body fitting is required for all resolution ranges. Even for high resolution maps, the fitting of subunit X-ray structures or homology models is often useful as an initial model for *de novo* building. Several free programs used for the fitting calculations are reviewed. For excellent overviews of *de novo* modeling and refinement methods, the reviews by DiMaio and Chiu (2016) and Cassidy et al. (2017) are recommended.

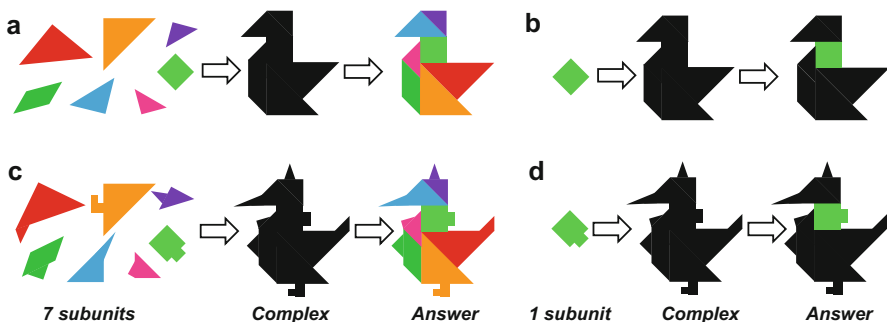
### 14.3 Type of Fitting Problem

The rigid body fitting problem is classified by several points (Fig. 14.1). The first point is the number of subunits to be fitted on the map. Fitting only one subunit on the map is called the *single* subunit fitting problem, whereas the fitting of more than one subunit is called the *multiple* subunit fitting problem. Another point is the locality of the map to be fitted by the subunits. The fitting of one or multiple subunits on the entire region of the given map is called a *global* fitting problem, whereas the fitting on part of the given map, is called a *local* fitting problem.

In view of the computation costs, the *single* fitting problem is much easier than the *multiple* problem. An exhaustive search is often possible for the *single* problem, if the six degrees of freedom are properly discretized. The *single global* problem is also solved by the principal-axes transformations (Pintilie et al. 2010; Suzuki et al. 2016). When the number of subunits becomes large (such as >10), the computation cost for the search increases exponentially.

The locality (*local* or *global*) of the problem often determines the required resolution. Solving the *local* problem often requires a better resolution map than the *global* problem. This situation is easy to understand by using the tangram puzzle, as an example.

The tangram is a tiling puzzle where seven flat pieces can be assembled in different ways to produce a target geometric shape (“silhouette puzzle”). These seven flat pieces are cut from a square, and thus they share edges with the same length and corners with the same angle. Since the 19-th century, many tangram books containing hundreds of problems (“silhouettes”), most of which have familiar shapes, such as birds, animals, people, houses and letters, have been published. An example of the “bird” tangram puzzles is shown in Fig. 14.2a. It is fun and challenging to assemble the seven pieces onto the given target shape. The tangram



**Fig. 14.2** Schematic views of the fitting problem using the tangram and the detailed tangram puzzle. Black shapes are density maps of the assembly of the subunits. Shapes with other colors are subunits. (a) A *multiple* and *global* problem of the tangram puzzle. (b) A *single* and *local* problem of the tangram puzzle. (c) A *multiple* and *global* problem of the detailed tangram puzzle. (d) A *single* and *local* problem of the detailed tangram puzzle



puzzle is similar to the *global* and *multiple* fitting problems. The seven flat pieces correspond to atomic models of subunits, and the target shape corresponds to a density map of the assembly of the subunits. Apparently, all seven pieces are required to solve the tangram puzzle. Putting only one piece on the target is often difficult, because many candidate positions are found with equally good fitness. In other words, solving a *single* and *local* tangram is almost impossible (Fig. 14.2b). This situation is similar to modeling on a low-resolution map.

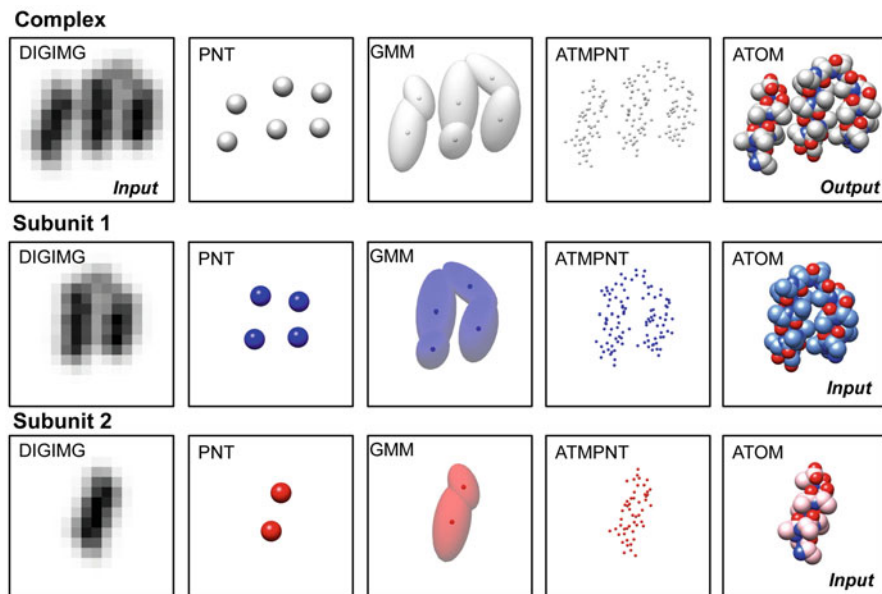
For an analogy of a higher resolution map, we have invented a new “detailed” tangram with more geometric details, as shown in Fig. 14.2c. Each piece has a characteristic additional fragment, and looks more like a piece from a jigsaw puzzles. In contrast to the standard tangram, we can easily determine the position of a piece on the detailed tangram shape, even if only one piece is available (Fig. 14.2d), although the new puzzle is too easy to solve for our entertainment. The original and detailed tangrams correspond to density maps with low and high resolutions, respectively.

The tangram examples shown in Fig. 14.2a, b suggest that all of the subunits are necessary for assembling subunits on a low resolution map. However, in most cases of low resolution maps, some of the subunit atomic structure are not available. To compensate for the missing structures, additional information about the configuration is often needed. That approach is called “hybrid integrative modeling” (Alber et al. 2008). In contrast, a higher resolution map allows us to fit the subunits locally. Correct local fittings lead to the correct multiple fittings; if *single local* fittings are solved correctly for all of the available subunit atomic models, then a *multiple* fitting problem can be also solved simply by assembling the solutions of the *single local* fittings. The *de novo* modeling is regarded as a type of *local* fitting, in which small fragments of secondary structure or polypeptide are fitted into a segmented local region of the map, using the stereo-chemical information.

## 14.4 Molecular Shape Representation

The algorithm of the fitting calculation strongly depends on the representation of the subunit atomic models and the density map. Typical representations of the molecule are shown in Fig. 14.3. A subunit atomic model is often considered as a set of spheres with a van der Waals radius (input ATOM in Fig. 14.3), and a density map of the complex is represented by a 3D digital image (input DIGIMG in Fig. 14.3). To enhance the computation speed, a more coarse-grained representation is often used for the fitting.

A point model (ATMPNT or PNT in Fig. 14.3) is often employed for the fitting due to its simplicity, and it includes a set of 3D points to represent an atomic structure or a density map. Many types of the point models have been proposed, with various levels of coarse-graining and different algorithms. The point model is also called the vector quantization model (Wriggers et al. 1998), beads model (Webb et al. 2018), and pseudo atomic model (Jonić and Sorzano 2016). The finest



**Fig. 14.3** Molecular representations for a density map of complex and atomic models of subunits. DIGIMG: 3D digital image model. Note that they are shown as 2D images for simplicity. PNT 3D points model, GMM Gaussian mixture model, ATMPNT 3D points model for atomic centers, ATOM van der Waals atomic spheres model

representation of this model is the 3D points of atomic centers for the atomic spheres (“ATMPNT” model in Fig. 14.3), ignoring the radii of the atoms. The “Fit-in-map” function in UCSF Chimera employs this representation (Goddard et al. 2007). Lower number of 3D points are also used for more coarse-grained representations. The SITUS package has several programs utilizing the vector quantization method to generate a given number of 3D points (Wriggers et al. 1998). The IMP package employs many types of “beads” models, which are essentially point models. It uses various levels of granularity, such as 1-residue beads or 10-residue beads. The advantage of the point model is that the fitting program can be solved by the discrete problem: matching the 3D points (Wriggers et al. 1998; Zhang et al. 2010; Pandurangan et al. 2015). A gradient-based fitting is also available for the point model. UCSF Chimera uses the gradient of the sum of the densities of the centers of atoms. If the point model is regarded as a set of isotropic Gaussian functions, then it can reproduce an approximated density map, and the correlation coefficient between the given map and subunit point models can be calculated.

The Gaussian mixture model (GMM) is a set of anisotropic Gaussian functions, and thus it represents the original density better than the point model. We will discuss GMM in a separated section.

For representing atomic models, 3D digital image representation is also applied (DIGIMG in Fig. 14.3), such as by the program *colores* in the Situs program

package. It has the advantage of uniform granularity, because both digital images have the same voxel width and resolution. In addition, the fast Fourier transfer (FFT) algorithm enhances the computation speed to calculate the cross correlations with all of the translations.

## 14.5 Tools and Programs for Rigid Body Fitting

This section describes several programs for rigid-body fitting.

### 14.5.1 UCSF Chimera

UCSF Chimera is one of the most popular graphic programs, and also provides both manual and automatic tools to manipulate atomic models and density maps (Pettersen et al. 2004). For the fitting an atomic model into a density map, Chimera provides a well-balanced approach between manual and automatic fittings. The manual fitting is accomplished with the help of the “Model” window. The automatic fitting tool “Fit in map” provides a quick gradient-based local optimization (steepest ascent method) of the atomic model. The atomic point model is employed for a subunit (ATMPNT in Fig. 14.3), whereas the complex is represented by the original digital image (DIGIMG in Fig. 14.3). The program maximizes the sum of densities on the centers of atoms using trilinear interpolation of the given map (Goddard et al. 2007). This tool is well-designed to iterate manual fitting and automatic refinement. One click of the “Fit in map” button moves the subunit by less than its diameter, and rotates it less than 90 degrees. Chimera also has the powerful segmentation tool “Segment Map”, and has a “Fit to Segments” tool, which is useful for fitting a model into part of the given map (Pintilie et al. 2010). The fitting function in UCSF Chimera is designed to solve the *single local* fitting problem. Fitting the multiple subunits can be performed by repeating the “Fit in map” of each single subunit.

The next generation software UCSF ChimeraX is now being developed, and its alpha release is available (Goddard et al. 2018). ChimeraX is designed to visualize CIF files of integrative hybrid modeling (IHM), provided by the PDB-Dev site (Burley et al. 2017).

### 14.5.2 Situs

The Situs program package, which was first released in 1998 (Wriggers 2012), is still widely used for rigid-body fitting against medium and low resolution maps. The source codes of the programs are written in C and C++, and are easy to compile and install in a Unix environment. Although it does not have a graphical

interface, its usage by command lines is simply designed and straightforward to use. Its computation speed is reasonably fast on standard desktop computers.

Among the many programs in the Situs package, the program *colores* is the most popular program for rigid-body fitting (Chacón and Wriggers 2002). It provides single local fitting using an exhaustive lattice-based search enhanced by the FFT and the off-lattice refinement. An atomic model of the subunit is changed to the density map in a digital image representation (DIGIMG in Fig. 14.3), and then two digital images are superimposed with translation to calculate the overlap. The computation of the optimal translation to maximize the overlap is accelerated in reciprocal space by the FFT algorithm, in which  $O(N^6)$  becomes  $O(N^3 \log N^3)$ , where  $N^3$  is the number of grid points. The optimal rotation is exhaustively searched with a given granularity (20~30 degrees). The off-lattice refinement is then performed, using the gradient-based local optimization (Powell's method).

Fitting tools, *quanpdb*, *quanvol* and *matchpt*, which use the vector quantization (VQ) method, are also included in the package. The VQ converts an atomic model or a density map into a set of representative 3D points (PNT model in Fig. 14.3). Fitting of a subunit atomic model into a map can be achieved by discrete point matching.

The Situs package is mainly designed to solve the *single* and *local* fitting problem. If a conformation of multiple subunits is provided by manual inspection or assembling the pose candidates generated by the single-subunit fitting for each subunit, it can be refined by the program *collage*. For the symmetric homo multimer, simply assembling the pose candidates of a single subunit can generate the conformation of the multimer.

### 14.5.3 Integrative Modeling Platform (IMP)

The integrative modeling platform (IMP) is the program package for modeling large macromolecular assemblies by integrating diverse experimental data, from not only electron microscopy, but also chemical crosslinking, FLET and SAXS (Russel et al. 2012). It employs several molecular representations: GMM, digital image, and point models, and various sampling algorithms with many types of spatial restraints. Most of the functions in IMP are provided as the Python module, called Python Modeling Interface (PMI), and the user writes a Python script with the header "import IMP" to use the full functions of the IMP package.

Some of the functions of IMP can be used as command-line tools. One of the command line tools is the program *multifit*, which is for fitting multiple subunits onto a density map (Lasker et al. 2009, 2010). It is designed to solve the *multiple global* problem, requiring the atomic structures for all components. The calculation by the program *multifit* consists of four steps: (1) segmentation of the map into anchor points generated by the Gaussian mixture model, (2) fitting each subunit to the map by an FFT search, (3) preparing a distance restraint file among the subunits, and (4) assembling the subunits by a branch-and-bound algorithm with the DOMINO optimizer.

The Python Modeling Interface (PMI) allows users to employ various types of molecular representations, spatial restraints, and sampling algorithms. The spherical beads model and Gaussian mixture model with varying sizes are available for molecular representation. Distance restraints for chain connectivity and chemical cross links, and restraints for overlap with a density map can be assigned. Monte Carlo and molecular dynamics methods have been implemented as basic sampling algorithms. Combining these algorithms, more advanced sampling procedures, such as simulated annealing, and replica-exchange methods, have been implemented.

In contrast to the *hand-built* and *auto-refined* strategy, the IMP aims at fully automatic modeling, partly because the human intuition cannot work to find the conformations satisfying hundreds of experimental restraints. IMP is also objective; if a PMI python script is available for the model, and the script is executed by the IMP program, the same model will be rebuilt, in principle. However, a rather long Python script (tutorial scripts often have more than 100 lines) and quite a long computation time are required.

#### 14.5.4 Fitting Using a Gaussian Mixture Model

A Gaussian mixture model (GMM) is a probabilistic model that assumes that all of the data points are generated from a mixture of a several number of Gaussian functions. Similar to the VQ (vector quantization) method, GMM has been used for representing the rough shapes of density maps and atomic models with relatively small numbers of parameters; VQ employs a set of 3D points, whereas GMM uses a set of 3D Gaussian functions. A 3D Gaussian mixture model is described as follows:

$$f(\mathbf{r}) = \sum_k^K w_k \cdot \phi_k(\mathbf{r}),$$

where  $\mathbf{r}$  is a 3D vector,  $\phi_k(\mathbf{r})$  is the  $k$ -th Gaussian function, and  $w_k$  is the weight for the  $k$ -th function. The Gaussian function  $\phi_k(\mathbf{r})$  is defined as follows:

$$\phi_k(\mathbf{r}) = \frac{1}{(2\pi)^{3/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{r} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{r} - \boldsymbol{\mu}_k)\right),$$

where  $\boldsymbol{\mu}_k$  is a 3D vector of the mean position, and  $\boldsymbol{\Sigma}_k$  is a 3x3 symmetric covariance matrix. The parameters required for a  $K$ -component GMM are  $K$  sets of  $(w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ . Several groups have used the Gaussian mixture models. The program *gmfit* employs GMM for multiple rigid-body fitting (Kawabata 2008). IMP also employs GMM as one of the molecular representations. The program *MultiFit* employs GMM to determine the anchor points. The structure of Mediator complex (PDBDEV\_00000003) was modeled using Gaussian mixture model (Robinson et al. 2015).

The GMM has the following properties that make it better than the other representation methods. (1) The parameters of GMM are fitted efficiently to maximize the

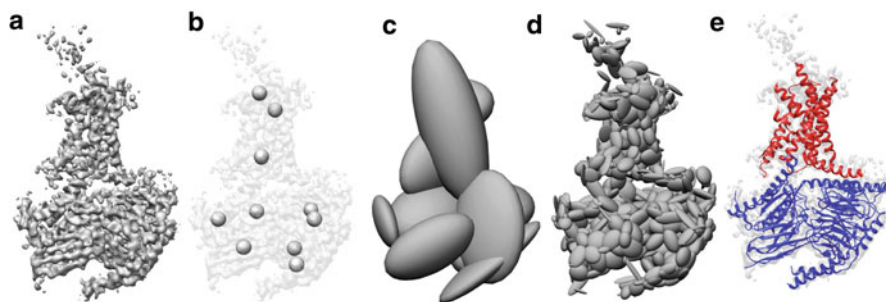
likelihood function by the expectation-maximization (EM) algorithm. (2) Similar density maps can be reproduced from the GMM with a relatively small number of Gaussian functions. Especially, the center of gravity, radius of gyration, and covariance matrix of the GMM are identical to the original map or atomic model, even if the GMM is one Gaussian function ( $K = 1$ ). The VQ method does not have this conservation property. (3) The overlap of two Gaussian functions can be analytically calculated as follows:

$$\int_{-\infty}^{\infty} \phi_A(\mathbf{r}) \phi_B(\mathbf{r}) d\mathbf{r} = \frac{1}{(2\pi)^{3/2} |\boldsymbol{\Sigma}_A + \boldsymbol{\Sigma}_B|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^T (\boldsymbol{\Sigma}_A + \boldsymbol{\Sigma}_B)^{-1} (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)\right)$$

This equation enhances the computation speed for the fitting calculation, because the overlap function has to be evaluated many times during a search for the optimal fitting position.

The fitting program *gmfit* uses the GMM for representing both maps and atomic models (Kawabata 2008) for multiple subunit fitting. The C source codes of *gmfit* and its accompanying program *gmconvert* are freely available (<http://pdj.org/gmfit>); the program *gmconvert* makes a GMM from a map or model. The service “*pairwise gmfit*” is also available through the Web, which quickly fit given two maps or models. This service can be accessed from a searching result page of the Omokage shape search (Suzuki et al. 2016; Kinjo et al. 2017; Kinjo et al. 2018).

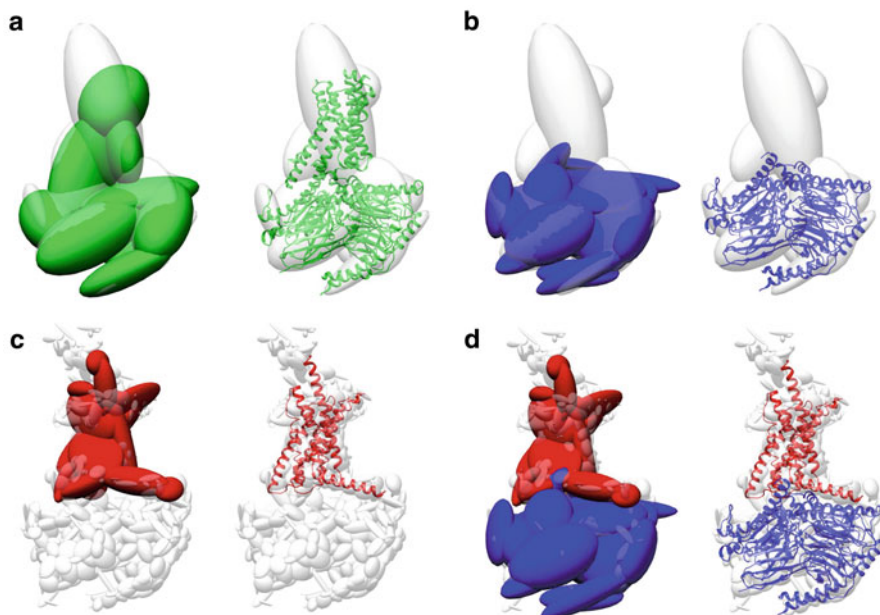
The GMM-based fitting is now being enhanced from several point of views. First, the EM algorithm implemented in *gmconvert* has been improved to consider the sizes of voxels and atoms. The standard EM algorithm only accepts points without their size as the input. However, the voxels and atoms are not actually points, as they have their own grid widths or atomic radii. We invented a new EM algorithm, called the Gaussian-input Gaussian mixture model, which accepts small Gaussian distribution functions that have identical radii of gyration to those of voxels or atoms (Kawabata 2018). Second, a down-sampled Gaussian mixture model is developed, by merging several neighboring voxels into one anisotropic Gaussian function (Kawabata 2018). This model is good for GMM with a large number of Gaussian functions, with small computation costs. Third, a new algorithm for multiple subunit fitting, so-called “segmentation-fitting” method, has been developed. It aims to efficiently cover a density map by given subunits, by repeating the “segmentation” and “fitting” procedures. This algorithm is now extended for fitting the subunits to a part of the density map, by introducing a mask region around each subunit. Finally, we have developed a helix detection program using GMM. The standard GMM does not have any restriction for its components of Gaussian function, and thus functions with any shapes and any sizes can be produced as components of GMM. We invented a new EM algorithm with the components of GMM restricted among a predefined library of Gaussian functions. We call this algorithm “library-GMM”. Library GMM works for detecting candidate regions for  $\alpha$ -helices, if the Gaussian functions in the library correspond to poly-Ala  $\alpha$ -helices.



**Fig. 14.4** Molecular representations for the 3D density map of a class B GPCR - G-protein complex (EMD-8623;  $200^3$  voxels; 4.1 Å; Liang et al. 2017). (a) Surface model of the density map. The author-recommended cutoff value of 0.05 is employed. (b) 3D point model generated by the vector quantization program *quanvol* in the SITUS package using 10 points (code book vectors). (c) Gaussian mixture model generated with the program *gmconvert* by the EM algorithm using 10 Gaussian functions. (d) Down-sampled Gaussian mixture model generated with the program *gmconvert* by merging  $8^3$  voxels into one Gaussian function. The number of Gaussian functions of the down-sampled GMM is 546. (e) The atomic model built by the authors (PDBcode:5uz7). The map consists of five protein chains (A:G $\alpha$ : GNAS2\_HUMAN, B:G $\beta$ : GBB1\_HUMAN, G:G $\gamma$ : GBG2\_HUMAN, N:nanobody 51, R:Calcitonin receptor: CALCR\_HUMAN). The chains A, B, G, N are colored by blue, and the chain R is colored red

Several representation of the map of the GPCR-G-protein complex(EMD-8623; Liang et al. 2017) are summarized in Fig. 14.4. The resolution of the map is medium-high (4.1 Å). The Gaussian mixture model with 10 Gaussian distribution function, derived by the EM algorithm for the Gaussian-input Gaussian mixture model, is shown in Fig. 14.4c. The down-sampled Gaussian mixture model, generated by merging  $8^3$  voxels into one Gaussian distribution function, is displayed in Fig. 14.4d. The number of Gaussian functions of the down-sampled GMM is 546. The atomic model built by the original authors (PDBcode:5uz7) is shown in Fig. 14.3e. The model in the soluble region was built by fitting the X-ray structure of G-protein complex (PDBcode:3sn6). The trans-membrane region was built by fitting the homology model based on the template X-ray structure of class A GPCR (PDBcode: 4l6r). The model was finally refined by the Phenix in the real space.

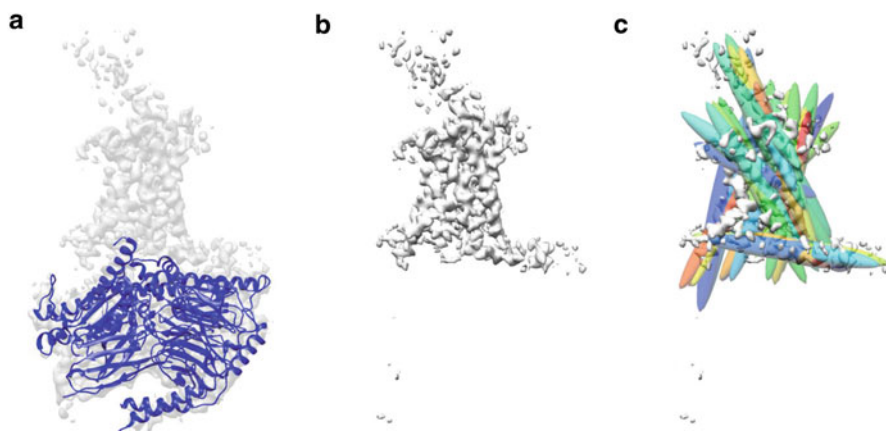
Various types of rigid-body fitting using GMM are shown in Fig. 14.5. A single global fitting is shown in Fig. 14.5a. The simple principal-component axis-based fitting with a small number of Gaussian functions is good enough for the single global fitting, if the two shapes are similar. Examples of single local fitting are displayed in Fig. 14.5b, c. For fitting the G-protein complex (Fig. 14.4b), GMM with 10 functions are sufficient, however, fitting the GPCR protein (Fig. 14.5c) requires more detailed GMM, if we regard the original authors' model (PDBcode:5uz7; Fig. 14.3e) as the correct standard. Generally speaking, fitting smaller subunits requires more detailed resolution of the map. An example of the multiple global fitting is shown in Fig. 14.4d. The detailed GMM also requires to the GPCR subunit to fit correctly.



**Fig. 14.5** Rigid body fitting of X-ray atomic models into the 3D density map of class B GPCR - G-protein complex (EMD-8623; Liang et al. 2017) calculated by the program *gmfit*. Several homologous X-ray structures were fitted on the map, without using the authors' model. Both GMMs (left) and corresponding atomic models (right) are shown. **(a)** Single global fitting. The atomic model of class A GPCR - G-protein complex (PDBcode:3sn6, chains A, B, G, N, R) is fitted into the GMM using 10 Gaussian functions (Fig. 14.4c). The atomic model consists of five chains; A:G $\alpha$ : GNAS2\_BOVIN, B:G $\beta$ : GBB1\_RAT, G:G $\gamma$ : GBG2\_BOVIN, N:Camelid antibody VHH fragment, R:beta-2 adrenergic receptor:ADRB2\_HUMAN). The residues 1002–1160 in chain R (ENLYS\_BPT4) have been removed. **(b)** Single local fitting of the complex of G $\alpha$ , G $\beta$ , G $\gamma$  and antibody (PDBcode:3sn6, chains A, B, G, N) into the GMM using 10 Gaussian functions (Fig. 14.4c). **(c)** Single local fitting of the class B GPCR (PDBcode:4l6r, chain A) into the down-sampled GMM (Fig. 14.4d). The residues 1001–1106 (C562\_ECOLX) have been removed. **(d)** Multiple global fitting of the two rigid body subunits into the down-sampled GMM (Fig. 14.4d). The first subunit is the complex of G $\alpha$ , G $\beta$ , G $\gamma$  and nanobody (PDBcode:3sn6, chains A, B, G, N), colored blue. The second subunit is the class B GPCR (PDBcode:4l6r, chain A), colored red

*De novo* modeling of trans-membrane helices is shown in Fig. 14.6. The density map of the trans-membrane regions is extracted by the fitted soluble G-protein complex (Fig. 14.6a, b). Then, candidates of trans-membrane helices are generated as a set of Gaussian functions by the library-GMM algorithm, as shown in Fig. 14.6c.





**Fig. 14.6** Detection of trans-membrane (TM) helices in the 3D density map. **(a)** The subunit is the complex of  $G\alpha$ ,  $G\beta$ ,  $G\gamma$  and nanobody (PDBcode:3sn6, chains A, B, G, N) is fitted into the map EMD-8623. **(b)** The region around the fitted subunit is erased. The remaining region is supposed to be the trans-membrane region. **(c)** The TM helix candidates are detected by the library-GMM algorithm. Each TM helix candidate is represented by one Gaussian function

## 14.6 Concluding Remarks

This chapter has surveyed the statistics of the atomic modeling programs for electron microscopy density maps, and mainly reviewed rigid-body fitting programs. The high resolution EM map allows us the *de novo* modeling using tools for X-ray crystallography. However, since biologically important complexes are often unstable and flexible, medium and medium-high resolution maps will be still common. Until the next revolution in single particle analysis for flexible complexes, specialized modeling tools must be developed for the medium and medium-high resolution map. Efficient and accurate rigid-body fitting programs must also be developed.

**Acknowledgements** This work was partially supported by JSPS KAKENHI Grants-in-Aid for Scientific Research (C), Grant Number JP26440078 and 17K07364, and the Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from Japan Agency for Medical Research and Development (AMED).

## References

- Adams PD, Afonine PV, Bunkóczy G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 266: 213–221
- Alber F, Förster F, Korkin D, Topf M, Sali A (2008) Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem* 77:443–477
- Bai XC, McMullan G, Scheres SH (2015) How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* 40:49–57
- Brunger AT (2007) Version 1.2 of the crystallography and NMR system. *Nat Protoc* 2:2728–2733
- Burley SK, Kurisu G, Markley JL, Nakamura H, Velankar S, Berman HM, Sali A, Schwede T, Trewhella J (2017) PDB-Dev: a prototype system for depositing integrative/hybrid structural models. *Structure* 25:1317–1318
- Cassidy CK, Himes BA, Luthey-Schulten Z, Zhang P (2017) CryoEM-based hybrid modeling approaches for structure determination. *Curr Opin Microbiol* 43:14–23
- Chacón P, Wriggers W (2002) Multi-resolution contour-based fitting of macromolecular structures. *J Mol Biol* 317:375–384
- DiMaio F, Chiu W (2016) Tools for model building and optimization into near-atomic resolution electron cryo-microscopy density maps. *Methods Enzymol* 579:255–276
- DiMaio F, Song Y, Li X, Brunner MJ, Xu C, Conticello V, Egelman E, Marlovits T, Cheng Y, Baker D (2015) Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat Methods* 12:361–365
- Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66:486–501
- Goddard TD, Huang CC, Ferrin TE (2007) Visualizing density maps with UCSF chimera. *J Struct Biol* 157:281–287
- Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE (2018) UCSF ChimeraX: meeting modern challenges in visualization and analysis. *Protein Sci* 27:14–25
- Jones TA (2004) Interactive electron-density map interpretation: from INTER to O. *Acta Crystallogr D Biol Crystallogr* 60:2115–2125
- Jonić S, Sorzano CÓS (2016) Coarse-graining of volumes for modeling of structure and dynamics in electron microscopy: algorithm to automatically control accuracy of approximation. *IEEE J Select Top Sig Process* 10:161–173
- Kawabata T (2008) Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a Gaussian mixture model. *Biophys J* 95:4643–4658
- Kawabata T (2018) Gaussian-input Gaussian mixture model for representing density maps and atomic models. *J Struct Biol* 203:1–16
- Kinjo AR, Bekker GJ, Suzuki H, Tsuchiya Y, Kawabata T, Ikegawa Y, Nakamura H (2017) Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Res* 45(D1):D282–D288
- Kinjo AR, Bekker GJ, Wako H, Endo S, Tsuchiya Y, Sato H, Nishi H, Kinoshita K, Suzuki H, Kawabata T, Yokochi M, Iwata T, Kobayashi N, Fujiwara T, Kurisu G, Nakamura H (2018) New tools and functions in data-out activities at Protein Data Bank Japan (PDBj). *Protein Sci* 27:95–102
- Lasker K, Topf M, Sali A, Wolfson HJ (2009) Inferential optimization for simultaneous fitting of multiple components into a cryoEM map of their assembly. *J Mol Biol* 388:180–194
- Lasker K, Sali A, Wolfson HJ. (2010) Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins* 278:3205–3211
- Lawson CL, Patwardhan A, Baker ML, Hryc C, Garcia ES, Hudson BP, Lagerstedt I, Ludtke SJ, Pintilie G, Sala R, Westbrook JD, Berman HM, Kleywegt GJ, Chiu W (2016) EM Data Bank unified data resource for 3DEM. *Nucleic Acids Res* 44(D1):D396–D403

- Liang YL, Khoshouei M, Radjainia M, Zhang Y, Glukhova A, Tarrasch J, Thal DM, Furness SGB, Christopoulos G, Coudrat T, Danev R, Baumeister W, Miller LJ, Christopoulos A, Kobilka BK, Wootten D, Skiniotis G, Sexton PM (2017) Phase-plate cryo-EM structure of a class B GPCR-G-protein complex. *Nature* 546:118–123
- Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* 67:355–367
- Nicholls RA, Long F, Murshudov GN (2012) Low-resolution refinement tools in REFMAC5. *Acta Crystallogr D Biol Crystallogr* 68:404–417
- Pandurangan AP, Vasishtan D, Alber F, Topf M (2015)  $\gamma$ -TEMPy: simultaneous fitting of components in 3D-EM maps of their assembly using a genetic algorithm. *Structure* 23:2365–2376
- Pettersen 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* 25:1605–1612
- Pintilie GD, Zhang J, Goddard TD, Chiu W, Gossard DC (2010) Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J Struct Biol* 170:427–438
- Robinson PJ, Trnka MJ, Pellarin R, Greenberg CH, Bushnell DA, Davis R, Burlingame AL, Sali A, Kornberg D. (2015) Molecular architecture of the yeast Mediator complex. *Elife* e08719
- Rossmann MG, Bernal R, Pletnev SV (2001) Combining electron microscopic with X ray crystallographic structures. *J Struct Biol* 136:190–200
- Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* 10:e1001244
- Siebert X, Navaza J (2009) UROX 2.0: an interactive tool for fitting atomic models into electron-microscopy reconstructions. *Acta Crystallogr D Biol Crystallogr* 65:651–658
- Suzuki H, Kawabata T, Nakamura H (2016) Omokage search: shape similarity search service for biomolecular structures in both the PDB and EMDb. *Bioinformatics* 32:619–620
- Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A (2008) Protein structure fitting and refinement guided by cryo-EM density. *Structure* 16:295–307
- Trabuco LG, Villa E, Mitra K, Frank J, Schulten K (2008) Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16:673–683
- van Zundert GCP, Melquiond ASJ, Bonvin AMJJ (2015) Integrative modeling of biomolecular complexes: HADDOCKing with Cryo-Electron microscopy data. *Structure* 23:949–960
- Webb B, Viswanath S, Bonomi M, Pellarin R, Greenberg CH, Saltzberg D, Sali A. (2018) Integrative structure modeling with the integrative modeling platform. *Protein Sci* 28:245–258
- Wriggers W (2012) Conventions and workflows for using situs. *Acta Crystallogr D Biol Crystallogr* 68:344–351
- Wriggers W, Milligan RA, Schulten K, McCammon JA (1998) Self-organizing neural networks bridge the biomolecular resolution gap. *J Mol Biol* 284:1247–1254
- Zhang S, Vasishtan D, Xu M, Topf M, Alber F (2010) A fast mathematical programming procedure for simultaneous fitting of assembly components into cryoEM density maps. *Bioinformatics* 26:i261–i268

# Chapter 15

## Hybrid Methods for Modeling Protein Structures Using Molecular Dynamics Simulations and Small-Angle X-Ray Scattering Data



Toru Ekimoto and Mitsunori Ikeguchi

**Abstract** Small-angle X-ray scattering (SAXS) is an efficient experimental tool to measure the overall shape of macromolecular structures in solution. However, due to the low resolution of SAXS data, high-resolution data obtained from X-ray crystallography or NMR and computational methods such as molecular dynamics (MD) simulations are complementary to SAXS data for understanding protein functions based on their structures at atomic resolution. Because MD simulations provide a physicochemically proper structural ensemble for flexible proteins in solution and a precise description of solvent effects, the hybrid analysis of SAXS and MD simulations is a promising method to estimate reasonable solution structures and structural ensembles in solution. Here, we review typical and useful *in silico* methods for modeling three dimensional protein structures, calculating theoretical SAXS profiles, and analyzing ensemble structures consistent with experimental SAXS profiles. We also review two examples of the hybrid analysis, termed MD-SAXS method in which MD simulations are carried out without any knowledge of experimental SAXS data, and the experimental SAXS data are used only to assess the consistency of the solution model from MD simulations with those observed in experiments. One example is an investigation of the intrinsic dynamics of *EcoO109I* using the computational method to obtain a theoretical profile from the trajectory of an MD simulation. The other example is a structural investigation of the vitamin D

---

T. Ekimoto  
Graduate School of Medical Life Science, Yokohama City University, Tsurumiku,  
Yokohama, Japan  
e-mail: [ekimoto@yokohama-cu.ac.jp](mailto:ekimoto@yokohama-cu.ac.jp)

M. Ikeguchi (✉)  
Graduate School of Medical Life Science, Yokohama City University, Tsurumiku,  
Yokohama, Japan

Medical Sciences Innovation Hub Program, RIKEN, Tsurumiku,  
Yokohama, Japan  
e-mail: [ike@yokohama-cu.ac.jp](mailto:ike@yokohama-cu.ac.jp)

receptor ligand-binding domain using snapshots generated by MD simulations and assessment of the snapshots by experimental SAXS data.

**Keywords** Small-angle X-ray scattering · Molecular dynamics simulation · Solution structure · Coarse-grained model · MD-SAXS · Endonuclease · Vitamin D receptor

## 15.1 Introduction for Small-Angle X-Ray Scattering

Small-angle X-ray scattering (SAXS) is an efficient experimental tool for measuring three-dimensional protein structures in solution (Svergun and Koch 2003; Rambo and Tainer 2013; Kikhney and Svergun 2015; Vestergaard 2016; Hammel 2012; Hura et al. 2009; Bernado 2010; Bernado and Svergun 2012). First, X-ray scattering data for both protein-solution samples and pure-buffer samples are obtained. Next, one-dimensional (1-D) scattering intensity is calculated by subtracting the scattering intensity for the buffer sample from that of the solution sample. The resulting 1-D scattering intensity includes information on the overall shape and size of proteins in the solution. Since proteins fluctuate and rotate freely in solution, the scattering intensity is an average quantity of rotational and conformational protein variants.

In contrast to X-ray crystallography, which is widely used for determination of protein atomic structures, SAXS data is limited to low resolution. However, structural information from SAXS data can capture bare structures in solution, and it is free from the effects of crystal packing; this is a major advantage of SAXS. There is no molecular-size limitation in SAXS experiments, and there is also a contrast advantage over nuclear magnetic resonance (NMR) that is widely used to capture solution structures. Since SAXS data can only estimate the overall shape of molecules due to its low resolution, high-resolution data obtained from other tools are necessary for determining atomic structures and dynamics. This means that SAXS is complementary to X-ray crystallography and NMR (Svergun and Koch 2003; Rambo and Tainer 2013; Hammel 2012; Grishaev et al. 2005; Venditti et al. 2016).

Basic analyses of SAXS data to understand solution structures are as follows (Svergun and Koch 2003; Rambo and Tainer 2013; Kikhney and Svergun 2015; Vestergaard 2016; Hammel 2012; Hura et al. 2009; Bernado 2010). From a 1-D scattering intensity, a radius of gyration ( $R_g$ ), molecular weight, maximal dimension ( $D_{max}$ ), excluded particle volume and flexibility can be estimated through the Guinier approximation at the small-angle range, the forward intensity, the pair distance function, the Porod volume and the Kratky plot, respectively. These analyses are used for validating monodispersity and interparticle interference (Jacques et al. 2012). When a high-resolution three-dimensional (3D) crystallography or NMR structure is available, a theoretical SAXS profile for the 3D structure is calculated using CRY SOL (Svergun et al. 1995). This is then compared to the experimental profile. The deviation between the two theoretical and experimental profiles can be structurally examined by superimposing the 3D structure onto a 3D envelope estimated from the experimental profile by *ab initio* methods (e.g., DAMMIN (Svergun

1999) and GASBOR (Svergun et al. 2001)). When the deviation seems to arise from flexible regions and/or relative arrangements of domains, *ab initio* modeling methods (e.g., SASREF (Petoukhov and Svergun 2005), BUNCH (Petoukhov and Svergun 2005), and CORAL (Petoukhov et al. 2012)) provide possible 3D structural models that are consistent with the experimental SAXS profile. A possibility of the effects of mixture on the SAXS profile is simply tested using OLIGOMER (Konarev et al. 2003) with arbitrary conformations.

Although information about solution structures is extracted from the basic analyses described above, there are essential difficulties to investigating solution structures from low-resolution SAXS data. Since the superimposition of the crystal structure onto the envelope estimated from the SAXS experimental data is not topologically unique, this approach makes it difficult in principle to figure out how the conformation is different from the crystal structure. In addition, misunderstandings may occur in the cases of very flexible proteins, proteins undergoing conformational changes and intrinsically disordered proteins/regions (Kikhney and Svergun 2015; Bernado and Svergun 2012; Wright and Dyson 1999), since those proteins do not adopt a specific and rigid conformation. In the case of these flexible proteins, SAXS data represent an average of scattering from various conformations in solution. This means that a set of conformations is necessary to analyze them using ensemble-modeling methods (e.g., EOM (Bernado et al. 2007)). However, the unique determination of a structural ensemble based on only SAXS data is still difficult due to the limited amount of available information. The generation of ensemble structures also requires careful handling, because proteins structurally fluctuate in specific ways based on their structural characteristics. To overcome the difficulty posed by the flexibility of proteins, an incorporation of physicochemical methods, such as molecular dynamics (MD) simulation, into SAXS analysis would allow us to generate physicochemically proper solution structures and structural ensembles. (See excellent reviews for more detailed information: Rambo and Tainer 2013; Kikhney and Svergun 2015; Hammel 2012; Schneidman-Duhovny et al. 2012; Boldon et al. 2015).

MD simulation is now becoming a powerful tool to study protein dynamics with increasing computational power (Dror et al. 2012; Goh et al. 2016; Lane et al. 2013). MD simulations can capture the conformational dynamics of proteins based on their structurally intrinsic characteristics. However, all-atom MD simulations suffer from two major problems. One is the limitation of the timescale, and the other is the accuracy of the force fields. To overcome the limitation of the timescale, many efforts have been made to improve the MD calculations. This includes developing highly parallelized algorithms (e.g., GROMACS (Abraham et al. 2015), AMBER (Case et al. 2017), NAMD (Phillips et al. 2005) and GENESIS (Kobayashi et al. 2017)) and specialized hardware (e.g., MD-GRAPE (Ohmura et al. 2014) and ANTON (Shaw et al. 2008)). In addition to conventional MD simulations, efficient sampling techniques with biased simulations (e.g., replica exchange (Sugita and Okamoto 1999), accelerated MD (Hamelberg et al. 2004), string methods (Weinan and Vanden-Eijnden 2010), and metadynamics (Piana and Laio 2007)) and statistical analysis of unbiased simulations (e.g., weighted ensemble simulation (Zuckerman and Chong 2017) and Markov state model (Harrigan et al. 2017; Scherer et al.

2015)) have been developed. Comparison of the simulated conformational dynamics with those observed in experiments allows us to assess the accuracy of the force fields (e.g., refs (Lindorff-Larsen et al. 2012; Beauchamp et al. 2012)), and improvements on the force field are still ongoing. For example, ensemble structures generated by various force fields have been examined and compared to SAXS and NMR data (Rauscher et al. 2015).

Hybrid analysis of SAXS and MD simulation is a promising method for estimating solution structures and structural ensembles of flexible proteins. Two approaches of the SAXS and MD combination have been proposed so far. The first approach uses the artificial forces to modify the weights of structures during MD simulations so that the generated structural-ensemble is consistent with the experimental SAXS data (e.g., SWAXS-driven MD (Chen and Hub 2015)). In the second approach, MD simulations are carried out without any knowledge of the experimental SAXS data (e.g., MD-SAXS (Oroguchi et al. 2009)). Then, the consistency of the theoretical SAXS profile that was calculated using MD trajectories from the experimental profiles is examined. In this method, the experimental SAXS data are used only to check the validity of MD simulations. Therefore, this approach can avoid the excessive modification of structural ensembles fitted to the experimental profile. These approaches and the typical methods for modeling protein structures with SAXS data analysis are reviewed in the next section.

## 15.2 Overview of Computational Methods for Modeling Protein Structures Using Small-Angle X-Ray Scattering Data

First, a complete 3D-structure of target proteins is necessary as a starting structure. When a crystal structure is available, missing regions for side chains, loops, and tags at N-terminal region should be added using homology modeling to make a complete structure such that the full length of the protein exactly agrees with that used in the SAXS experiment. The theoretical intensity depends on the length of the protein used in the calculation. At the small-angle region,  $R_g$  depends on the total number of atoms in the calculated protein, and the molecular shape created by flexible regions affects the shape of the intensity at the middle- to high-angle regions. The homology modeling (Fiser 2010) can be executed by MODELLER (Sali and Blundell 1993) (implemented in Chimera (Yang et al. 2012)) or web-based tools (e.g., SWISS-MODEL (Kiefer et al. 2009), HHpred (Alva et al. 2016), Robetta (Kim et al. 2004)). Partially unfolded, multi-domain and complex structures can be modeled through the combined use of the template structures. When the relative position of the domains seems to be flexible, possible relative positions are estimated using docking simulations (e.g., ClusPro (Kozakov et al. 2017)) and modeling linker parts. Protein-protein docking simulations are also done with experimental SAXS data (e.g., pyDockSAXS (Pons et al. 2010), FoXSDock (Schneidman-Duhobny et al. 2011)). If no crystal structures are available, a structure can be provided by homology modeling (Fiser 2010), and a template search combined with SAXS data

(e.g., SAXSTER (dos Reis et al. 2011)) is informative. When only the rough shape is necessary, the coarse-grained (CG) model of proteins is a useful choice (Saunders and Voth 2013). For an example of a CG model, a residue is represented as a bead at its  $C\alpha$  position. Due to the coarse-grained representation, detailed discoveries of interactions between residues are impossible. However, efficient samplings of very flexible proteins are possible.

The theoretical SAXS profile of the provided 3D-structure is compared to the experimental profile. The experimental SAXS data include not only the solute itself but also the solvent effects such as the solvent-excluded volume and the hydration water. Because the electron density of hydration water is larger than that of bulk water, the scattering from the hydration water around the solute significantly contributes to the SAXS profile. All methods introduced here take into account such solvent effects and the difference is their treatments: the implicit hydration model or explicit model. The implicit representation of hydration water is used in CRY SOL (Svergun et al. 1995), FoXS (Schneidman-Duhovny et al. 2013), AquaSAXS (Poitevin et al. 2011), Zernike polynomials-based method (Liu et al. 2012), SWAXS with HyPred (Virtanen et al. 2011), and RISM-SAXS (Nguyen et al. 2014). The explicit representation is used in AXES (Grishaev et al. 2010), Park et al. (2009), MD-SAXS (Oroguchi et al. 2009), Hummer et al. (Köfinger and Hummer 2013), WAXSiS (Knight and Hub 2015), and PM-SAXS (Marchi 2016). For use of the coarse-grained model of proteins or protein-DNA/RNA complexes, the pre-calculated model is used in Stovgaard et al. (Stovgaard et al. 2010), and the explicit hydration model with a dummy water molecule is used in Fast-SAXS (Yang et al. 2009) and Fast-SASXS-pro (Ravikumar et al. 2013).

In methods based on the implicit solvent model, a uniform hydration model with adjustable parameters (CRY SOL (Svergun et al. 1995) and FoXS (Schneidman-Duhovny et al. 2013)), a pre-calculated average density (AquaSAXS (Poitevin et al. 2011), Zernike polynomials-based method (Liu et al. 2012), and SWASX with HyPred (Virtanen et al. 2011)) or a theoretically calculated solvent density (RISM-SAXS (Nguyen et al. 2014)) is used for calculations of the excluded-volume term and the hydration shell term in the form factor. For example, in CRY SOL (Svergun et al. 1995), the hydration water is modeled as the hydration shell of proteins, which has a higher electron density than the bulk water region, and the solvent-excluded volume is modeled as a Gaussian sphere with effective radii. However, the estimation of the increment of electron density in the hydration shell and the determination of the effective radius of the Gaussian spheres are difficult because they depend on the nature of the protein surface, the packing of protein interiors and solvent compositions. Therefore, two parameters, i.e., the increment of electron density in the hydration shell and the effective radius of the Gaussian spheres, are adjusted for fitting to experimental SAXS profiles. In FoXS (Schneidman-Duhovny et al. 2013), the formulation of the form factor is like that of CRY SOL (Svergun et al. 1995). However, the fraction of the solvent-accessible surface is introduced in the hydration shell term. In pre-calculated average density models (Poitevin et al. 2011; Liu et al. 2012; Virtanen et al. 2011), the solvent density around solute is numerically calculated by a 3-Dgrid-based approach before its density



map at each grid is used to calculate the form factors. In RISM-SAXS (Nguyen et al. 2014), a thermally averaged distribution of water and ions around a protein is theoretically obtained using the three-dimensional reference interaction model (3D-RISM). According to a comparison between methods (Schneidman-Duhovny et al. 2012), the discrepancy between theoretical and experimental profiles  $\chi$  for CRY SOL (Svergun et al. 1995) and FoXS (Schneidman-Duhovny et al. 2013) is reasonable despite the uniform hydration shell used in CRY SOL (Svergun et al. 1995) and FoXS (Schneidman-Duhovny et al. 2013).

In the methods using the explicit solvent model, explicitly water molecules are placed around a protein with a superimposition (AXES (Grishaev et al. 2010)), or explicit coordinates of water molecules around a protein are generated using an all-atom MD simulation (Park et al. (2009), MD-SAXS (Oroguchi et al. 2009), Hummer et al. (Köfing and Hummer 2013), WAXSiS (Knight and Hub 2015), and PM-SAXS (Marchi 2016)). In AXES (Grishaev et al. 2010), the excluded and surface solvent molecules are determined by superimposition of a protein onto snapshots generated by MD simulations of the bulk system. Compared with CRY SOL (Svergun et al. 1995), AXES (Grishaev et al. 2010) uses explicit configurations of water molecules and shows an improvement in  $\chi$ . However, fluctuation of water molecules around a protein is not considered. In contrast, all-atom MD simulations can incorporate the fluctuation of water around a flexible protein, and the MD-based methods (Park et al. 2009; Oroguchi et al. 2009; Köfing and Hummer 2013; Knight and Hub 2015; Marchi 2016) treat solvent effects at an atomic level. In SAXS experiments, X-ray scattering from the buffer-only solution is measured as well as that of the protein solution before the scattering intensity of the buffer solution is subtracted from those of protein solution. In the MD-based methods (Park et al. 2009; Oroguchi et al. 2009; Köfing and Hummer 2013; Knight and Hub 2015; Marchi 2016), the MD simulation for the pure solvent and the protein solution is performed as an experiment. Then, the theoretical SAXS profile is obtained by subtracting the two theoretical scattering intensities of protein-solution and pure-solvent MD simulations. Thus, the solvent effects on SAXS profiles, i.e., hydration water and the solvent-excluded volume of proteins, are considered at the atomic level. In addition, since the electron density of bulk solvent depends on the ion concentration, ions in bulk significantly affect SAXS profiles. The ion effects are also considered (Oroguchi and Ikeguchi 2011). The computational method for the rotational average of the form factor is the main difference between the MD-based methods (Park et al. 2009; Oroguchi et al. 2009; Köfing and Hummer 2013; Knight and Hub 2015; Marchi 2016), and the modulation of the excluded-volume term is in the method (Köfing and hummer 2013) to improve the accuracy for the WAXS region.

Due to the limitation of a time scale in the use of all-atom MD simulations, the CG representation model (Saunders and Voth 2013) is still useful for very large and/or very flexible proteins. Under CG-MD simulations, the effect of water molecules is implicitly incorporated, and it is necessary to develop a method for incorporating solvent effects on theoretical SAXS profiles in a CG manner. In the method by Stovgaard et al. (2010), the form factor of CG particles is estimated by averaging form factors calculated using CRY SOL (Svergun et al.

1995) based on structural data in the Protein Data Bank (PDB). In contrast, in Fast-SAXS (Yang et al. 2009) and Fast-SAXS-pro (Ravikumar et al. 2013), the form factor of CG particles is estimated using an average of residues in high-resolution structures in PDB, and the contributions of solvent effects are estimated using explicit placements of dummy water molecules around the protein. Here, in the CG-MD based approaches, the ensemble generated by CG-MD simulations should be checked by any experimental results, including SAXS, because the CG representation includes many adjustable parameters about structures and dynamics due to the coarse-grained model.

The implicit or explicit treatment of solvent effects influences both the accuracy and computational cost for theoretical SAXS profiles. This is a trade-off, and the choice of the method depends on the purpose. For example, when the first priority is a structural investigation to discover the structural characteristics consistent with the experimental SAXS profile, the method based on the implicit model (e.g., CRY SOL (Svergun et al. 1995)) is adequate because the fast calculation allows us to calculate a large number of structures. When an accurate SAXS profile is necessary, the method based on the explicit model with MD simulations (e.g., MD-SAXS (Oroguchi et al. 2009)) is adequate because it explicitly treats the dynamics of water molecules. However, the computational cost is high, and the solvent effects of both water molecules and ions are considered.

Because the experimental SAXS data is obtained as an averaged quantity over conformations of proteins, a consideration of the ensemble structure is necessary. Several approaches have been introduced so far (e.g., MD-SAXS (Oroguchi et al. 2009), Lau et al. (Lau and Roux 2007), EOM (Bernado et al. 2007), MES (Pelikan et al. 2009), BSS-SAXS (Yang et al. 2010), and EROS (Rozycki et al. 2011)). Procedures in these methods are as follows. First, the resolution of proteins and solvent molecules is chosen as an all-atom or CG representation. Using MD-based samplings or topologically random generations, a set of conformations is generated. A theoretical profile is calculated as an average of the weighted profiles of each conformation. A major difference among the methods is how to use experimental SAXS profiles. In MD-SAXS (Oroguchi et al. 2009) and Lau et al. (Lau and Roux 2007), the weight in the average process is determined by the force field or free energy landscape, respectively. In MD-SAXS (Oroguchi et al. 2009), a trajectory generated by an all-atom MD simulation is directly used in the calculation of the theoretical profile, and a simple average is taken because the conformations naturally appear in accordance with weights defined in the force field. In Lau et al. (Lau and Roux 2007), a free energy landscape is obtained using all-atom MD simulations and umbrella sampling. The theoretical profile is obtained using an average of the profiles for conformations near the free energy minimum with the Boltzmann distribution of their energies. In both MD-SAXS (Oroguchi et al. 2009) and the method by Lau et al. (Lau and Roux 2007), the experimental SAXS profile is used only for validation. In contrast, other methods use the experimental SAXS profile in the optimization process of the weights of conformations so that the difference between the theoretical and experimental SAXS profiles is minimized. Two methods have been introduced to avoid overfitting. The one way is by reducing

the size of the ensemble using in the average process. Trial for the ensemble selection is done by iterations on the subset selection in EOM (Bernado et al. 2007) and MES (Pelikan et al. 2009). The reduction is also done by clustering processes in terms of structural similarity and SAXS intensity similarity, and the number of structures and their weights are determined by the Bayesian-based Monte Carlo in BSS-SAXS (Yang et al. 2010). The other way is through reweighting using a maximum-entropy method in EROS (Rozycki et al. 2011). The pseudo free energy is defined as  $\chi^2 - \theta S$ , where  $\theta$  is a control parameter and  $S$  is the relative entropy representing the change in total weights from the initial weight. By changing the weights (entropy) at an adequate  $\theta$ , the relative weight of conformations is optimized such that the free energy is minimized. Similar approaches are used in EOM2 (Tria et al. 2015) and the ensemble-fit procedure in AquaSAXS (Poitevin et al. 2011).

When a representative structure consistent with experimental SAXS data is required, the use of SAXS-driven structural-optimizations may be a better choice. For example, the Monte Carlo (MC) based approach (Förster et al. 2008), the normal-mode flexible fitting (Gorba and Tama 2010), the CG elastic network model (Zheng and Tekpinar 2011), SAXS\_MD (Kojima et al. 2004; Morimoto et al. 2013), SWAXS-driven MD (Chen and Hub 2015), and SAXS-guided metadynamics (Kimanius et al. 2015) have been proposed. A common strategy in these methods is to incorporate the experimental SAXS profile into the scoring function or the potential energy as a bias so that the input structure is forced to undergo conformational change toward a structure consistent with the experimental SAXS profile. In the MC-based approach (Förster et al. 2008), candidate structures are modeled via rigid-body simulations. In the methods using CG representations (Gorba and Tama 2010; Zheng and Tekpinar 2011), a protein is treated as a chain of  $C\alpha$  atoms, and the positions of the atoms are moved according to the low frequency normal mode (Gorba and Tama 2010) or the minimum of the elastic network model energy (Zheng and Tekpinar 2011). In contrast, SAXS\_MD (Kojima et al. 2004), SWAXS-driven MD (Chen and Hub 2015), and SAXS-guided metadynamics (Kimanius et al. 2015) use all-atom MD simulations with an additional potential. Owing to the all-atom treatment of proteins and solvent molecules, the resulting structures may be more plausible than those generated by other methods. The solvent effects on the theoretical SAXS profiles are also explicitly treated in SWAXS-driven MD (Chen and Hub 2015). Additional information derived from NMR, e.g., distances, is incorporated in SAXS\_MD (Morimoto et al. 2013). Note that the resulting structure in these methods is obtained using artificial forces, so structural validity should be checked.

### 15.3 Applications of the Hybrid Method of Molecular Dynamics Simulations and Small-Angle X-Ray Scattering

The hybrid methods of MD simulations and SAXS experiments enable us to discuss their functions via 3D solution structures observed in SAXS experiments on a physicochemically rational basis. Such analyses will be helpful for even the

following difficult cases. (i) Solution structures appear to be flexible and are hardly crystallized. (ii) Solution structures appear to adopt a different conformation from the crystal structures. (iii) Proteins undergo conformational changes upon ligand binding. However, structures after the conformational change cannot be determined using crystallography. (iv) Structures in the apo state cannot be determined. (v) Only crystal structures of homologous proteins are determined. (vi) Only parts of the domain structures are determined. However, full-length multi-domain structures cannot be determined.

In the following sections, two applications of the hybrid approach of MD simulations and SAXS are reviewed as an example. In these studies, MD simulations are carried out without any knowledge of experimental SAXS data. Then, the consistency of the theoretical SAXS profile calculated from MD trajectories with the experimental profiles is examined. In this method, the experimental SAXS data are used only to check the validity of MD simulations. This approach is referred to as “the MD-SAXS method”. The MD-SAXS method was applied to endonuclease *EcoO109I* (Oroguchi et al. 2009) and vitamin D receptor ligand-binding domain (Anami et al. 2016).

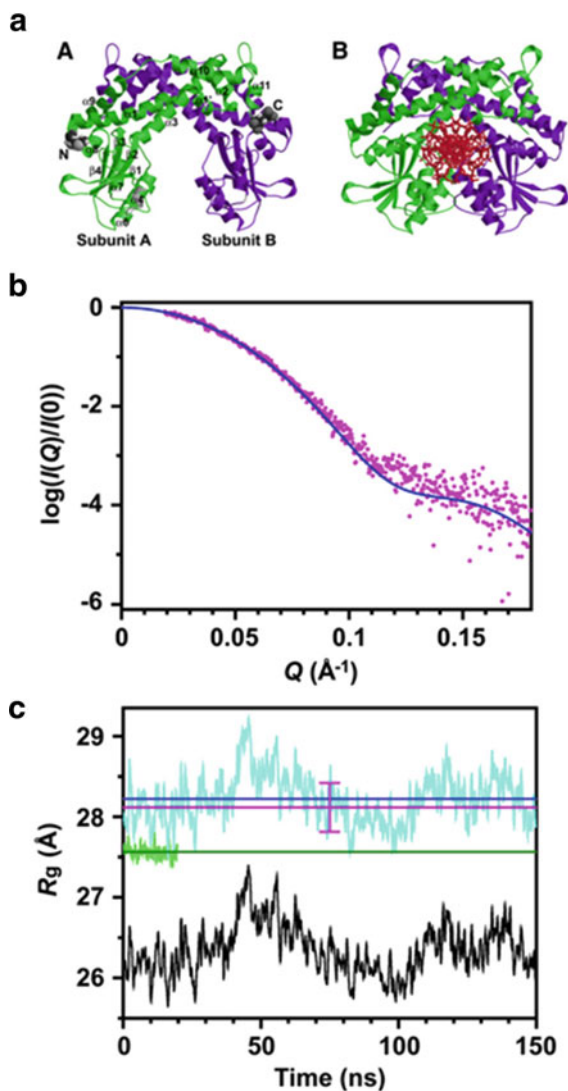
### ***15.3.1 Investigation of Intrinsic Dynamics of EcoO109I and Extensions of MD-SAXS Methods***

*EcoO109I* is a type II restriction endonuclease that recognizes specific nucleotide sequences. The crystal structures of both the DNA-free and DNA-bound forms have been determined (Fig. 15.1a), and SAXS measurements for the DNA-free form have been carried out. According to experiments, a homodimer is a functional unit in solution, and *EcoO109I* consists of the dimerization domain and the catalytic domain. In the DNA-bound form, each catalytic domain tucks a double-stranded DNA such as a scissor. By comparing the crystal structures of the DNA-free and DNA-bound forms, *EcoO109I* is supposed to undergo a conformational change after binding DNA. However, the space between the two catalytic domains is not large enough to bind DNA in the crystal structure for the DNA-free form.

To probe the solution structure of the DNA-free form, an all-atom MD simulation was carried out. To understand intrinsic dynamics, a structural ensemble consistent with the experimental SAXS data was necessary. To this end, a computational method was developed to calculate a theoretical SAXS profile by a structural ensemble, termed MD-SAXS, and it was used to assess the structural ensemble.

The formulation and procedure of MD-SAXS is as follows. In SAXS experiments, scattering from both the buffer-only solution and the protein solution is measured to subtract the effect of the solvent-excluded volume from the scattering intensity of the protein solution, and then, the scattering intensity of the buffer solution is subtracted from that of the protein solution. In MD-SAXS, just as in experiments, the two MD simulations for pure solvent and protein solution were

**Fig. 15.1** (a) Crystal structures of *EcoO109I* of DNA-free (sub-letter A) and DNA-bound (sub-letter B) forms. (Oroguchi et al. 2009) (b) Experimental (pink dots) and theoretical profiles (blue curve) of the DNA-free *EcoO109I*. (Oroguchi et al. 2009) (c) Simulation time dependence of protein-water  $R_g$  (cyan dots) estimated from Guinier plot for the theoretical profile obtained from a 150 ns simulation, protein-water  $R_g$  calculated from a restraint-MD trajectory (green dots), and protein-only  $R_g$  calculated from the 150 ns simulation. Horizontal blue and green lines represent averages of protein-water  $R_g$  over the 150 ns simulation and the restraint-MD trajectory, respectively. The pink line and the error bar show  $R_g$  estimated from Guinier plot for the experimental profile and its error. (Oroguchi et al. 2009)



performed, and then the theoretical SAXS profile ( $I(\mathbf{Q})$ ) was obtained by subtracting the two theoretical scattering intensities of protein-solution ( $I^U(\mathbf{Q})$ ) and pure-solvent ( $I^V(\mathbf{Q})$ ) MD simulations as

$$I(\mathbf{Q}) = I^U(\mathbf{Q}) - I^V(\mathbf{Q}),$$

where  $\mathbf{Q}$  is the scattering vector. Because the experimental SAXS profile is an averaged quantity over the orientational and configurational degree of freedom of a protein in solution, the theoretical scattering intensity is defined by

$$I(\mathbf{Q}) = \left\langle \left\langle I'(\mathbf{Q}) \right\rangle_{\Omega_{\mathbf{Q}}} \right\rangle_{\text{MD}},$$

where  $I'$  is the instantaneous scattering intensity,  $\langle X \rangle_{\Omega_{\mathbf{Q}}}$  represents the orientational average, and  $\langle X \rangle_{\text{MD}}$  represents the configurational average of all snapshots in the trajectory. The instantaneous intensity corresponding to a snapshot in the trajectory is given by

$$I'(\mathbf{Q}) = \iint (\rho'(\mathbf{r}) - \rho_0) (\rho'(\mathbf{r}') - \rho_0) e^{-i\mathbf{Q} \cdot (\mathbf{r} - \mathbf{r}')} d\mathbf{r} d\mathbf{r}'$$

where  $\rho'(\mathbf{r})$  is the instantaneous electron density at position  $\mathbf{r}$ . A fictive 3D sphere is defined such that the protein is centered in the sphere, and the sphere includes the protein and the solvent molecules around the protein. Using the sphere, the coordinates for the protein and water molecules in a snapshot can be classified into the areas inside ('V') and outside the sphere. When the size of the sphere is large enough, the instantaneous scattering intensity with the orientational average can be given by

$$\langle I'(\mathbf{Q}) \rangle_{\Omega_{\mathbf{Q}}} \sim \left\langle \iint_{\text{V}} (\rho'(\mathbf{r}) - \rho_0) (\rho'(\mathbf{r}') - \rho_0) e^{-i\mathbf{Q} \cdot (\mathbf{r} - \mathbf{r}')} d\mathbf{r} d\mathbf{r}' \right\rangle_{\Omega_{\mathbf{Q}}}$$

where the integration is executed in the inside of the sphere (V). This assumption is valid when the density fluctuation in the area outside the sphere has no correlation with that inside the sphere. In the result, the scattering intensity can be calculated only by the contributions inside the sphere. For fast computation, the orientational average is calculated by a multipole expansion. The trajectory of a 150-ns MD simulation was used in the configurational average.

The theoretical SAXS profile calculated using MD-SAXS with the structural ensemble agreed with the experimental profile (Fig. 15.1b). In particular,  $R_g$  estimated from the Guinier approximation of the theoretical profile ( $\sim 28.2 \text{ \AA}$ ) was within errors of the experimental  $R_g$  ( $28.1 \pm 0.3 \text{ \AA}$ ). From the simulation-time dependence of  $R_g$  obtained by the SAXS profile in each 500 ps time window,  $R_g$  varied within  $\sim 1.8 \text{ \AA}$ , which was reflected in the fluctuation between the most expanded and closed conformations (Fig. 15.1c). This shows that the agreement of  $R_g$  is provided by the configurational average over conformations in the trajectory. This point also shows the importance of the protein flexibility treatment;  $R_g$  estimated from a restraint MD simulation, in which the protein structure was restrained to the crystal structure, was close to that at the most closed conformation and smaller than the experimental  $R_g$  (Fig. 15.1c). In addition,  $R_g$  estimated using only the protein ( $\sim 26.2 \text{ \AA}$ ) was smaller than those of those given by the theoretical and experimental profiles, indicating that the explicit treatment of water molecules in MD-SAXS provides an adequate description of solvent effects (Fig. 15.1c).

From the structural ensemble consistent with the experimental SAXS profile, the intrinsic dynamics of *EcoO109I* were revealed. The large motions in the trajectory were derived using principal component analysis, and the largest- and the second-largest motions were the open-close motion and the twisting motion, respectively. The motions were relevant to the function as follows. The first motion allows *EcoO109I* to interact with DNA like a scissor, and the second motion allows the two catalytic domains to fit together on the major groove of DNA from both sides. The MD simulation revealed the intrinsic dynamics, including the transiently open conformation that was necessary to access the DNA.

In ref. (Oroguchi and Ikeguchi 2011), the formulation of MD-SAXS was extended to the buffer with ions. Because the electron density of the bulk solvent depends on the ion concentration, ions in bulk significantly affect SAXS profiles. The effect of the ions was incorporated via the form factor, and the instantaneous intensity for the buffer with ions was then given by

$$\langle I'(\mathbf{Q}) \rangle_{\Omega_{\mathbf{Q}}} = \left\langle \iint_V (\rho'(\mathbf{r}) - \rho_0 f_V(\mathbf{Q})) (\rho'(\mathbf{r}') - \rho_0 f_V(\mathbf{Q})) e^{-i\mathbf{Q} \cdot (\mathbf{r} - \mathbf{r}')} d\mathbf{r} d\mathbf{r}' \right\rangle_{\Omega_{\mathbf{Q}}}$$

where  $f_V(\mathbf{Q})$  is the form factor for solvent molecules defined as

$$f_V(\mathbf{Q}) = \sum_{\mu} x_{\mu} f_{\mu}(\mathbf{Q}) / \sum_{\mu} x_{\mu} f_{\mu}(0)$$

where  $x_{\mu}$  is the number concentration of the solvent species  $\mu$ , and  $f_{\mu}(\mathbf{Q})$  is the atomic form factor. They applied the expanded MD-SAXS to hen egg white lysozyme solutions for various concentrations of NaCl. The calculated intensities showed the effects on the ion strength. The decrease in  $I(0)$  (or  $R_g$ ) was observed as the ion concentration increased, and the shape of the SAXS curves varied for different concentrations. Due to a slow mobility of ions, the convergence of the scattering intensity depends on the ion concentrations, and at least  $\sim 20$  ns simulation was necessary for a converged profile even if a protein structure was fixed. In 0 mM NaCl buffer, a simulation of  $\sim 0.2$  ns was sufficient for the convergence. To overcome the slow convergence, they developed a novel fitting method that produced the profile in the presence of ions from the MD simulation in the pure water buffer, i.e., 0 mM NaCl. Dividing the solvent region in the sphere region denoted as  $V$  into several spherical layers of thickness  $\Delta d$  in the direction perpendicular to the protein surface, they defined the density distribution of solvent molecules as

$$\rho^{\text{fit}}(i\Delta d) = \rho_{\text{water}}^{\text{fit}}(i\Delta d) + \rho_{\text{cs}}^{\text{fit}}(i\Delta d)$$

where  $\rho_{\text{water}}^{\text{fit}}$  or  $\rho_{\text{cs}}^{\text{fit}}$  are the density of water or ions in  $i$ -th layer. To obtain  $\rho_{\text{cs}}^{\text{fit}}(i\Delta d)$ , they estimated it by the sigmoid function model for the virtual density distribution of ions. The sigmoid model smoothly connects from the density of

the inside region without ions to the bulk density with ions, and there are three adjustable parameters; One parameter is the distance from the protein at which ions can approach the protein; another parameter is the smoothness of the density change of the buffer with ions; and the other parameter is the bulk density of the buffer with ions. These parameters were fitted such that the similarity score between the theoretical and experimental profiles was minimized. The fitting method successfully reproduced the theoretical profiles with the presence of ions for various concentrations from the MD simulations with the absence of ions. Furthermore, the density distribution of solvents in real space was reproduced well by the fitting method even though the fitting was carried out for the scattering intensity. These results indicate the applicability of the method and the validity of the solvent model.

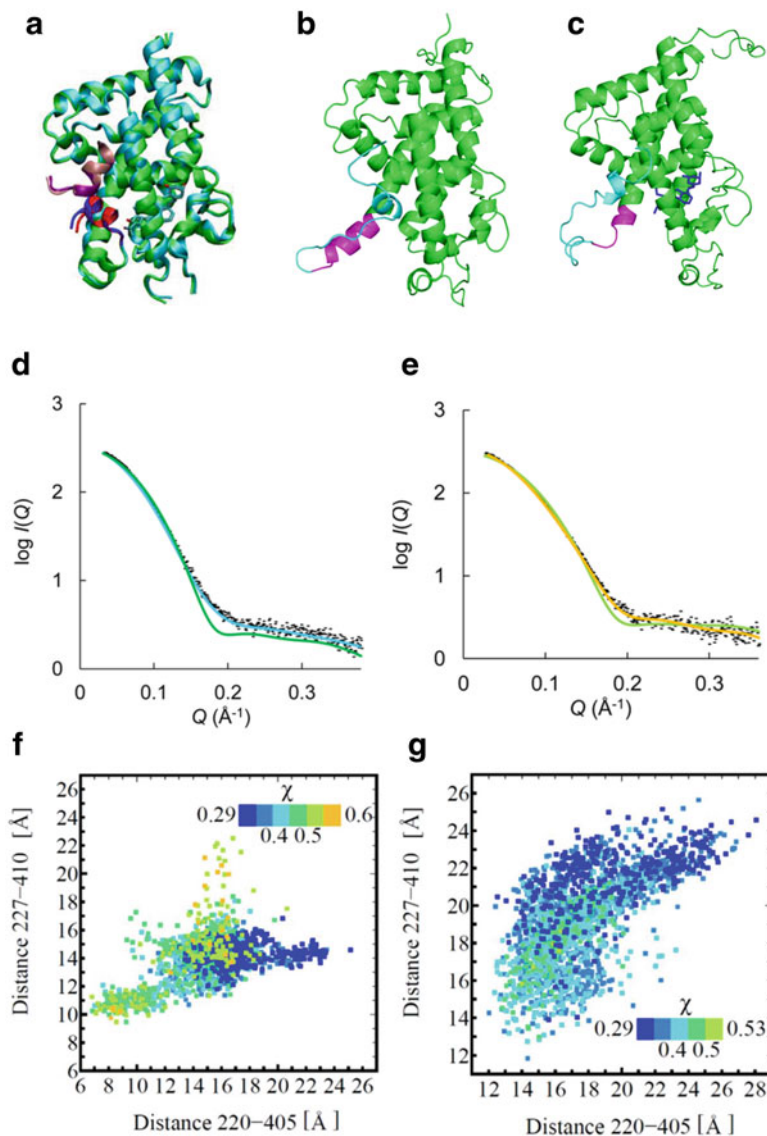
As for the limitation of the spherical boundary denoted as  $V$ , the formulation was extended to the non-spherical boundaries (Oroguchi and Ikeguchi 2012). When MD-SAXS is applied to elongate proteins, many solvent molecules are necessary in the calculation due to the limitation of the spherical region, and a cube box must be employed. However, solvent molecules at least in a rectangular box or a cylinder are sufficient for typical MD simulations in terms of accuracy and efficiency. The integration parts in instantaneous intensity were formulated, and the spherical region was changed to the box region or the cylinder region. The scattering intensities calculated from the rectangular and the cylinder regions agreed with the intensity calculated from the original sphere region. Owing to this formulation, MD-SAXS can be applied to typical settings of MD simulations for elongate proteins.

### ***15.3.2 Structural Investigation of the Vitamin D Receptor Ligand-Binding Domain***

The vitamin D receptor (VDR) is a member of the nuclear receptor (NR) family. VDR is a ligand-dependent transcription factor that regulates the expression of genes related to calcium homeostasis, immunomodulation, cell differentiation, and cell proliferation. Because the functions of VDR and other NRs are involved in human diseases, understanding their regulation by ligand binding contributes to structure-based drug design. Transcriptional regulation is conducted by sequential molecular events: ligand binding, dimerization with a partner receptor, recruitment of coregulators (coactivators/corepressors), and binding to DNA. NRs have a highly conserved DNA-binding domain and a moderately conserved ligand-binding domain (LBD). Transactivation is initiated by the conformational change of LBD induced by ligand binding. According to X-ray crystal structures of NRs and other experimental results, a local conformational change around helix 12 in the LBD is key to regulating agonism/antagonism.

Many crystal structures of agonist/antagonist-binding VDR-LBD have been solved so far. However, all the crystal structures are almost identical, regardless of agonist/antagonist binding (Fig. 15.2a). Those crystal structures are considered





**Fig. 15.2** (a) Superimposition of the agonist-binding VDR-LBD (PDB id: 2ZLC) and the antagonist-binding VDR-LBD (PDB id: 2ZXM). Helix 12 of the agonist- and antagonist-binding form is colored by red and blue, respectively, and coactivator is shown by dark-pink and purple color. (b) A solution model agreed with the experimental SAXS data for the apo form, termed ApoMD-open structure. (c) A solution model for the antagonist-binding form, termed AntagoMD-open structure. (d) Experimental profile of the apo form (black dots) and theoretical profiles of

to be the agonist form. Because the conformation of helix 12 is key, the crystal structures do not provide structural insight into the mechanism of antagonist activity. In addition, no crystal structure of the apo form has been reported, and the exact conformation of the apo form remains unknown.

To reveal the apo and antagonist-binding forms of VDR-LBD, a hybrid approach of SAXS and MD was used (Anami et al. 2016). SAXS experiments can reveal an overall shape of molecular structures in solution, and can capture both the flexible structure of the apo form and the conformational change in response to antagonist binding. SAXS profiles of apo and antagonist-binding rat VDR-LBDs were obtained. However, the profiles were different from the theoretical profiles calculated from crystal structures. The discrepancy between the theoretical and experimental profiles  $\chi$  calculated by CRY SOL (Svergun et al. 1995) showed high values ( $\chi = 0.8$  for the apo form (Fig. 15.2d),  $\chi = 0.7$  for the antagonist-binding form (Fig. 15.2e)). Although all the reported crystal structures of VDR-LBD were also fit to each experimental profile, all the theoretical curves deviated from the experimental profile. This result showed that the solution structures of both forms were different from the active (agonist) forms reported previously, and the solution structures captured by SAXS experiments reflected the inactive states.

To clarify the solution structures at atomic resolution, they conducted MD simulations and collected each structural ensemble. In this study, all-atom MD simulations were carried out for the structural investigation, and the experimental SAXS profiles were used only to judge whether the solution models generated by MD simulations were close to those in the experiment or not. To improve the generation of a structural ensemble, various initial models were prepared using homology modeling, and multiple MD simulations from the initial models were carried out. A 100-ns MD simulation was performed in each initial model, and a snapshot was saved every 50 ps (2000 snapshots in total). The theoretical profile of each snapshot structure was calculated using CRY SOL (Svergun et al. 1995), and it was compared to the experimental profile. Then, a reliable structure for each apo and antagonist-binding form was reported.

The structural investigation by MD simulations successfully provided a solution model of each apo and antagonist-binding form that were consistent with the experimental SAXS profile. Compared to the simulation time dependences of  $\chi$  for various MD simulations, an MD simulation generated solution structures with low  $\chi$ , and then the ensemble generated by the MD simulation was selected.

---

**Fig. 15.2** (continued) the crystal structure (green curve) and the ApoMD-open structure (cyan curve). (Anami et al. 2016) (e) Experimental profile of the antagonist-binding form (black dots) and theoretical profiles of the crystal structure (light-green curve) and the AntagoMD-open structure (orange curve). (Anami et al. 2016) (f) Distribution of  $\chi$  for various snapshots generated from an apo-form MD simulation on the distance map. (Anami et al. 2016) (g) Distribution of  $\chi$  for various snapshots generated from an MD simulation of the antagonist-binding form on the distance map. (Anami et al. 2016). (Fig. 15.2d–g: Reprinted with permission from Anami et al. *J. Med. Chem.* 59:7888–7900 (2016). Copyright 2016 American Chemical Society)

In the ensemble, a solution model was selected as a snapshot with the lowest  $\chi$ , and it was referred to as the ApoMD-open (Fig. 15.2b) and AntagoMD-open (Fig. 15.2c) structures. The theoretical SAXS profiles of both forms agreed with each experimental profile ( $\chi = 0.29$  for both forms (Fig. 15.2d, e)). In both forms, helix 12 was partially unraveled and did not adopt the active form. In the ApoMD-open structure, helix 11 bent outward in a kink-centered hinge-bending motion, and the motion created a wide entrance leading to the ligand-binding pocket (LBP). In the AntagoMD-open structure, the wide entrance of the LBP was created by wide and flexible loop between helices 11 and 12 (loop 11–12).

To check how the structural feature was invariant against the various possible conformations sampled by the MD simulation, the tendency between the structural feature and  $\chi$  was analyzed. The common structural feature in both solution models was the wide entrance. To characterize the entrance width, two distances between residues were defined, and  $\chi$  for every snapshot was mapped onto the surface of the distances (Fig. 15.2f, g). The distribution of  $\chi$  showed that  $\chi$  negatively correlated with the entrance width. The structures with a wide entrance, like Apo/AntagoMD-open, showed low  $\chi$ . This result showed that the solution structures observed in SAXS experiments mainly fluctuated near the ApoMD-open or AntagoMD-open structures. To further check the validity of the relationship, cross validation analysis was performed. The snapshots generated by the MD simulation for the apo (or antagonist-binding) form were fitted to the experimental SAXS profile for the antagonist-binding (or apo) form, and vice versa. The cross-validation analysis showed distinctively different behavior in the fluctuation around the LBP in each apo and antagonist-binding form. Interestingly, several snapshots for the apo form showed low  $\chi$  values when they were fitted to the SAXS profile of the antagonist-binding form. However, the structures of the apo form were not suitable for antagonist binding because clashes among the antagonist and residues occurred. Thus, the structure around the LBP sampled by the MD simulation for the antagonist-binding form was essential for retaining ligand binding. This analysis demonstrated an advantage of this hybrid approach. When the structural investigation is done using only an experimental SAXS profile, the unsuitable structure can be selected as a consistent model to satisfy the SAXS profile. In this hybrid approach, the unsuitable structure is never selected because the unsuitable structure does not appear in MD simulations of the agonist-binding form.

This hybrid approach of SAXS and MD simulation provided a solution model of apo and antagonist-binding structures. According to the SAXS experiments of the agonist-binding form by Rochel et al. (2001), the obtained SAXS profile was consistent with the theoretical profile calculated from the crystal structure, indicating that the solution structure of the agonist-binding form is identical to the crystal structure. Integrating the structural information about the apo and agonist/antagonist-binding forms, they proposed a model for agonist/antagonist activity controlled by ligand binding called the “folding-door model” in which helix 11 acts as a door for ligand-binding unlike the mouse-trap model (Moras and Gronemeyer 1998). However, further analyses for the structural ensemble and the whole VDR are the next steps. Helix 12 and helix 11 may fluctuate between apo and antagonist-binding forms,

and structural information about their structural ensembles allow us to further understand conformations influenced by ligands. As a straightforward expansion for the MD-SAXS method, a comparison of the theoretical profile for an ensemble structure weighted by the existing probability estimated by the suitable techniques (e.g., Markov state model) with the experimental profile will be suitable. Cross-validations from the analyzing techniques of SAXS (e.g., EROS (Rozycki et al. 2011)) and SAXS-driven MD (e.g., SWAXS-driven MD (Chen and Hub 2015)) will be applicable. SAXS experiments for the whole heterodimer VDR-RXR have been performed by Rochel et al. (Rochel et al. 2011). Since the dimer is large and flexible, CG representation will be useful. The analyzing technique with a CG model (e.g., Fast-SAXS (Yang et al. 2009) and BSS-SAXS (Yang et al. 2010)) will be applicable.

## 15.4 Conclusion

Small-angle X-ray scattering (SAXS) is an efficient experimental tool to measure the overall shape of macromolecular structures under nearly physiological aqueous conditions. Due to the low resolution of SAXS data, high-resolution data obtained from X-ray crystallography, NMR, or other physicochemical methods is necessary to understand protein functions based on structures at atomic resolution. Thus, SAXS is complementary to other methods. In this review, we focused on hybrid approaches of SAXS and *in silico* methods, and typical and effective methods were introduced. The methods will be useful for obtaining theoretical SAXS profiles from (ensemble) structures with adequate treatments of solvent effects and to estimate reasonable structures consistent with experimental SAXS profiles. The combination analysis of SAXS and molecular dynamics (MD) simulations is a promising method to estimate solution structures and structural ensembles for flexible proteins. The use of MD simulations provides a physicochemically proper structural ensemble in solution and a precise description of solvent effects. Two approaches of such combination analysis have been proposed so far. The first approach is the SAXS-driven MD simulation in which artificial forces defined by experimental SAXS data are employed to modify the weights of structural clusters during MD simulations. The second approach is the MD-SAXS method in which MD simulations are carried out without any knowledge of experimental SAXS data, and the experimental SAXS data are used only to assess the consistency of the solution model from MD simulations with those observed in experiments. Since the second approach can avoid the excessive modification of structural ensembles fitted to the experimental profile, we reviewed examples using the second approach. The first example is an investigation of the intrinsic dynamics of *EcoO109I* (Oroguchi et al. 2009). To investigate dynamics, the computational method to obtain a theoretical SAXS profile from the trajectory of an MD simulation was developed. The method provides accurate profiles from a structural ensemble, and intrinsic dynamics are revealed from analyses of the ensemble consistent with the experimental profile. The second example is a structural investigation of the vitamin D receptor ligand-

binding domain for the apo and antagonist-binding forms (Anami et al. 2016). Theoretical SAXS profiles for all the reported crystal structures deviate from the experimental profiles. However, MD simulations successfully provided solution models consistent with the experimental profiles. The structural features of the solution models are reasonable from the viewpoint of their functions. These examples demonstrate the applicability of the hybrid approach of SAXS and MD simulations. This approach and other related methods allow us to understand the relationship between functions and structures on the basis of experimental and physicochemical rationales.

**Acknowledgements** This work was financially supported by Innovative Drug Discovery Infrastructure through Functional Control of Biomolecular Systems, Priority Issue 1 in Post-K Supercomputer Development from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) to M.I. (Project ID: hp150269, hp160223, hp170255, and hp180191); by Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS) (Project ID: JP17am0101109) from Japan Agency for Medical Research and Development (AMED) to M.I.; and by RIKEN Dynamic Structural Biology Project to M.I. We further thank collaborators, Dr. Tomotaka Oroguchi (Keio Univ.), Prof. Hiroshi Hashimoto (Univ. of Shizuoka), Prof. Toshiyuki Shimizu (Tokyo Univ.), Prof. Mamoru Sato (Yokohama City Univ.), Dr. Yasuaki Anami (Univ. of Texas), Dr. Nobutaka Shimizu (KEK), Dr. Daichi Egawa (Showa Pharmaceutical Univ.), Dr. Toshimasa Itoh (Showa Pharmaceutical Univ.), and Prof. Keiko Yamamoto (Showa Pharmaceutical Univ.).

## References

- Abraham MJ, Murtola T, Schulz R, Pall S, Smith JC, Hess B, Lindahl E (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1-2:19–25
- Alva V, Nam SZ, Söding J, Lupas AN (2016) The MPI bioinformatics toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nuc Acid Res* 44:W410–W415
- Anami Y, Shimizu N, Ekimoto T, Egawa D, Itoh T, Ikeguchi M, Yamamoto K (2016) Apo- and antagonist-binding structures of vitamin D receptor ligand-binding domain revealed by hybrid approach combining small-angle x-ray scattering and molecular dynamics. *J Med Chem* 59:7888–7900
- Beauchamp KA, Lin YS, Das R, Pande VS (2012) Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J Chem Theory Comput* 8:1409–1414
- Bernado P (2010) Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering. *Eur Biophys J* 39:769–780
- Bernado P, Svergun DI (2012) Structural analysis of intrinsically disordered proteins by small-angle x-ray scattering. *Mol Bio Syst* 8:151–167
- Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle x-ray scattering. *J Am Chem Soc* 129:5656–5664
- Boldon L, Laliberte F, Liu L (2015) Review of the fundamental theories behind small angle x-ray scattering, molecular dynamics simulations, and relevant integrated application. *Nano Rev* 6:25661

- Case DA, Cerutti DS, Cheatham TE III, Darden TA, Duke RE, Giese TJ, Gohlke H, Goetz AW, Greene D, Homeyer N, Izadi S, Kovalenko A, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Mermelstein D, Merz KM, Monard G, Nguyen H, Omelyan I, Onufriev A, Pan F, Qi R, Roe DR, Roitberg A, Sagui C, Simmerling CL, Botello-Smith WM, Swails J, Walker RC, Wang J, Wolf RM, Wu X, Xiao L, York DM, Kollman PA (2017) AMBER 2017 University of California, San Francisco
- Chen P, Hub JS (2015) Interpretation of solution x-ray scattering by explicit-solvent molecular dynamics. *Biophys J* 108:2573–2584
- dos Reis MA, Apricio R, Zhang Y (2011) Improving protein template recognition by using small-angle x-ray scattering profiles. *Biophys J* 101:2770–2781
- Dror RO, Dirks RM, Grossman JP, Xu H, Shaw DE (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys* 41:429–452
- Fiser A (2010) Template-based protein structure modeling. *Methods Mol Biol* 673:73–94
- Förster F, Webb B, Krukenberg KA, Tsuruta H, Agard DA, Sali A (2008) Integration of small angle x-ray scattering data into structural modeling of proteins and their assemblies. *J Mol Biol* 382:1089–1106
- Goh BC, Hadden JA, Bernardi RC, Singharoy A, McGreevy R, Rudack T, Cassidy CK, Schulten K (2016) Computational methodologies for real-space structural refinement of large macromolecular complexes. *Annu Rev Biophys* 45:253–278
- Garba C, Tama F (2010) Normal mode flexible fitting of high-resolution structures of biological molecules toward SAXS data. *Bioinform Biol Insights* 4:43–54
- Grishaev A, Wu J, Trehwella J, Bax A (2005) Refinement of multidomain protein structures by combination of solution small-angle x-ray scattering and NMR data. *J Am Chem Soc* 127:16621–16628
- Grishaev A, Guo L, Irving T, Bax A (2010) Improved fitting of solution x-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. *J Am Chem Soc* 132:15484–15486
- Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J Chem Phys* 120:11919–11929
- Hammel M (2012) Validation of macromolecular flexibility in solution by small-angle x-ray scattering (SAXS). *Eur Biophys J* 41:789–799
- Harrigan MP, Sultan MM, Hernandez CX, Husic BE, Eastman P, Schwantes CR, Beauchamp KA, McGibbon RT, Pande VS (2017) MSMBuilder: statistical models for biomolecular dynamics. *Biophys J* 112:10–15
- Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL, 2nd, Tsutakawa SE, Jenny FE, Jr., Classen S, Frankel KA, Hopkins RC, Yang Sj, Scott JW, Dillard BD, Adams MW, Tainer, JA (2009) Robust, high-throughput solution structural analysis by small angle X-ray scattering (SAXS). *Nat Methods* 6(8):606–612
- Jacques DA, Guss JM, Svergun DI, Trehwella J (2012) Publication guidelines for structural modeling of small-angle scattering data from biomolecules in solution. *Acta Cryst D* 68:620–626
- Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T (2009) The SWISS-MODEL repository and associated resources. *Nuc Acid Res* 37:D387–D392
- Kikhney AG, Svergun DI (2015) A practical guide to small angle x-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* 589:2570–2577
- Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. *Nuc Acid Res* 32:W526–W531
- Kimanius D, Pettersson I, Schluchebier G, Lindahl E, Andersson M (2015) SAXS-guided metadynamics. *J Chem Theory Comput* 11:3491–3498
- Knight CJ, Hub JS (2015) WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. *Nuc Acid Res* 43:W225–W230
- Kobayashi C, Jung J, Matunaga Y, Mori T, Ando T, Tamura K, Kamiya M, Sugita Y (2017) GENESIS 1.1: a hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms. *J Comput Chem* 38:2193–2206

- Köfinger J, Hummer G (2013) Atomic-resolution structural information from scattering experiments on macromolecules in solution. *Phys Rev E* 87:052712
- Kojima M, Timchenko AA, Higo J, Ito K, Kihara H, Takahashi K (2004) Structural refinement by restrained molecular-dynamics algorithm with small-angle x-ray scattering constraints for a biomolecule. *J Appl Cryst* 37:103–109
- Konarev PV, Volkov VV, Sokolova AV, Koch MHJ, Svergun DI (2003) PRIMUS – a windows-pc based system for small-angle scattering data analysis. *J Appl Cryst* 36:1277–1282
- Kozakov D, Hall DR, Xia B, Porter KA, Pothorny D, Yueh C, Beglov D, Vajda S (2017) The ClusPro web server for protein-protein docking. *Nat Protoc* 12:255–278
- Lane TJ, Shukla D, Beauchamp KA, Pande VS (2013) To milliseconds and beyond: challenges in the simulation of protein folding. *Curr Opin Struct Biol* 23:58–65
- Lau AY, Roux B (2007) The free energy landscapes governing conformational changes in a glutamate receptor ligand-binding domain. *Structure* 15:1203–1214
- Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE (2012) Systematic validation of protein force fields against experimental data. *PLoS One* 7:e32131
- Liu H, Morris RJ, Hexemer A, Grandison S, Zwart PH (2012) Computation of small-angle scattering profiles with three-dimensional Zernike polynomials. *Acta Cryst A* 68:278–285
- Marchi M (2016) A first principle particle mesh method for solution SAXS of large bio-molecular systems. *J Chem Phys* 145:045101
- Moras D, Gronemeyer H (1998) The nuclear receptor ligand-binding domain: structure and function. *Curr Opin Cell Biol* 10:384–391
- Morimoto Y, Nakagawa T, Kojima M (2013) Small-angle x-ray scattering constraints and local geometry like secondary structures can construct a coarse-grained protein model at amino acid residue resolution. *Biochem Biophys Res Commun* 431:65–69
- Nguyen HT, Pabit SA, Meisburger SP, Pollack L, Case DA (2014) Accurate small and wide angle x-ray scattering profiles from atomic models of proteins and nucleic acids. *J Chem Phys* 141:22D508
- Ohmura I, Morimoto G, Ohno Y, Hasegawa A, Taiji M (2014) MDGRAPE-4: a special-purpose computer system for molecular dynamics simulations. *Philos Trans A Math Phys Eng Sci* 372:20130387
- Oroguchi T, Ikeguchi M (2011) Effects of ionic strength on SAXS data for proteins revealed by molecular dynamics simulations. *J Chem Phys* 134:025102-1-14
- Oroguchi T, Ikeguchi M (2012) MD-SAXS method with nonspherical boundaries. *Chem Phys Lett* 541:117–121
- Oroguchi T, Hashimoto H, Shimizu T, Sato M, Ikeguchi M (2009) Intrinsic dynamics of restriction endonuclease EcoO109I studied by molecular dynamics simulations and x-ray scattering data analysis. *Biophys J* 96:2808–2822
- Park S, Bardhan JP, Roux B, Makowski L (2009) Simulated x-ray scattering of protein solutions using explicit-solvent models. *J Chem Phys* 130:134114-1-8
- Pelikan M, Hura GL, Hammel M (2009) Structure and flexibility within proteins as identified through small angle x-ray scattering. *Gen Physiol Biophys* 28:174–189
- Petoukhov MV, Svergun DI (2005) Global rigid body modelling of macromolecular complexes against small-angle scattering data. *Biophys J* 89:1237–1250
- Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, Gorba C, Mertens HDT, Konarev PV, Svergun DI (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Cryst* 45:342–350
- 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* 26:1781–1802
- Piana S, Laio A (2007) A bias-exchange approach to protein folding. *J Phys Chem B* 111:4553–4559
- Poitevin F, Orland H, Doniach S, Koehl P, Delarue M (2011) AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformly hydrated atomic models. *Nuc Acid Res* 39:W184–W189

- Pons C, D'Abramo M, Svergun DI, Orozco M, Bernado P, Fernandez-Recio J (2010) Structural characterization of protein-protein complexes by integrating computational docking with small-angle scattering data. *J Mol Biol* 403:217–230
- Rambo RP, Tainer JA (2013) Super-resolution in solution x-ray scattering and its applications to structural systems biology. *Annu Rev Biophys* 42:415–441
- Rauscher S, Gapsys V, Gajda MJ, Zweckstetter M, de Groot BL, Grubmüller H (2015) Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J Chem Theory Comput* 11:5513–5524
- Ravikumar KM, Huang W, Yang S (2013) Fast-SAXS-pro: a unified approach to computing SAXS profiles of DNA, RNA, protein, and their complexes. *J Chem Phys* 138:024112-1-7
- Rochel N, Tocchini-Valentini G, Egea PF, Juntunen K, Garnier JM, Vihko P, Moras D (2001) Functional and structural characterization of the insertion region in the ligand binding domain of vitamin D nuclear receptor. *Eur J Biochem* 268:971–979
- Rochel N, Ciesielski F, Godet J, Moman E, Roessle M, Peluso-Iltis C, Moulin M, Haertlein M, Callow P, Mely Y, Svergun DI, Moras D (2011) Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings. *Nat Struct Mol Biol* 18:564–570
- Rozycki B, Kim YC, Hummer G (2011) SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* 19:109–116
- Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815
- Saunders MG, Voth GA (2013) Coarse-graining methods for computational biology. *Annu Rev Biophys* 42:73–93
- Scherer MK, Trendelkamp-Schroer B, Paul F, Perez-Hernandez G, Hoffmann M, Plattner N, Wehmeyer C, Prinz JH, Noe F (2015) PyEMMA2: a software package for estimation, validation, and analysis of Markov models. *J Chem Theory Comput* 11:5525–5542
- Schneidman-Duhobny D, Hammel M, Sali A (2011) Macromolecular docking restrained by a small angle x-ray scattering profile. *J Struct Biol* 173:461–471
- Schneidman-Duhovny D, Kim SJ, Sali A (2012) Integrative structural modeling with small angle x-ray scattering profiles. *BMC Struct Biol* 12:7
- Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2013) Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys J* 105:962–974
- Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ, Chao JC (2008) Anton, a special-purpose machine for molecular dynamics simulation. *Commun Acm* 51:91–97
- Stovgaard K, Andreetta C, Ferkinghoff-Borg J, Hamelryck T (2010) Calculation of accurate small angle x-ray scattering curves from coarse-grained protein models. *BMC Bioinform* 11:429
- Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314:141–151
- Svergun DI (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys J* 77:2879–2886
- Svergun DI, Koch MJH (2003) Small-angle scattering studies of biological macromolecules in solution. *Rep Prog Phys* 66:1735–1782
- Svergun DI, Barberato C, Koch MJH (1995) CRY SOL – a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Cryst* 28:768–773
- Svergun DI, Petoukhov MV, Koch MHJ (2001) Determination of domain structure of proteins from x-ray solution scattering. *Biophys J* 80:2946–2953
- Tria G, Mertens HD, Kachala M, Svergun DI (2015) Advanced ensemble modeling of flexible macromolecules using x-ray solution scattering. *IUCrJ* 26:207–217
- Venditti V, Egner TK, Clore GM (2016) Hybrid approaches to structural characterization of conformational ensembles of complex macromolecular systems combining NMR residual dipolar couplings and solution x-ray scattering. *Chem Rev* 116:6305–6322
- Vestergaard B (2016) Analysis of biostructural changes, dynamics, and interactions – small-angle x-ray scattering to the rescue. *Arch Biochem Biophys* 602:69–79



- Virtanen JJ, Makowski L, Sosnick TR, Freed KF (2011) Modeling the hydration layer around proteins: applications to small- and wide-angle x-ray scattering. *Biophys J* 101:2061–2069
- Weinan E, Vanden-Eijnden E (2010) Transition-path theory and path-finding algorithms for the study of rare events. *Annu Rev Phys Chem* 61:391–420
- Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293:321–331
- Yang S, Park S, Makowski L, Roux B (2009) A rapid coarse residue-based computational method for x-ray solution scattering characterization of protein folds and multiple conformational states of large protein complexes. *Biophys J* 96:4449–4463
- Yang S, Blachowicz L, Makowski L, Roux B (2010) Multidomain assembled states of Hck tyrosine kinase in solution. *Proc Natl Acad Sci U S A* 107:15757–15762
- Yang Z, Lasker K, Schneidman-Duhovny D, Webb B, Huang CC, Pettersen EF, Goddard TD, Meng EC, Sali A, Ferrin TE (2012) UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J Struct Biol* 179:269–278
- Zheng W, Tekpinar M (2011) Accurate flexible fitting of high-resolution protein structures to small-angle x-ray scattering data using a coarse-grained model with implicit hydration shell. *Biophys J* 101:2981–2991
- Zuckerman DM, Chong LT (2017) Weighted ensemble simulation: review of methodology, applications, and software. *Annu Rev Biophys* 46:43–57

**Part IV**  
**Data Validation and Archives**  
**for Hybrid Methods**

# Chapter 16

## Archiving of Integrative Structural Models



Helen M. Berman, Jill Trehwella, Brinda Vallat, and John D. Westbrook

**Abstract** Integrative or hybrid structural biology involves the determination of three-dimensional structures of macromolecular assemblies by combining information from a variety of experimental and computational methods. Archiving the results of integrative/hybrid modeling methods have complex requirements and existing archiving mechanisms are insufficient to handle these pre-requisites. Three concepts important for archiving integrative/hybrid models are presented in this chapter: (1) building a federated network of structural model and experimental data archives, (2) development of a common set of data standards, and (3) creation of mechanisms for interoperation and data exchange among the repositories in a federation. Methods proposed for achieving these objectives are also discussed.

**Keywords** Protein Data Bank · Integrative/hybrid modeling methods · PDBx/mmCIF · Data standards · Data exchange · Structural biology federation

---

H. M. Berman (✉)

RCSB Protein Data Bank, Department of Chemistry and Chemical Biology, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA  
e-mail: [berman@rcsb.rutgers.edu](mailto:berman@rcsb.rutgers.edu)

J. Trehwella

School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW, Australia

Department of Chemistry, University of Utah, Salt Lake City, UT, USA

B. Vallat

RCSB, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

J. D. Westbrook

RCSB Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

## 16.1 Introduction

The field of structural biology has undergone dramatic growth and change in the 60 plus years since Kendrew determined the structure of myoglobin (Kendrew et al. 1958) and Perutz the structure of hemoglobin (Perutz et al. 1960) – the first atomic structures of macromolecular proteins determined using X-ray crystallography. Today, while individual biomolecular structures of the highest resolution and accuracy remain central to the field, the next frontier in structural molecular biology is characterization of the large, complex and dynamic macromolecular networks and machinery that drive fundamental biological processes such as replication, transcription, concerted movement, defense against infection, etc. These targets are elusive to traditional approaches to structure determination that use a single technique, such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy or 3D Electron Microscopy (3DEM). To address this problem, integrative or hybrid (I/H) methods are being developed that combine data from complementary experimental techniques and computational models in innovative ways (Sali et al. 2015, Ward et al. 2013). For example, I/H methods have been used to develop detailed molecular models of the molecular machines and assemblies that control protein biosynthesis (ribosome) (Leitner et al. 2016), the movement of proteins across the nuclear membrane in a cell (nuclear pore complex) (Kim et al. 2018), sensing in pathogenic bacteria that enables infection (bacterial type III secretion system) (Loquet et al. 2012), and the regulation of the degradation of damaged, malfunctioning or toxic proteins in the cell (proteasomal lid sub-complex) (Politis et al. 2014).

The Protein Data Bank (PDB), founded in 1971 with only seven protein structures (Protein Data Bank 1971), is today a searchable, open global archive that holds more than 140,000 structures of biological macromolecules and their complexes, all of which are freely accessible. The vast majority of deposited structures have been determined by a single technique: X-ray crystallography, NMR spectroscopy or 3D electron microscopy. The Model Archive (MA) (Haas et al. 2013; Haas and Schwede 2013), managed by the Protein Model Portal (PMP), archives about 1400 *in silico* models derived using purely computational techniques. Well-developed infrastructure is in place for these structural model archives, with efficient deposition and data processing procedures along with data standards, validation and curation methods.

The increasingly diverse data types used in I/H methods has led to models that can span multiple spatiotemporal scales and conformational states. Therefore, existing archiving mechanisms that are designed for individual atomistic structures, are insufficient to capture the details of an I/H model. The necessary requirements for processing and archiving I/H models have yet to be fully established. In recognition of this problem, the worldwide PDB (wwPDB) (Berman et al. 2007) established the I/H Methods Task Force, and in October 2014, a workshop was held (Sali et al. 2015) at the European Bioinformatics Institute, Hinxton, UK. Thirty-eight leaders in experimental structural biology, *in silico* and integrative modeling,

visualization, and data archiving discussed the steps required to make the results of I/H modeling publicly available. They converged on the set of recommendations summarized below (Sali et al. 2015):

**Recommendation 1:** In addition to archiving the models themselves, all relevant experimental data and metadata as well as experimental and computational protocols should be archived; inclusivity is key.

**Recommendation 2:** A flexible model representation needs to be developed, allowing for multi-scale models (with atomistic and non-atomistic representations), multi-state models (existing in various conformations), ensembles of models, and models related by time or other order.

**Recommendation 3:** Procedures for estimating the uncertainty of integrative models should be developed, validated, and adopted.

**Recommendation 4:** A federation of model and data archives should be created.

**Recommendation 5:** Publication standards for integrative models should be established.

Implementation of these recommendations will take years of research and community building efforts. However, the key recommendations involving the creation of a federated system of model and data archives and the development of a flexible data representation are crucial for archiving I/H models and hence are being addressed presently.

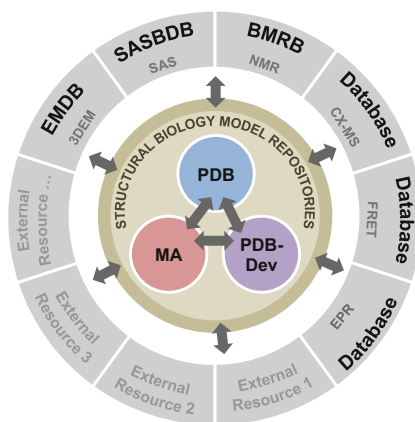
## 16.2 The Structural Biology Federation

Models determined by I/H methods utilize the data from a wide range of biophysical methodologies, including but not limited to: X-ray crystallography, NMR spectroscopy, 3DEM, Small Angle Scattering (SAS), Förster Resonance Energy Transfer (FRET), Chemical Crosslinking and Mass Spectrometry (CX-MS), Electron Paramagnetic Resonance (EPR) spectroscopy, Atomic Force Microscopy (AFM), deep sequencing and coevolution methods and other proteomics and bioinformatics techniques (Ward et al. 2013; Whitehead et al. 2012; Hopf et al. 2014). Experimental data from complementary methods are combined to provide a set of spatial restraints and structural information that are used in the determination of the three-dimensional structures of macromolecular assemblies. Currently, these data are stored in a variety of places. The atomic coordinates of structural models derived by X-ray crystallography, NMR spectroscopy, and 3DEM are archived in the *PDB* (Berman et al. 2000) along with data needed for model validation such as the structure factors from X-ray crystallography and NMR chemical shifts. There are also several experimental data repositories that store information belonging to the particular domain: the Electron Microscopy Data Bank (*EMDB*) (Patwardhan and Lawson 2016) (Lawson et al. 2011) archives the 3DEM maps as well as extensive metadata; BioMagResBank (*BMRB*) (Ulrich et al. 2008) contains NMR spectra, chemical shifts and other NMR-derived information such as NOE restraints and coupling constants; Small Angle Scattering Biological Data Bank (*SASBDB*) (Valentini et al. 2015) and *BIOISIS* (Rambo et al. 2017) contain small-angle scattering data

and models; members of the ProteomeXchange consortium (Vizcaino et al. 2014) including PRIDE (Vizcaino et al. 2016) and PeptideAtlas (Desiere et al. 2006) archive proteomics data as well as results from chemical crosslinking and mass spectrometry experiments. For other experimental methods, such as FRET and EPR, there are no standard mechanisms to archive the experimental data. As a result, there may be cloud-hosted data sets on external sites such as GitHub (GitHub Inc. 2007), or perhaps most commonly, un-hosted data sets not usually accessible to the public that reside in individual research laboratories.

In addition to archiving the three-dimensional coordinates of structural models, it is necessary to archive metadata describing the chemistry and the protocols used to determine the model, as well as the subset of experimental data needed to validate the models. Furthermore, many communities want and need a broader set of experimental data and metadata archived so that they can be available for future research.

To accommodate the need for an archive of validated models, and archives for the different experimental methods used to compute these models, a federated system of model and data archives was recommended by the I/H Methods Task Force (Sali et al. 2015). A conceptual diagram of this Federation is shown in Fig. 16.1. At the center of the figure are the principal structural biology model repositories, including the existing *PDB* and *MA* archives, along with a prototype *PDB-Dev* system, which hosts I/H models and associated spatial restraints (Vallat et al. 2016c, 2018; Burley et al. 2017). The outside ring includes complementary experimental data repositories that would share a subset of experimental data and metadata with the structural model repositories at the center, while continuing to provide the full complement of data for their specialist communities. An important component



**Fig. 16.1** A conceptual diagram of the proposed members of the federation. Repositories that focus on macromolecular structural models are shown in the center of the figure (structural biology model repositories), while examples of repositories that contain primary experimental data and/or derived restraints and associated metadata are shown in the outer circle. This outer circle contains only some examples of experimental data archives

of this federation is the establishment of methods for data exchange among the individual repositories. The data definitions supporting these repositories need to be well-aligned and software tools required for this purpose need to be developed. The I/H models of complex biological systems will likely evolve with time as new and different kinds of data become available. Therefore, the data exchange mechanisms should be able to support these evolutionary improvements. The creation of a Federation will provide a unified network of resources for structural biology models and data and will further enable the development of mechanisms for communication and interoperation among the different scientific communities contributing to structural biology.

### 16.3 Creation of Data Standards

One of the important pre-requisites for building an archive is the creation of data standards. The data standards, usually defined in a “dictionary” of data terms, provide the descriptions and specifications for the information stored in an archive. These data specifications include precise definitions for the data terms including their units and allowed ranges, software features, storage data formats, and data relationships and dependencies. To build an interoperable federated system of structural biology resources, it is necessary that each participating repository has well-defined data standards.

The scope of the contents to be archived varies among the data repositories. Ideally, the archived content contains the minimum information needed to accurately represent a complete and reproducible experiment. Experimental data repositories typically capture the sample conditions, the experimental methods and software tools used, the primary results and derived data, and associated metadata. Structural model repositories capture atomic and molecular descriptions along with metadata related to the structure determination method. The scope of the data content and formats for data standards among different repositories are not always the same.

The *PDB* archive uses the PDBx/mmCIF data standard (Fitzgerald et al. 2005) that grew out of an effort by the crystallographic community to define the many elements of the crystallographic experiments and the results derived from those experiments. The initial dictionary contained about 3000 data items, which is now expanded to about 6500. Terms specific for NMR and for 3DEM were added as structural models derived from those methods were deposited and processed by the *PDB*. In addition to the atomic coordinates of the models, the *PDB* also stores experimental data that are essential for validating these structures. These include X-ray structure factors, NMR chemical shifts and restraints, all of which are defined in the PDBx/mmCIF data dictionary (Fitzgerald et al. 2005).

The experimental data repositories that are members of the Federation, archive method-specific data and metadata. They require a compatible data representation that serves the needs of the community. BMRB (Ulrich et al. 2008) has a large array of NMR specific spectral data such as the chemical shifts, NOE restraints

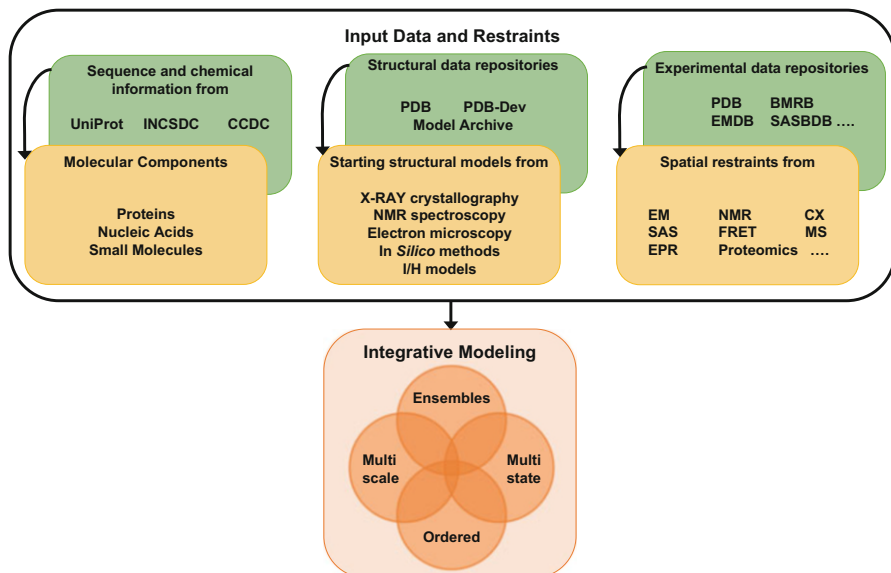
and coupling constants. The underlying data representation is based on the NMR-Star format (BioMagResBank 2004), which is a close relative of the PDBx/mmCIF data representation. EMDB (Patwardhan and Lawson 2016) (Lawson et al. 2011) contains 3DEM-derived maps expressed in CCP4 format (Winn et al. 2011) and a database that follows an internally defined XML format. SASBDB (Valentini et al. 2015) archives the results of solution scattering experiments and has adopted an extension of the PDBx/mmCIF dictionary, called sasCIF (Kachala et al. 2016; Malfois and Svergun 2000). The sasCIF extension provides SASBDB the advantages of pre-aligned data definitions and seamless interoperability with the *PDB*. Other communities that generate *in silico structural models*, CX-MS data, FRET data, EPR data, and deep genome sequencing are in various stages of creating standards for their disciplines.

The creation of an I/H model archive requires the development of a flexible data representation as recommended by the wwPDB I/H Methods Task Force. The existing data pipeline of the *PDB* archive is insufficient to handle I/H models because the *PDB* currently handles mono-scale atomistic structures derived from experimental techniques such as X-ray crystallography, NMR spectroscopy, and 3DEM. The data representation for I/H models should account for ensembles of multi-scale structural models (comprising of atomistic and coarse-grained representations of macromolecular assemblies), conformations in multiple states and models related by time or other order. It is envisioned that multi-scale I/H models can span a broad range of structures including those of individual molecules, their complexes, cellular neighborhoods, and even the entire cell. Furthermore, the input spatial restraints used in I/H modeling can be obtained from a variety of experimental and computational techniques and hence, the data representation should be able to comprehensively capture this information together with details of modeling workflows and other relevant metadata.

An I/H methods data dictionary has been created (Berman et al. 2016, Vallat et al. 2016a, b, 2018) that defines the data contents from an I/H investigation to be archived. This dictionary is an extension of the PDBx/mmCIF dictionary (Fitzgerald et al. 2005) and therefore is complementary to the definitions already present in the PDBx/mmCIF dictionary such as descriptions of the molecular system, atomic coordinates, metadata related to authors, citations, and software use. New definitions have been created to represent multi-scale structural models (including coarse-grained spheres and three-dimensional Gaussian volumes), multi-state and time ordered ensembles, starting structural models used as input in the I/H modeling and restraints derived from experimental methods such as CX-MS, 2DEM, 3DEM and SAS. Preliminary information regarding the modeling workflows and validation metrics are also defined in the dictionary. The initial set of definitions have been created based on the I/H models obtained from the Integrative Modeling Platform (IMP, (Russel et al. 2012)) software package. Figure 16.2 shows a schematic representation of the contents of the I/H methods data dictionary.

The PDB-Dev system (Vallat et al. 2016c, 2018; Burley et al. 2017) has been built based on the new I/H methods extension dictionary. At present, twenty two structures covering a variety of I/H modeling software and experimental data types





**Fig. 16.2** Illustration of the data content captured in the integrative/hybrid methods dictionary (Berman et al. 2016; Vallat et al. 2016a, b, 2018). The green boxes represent existing external repositories that archive sequence, chemical, structural, and experimental data for biological macromolecules. The yellow, orange, and blue boxes represent the information captured in the recently developed I/H methods dictionary. This information includes details of the molecular components, the starting structural models of individual molecular components, and the spatial restraints derived from various experimental methods. The details of the integrative modeling algorithm are also captured in the dictionary including definitions for multi-scale, multi-state and ordered structural ensembles of macromolecular assemblies

have been deposited into PDB-Dev. These structures and associated spatial restraints are available from the *PDB-Dev* website (Vallat et al. 2016c, 2018; Burley et al. 2017) in a format compliant with the new I/H methods data dictionary (Berman et al. 2016; Vallat et al. 2016a, b, 2018). The ChimeraX visualization software (Ferrin et al. 2017) provides basic support to visualize the multi-scale I/H models obtained from *PDB-Dev*.

Following the recommendations of the wwPDB I/H Methods Task Force, we have assembled a set of data standards and a prototype deposition and archiving system that lays the foundation for building a full-fledged archive for I/H models. The development of a comprehensive data pipeline to curate and validate these I/H structural models to provide cleaner and richer data content to the users, is the focus of ongoing research projects.

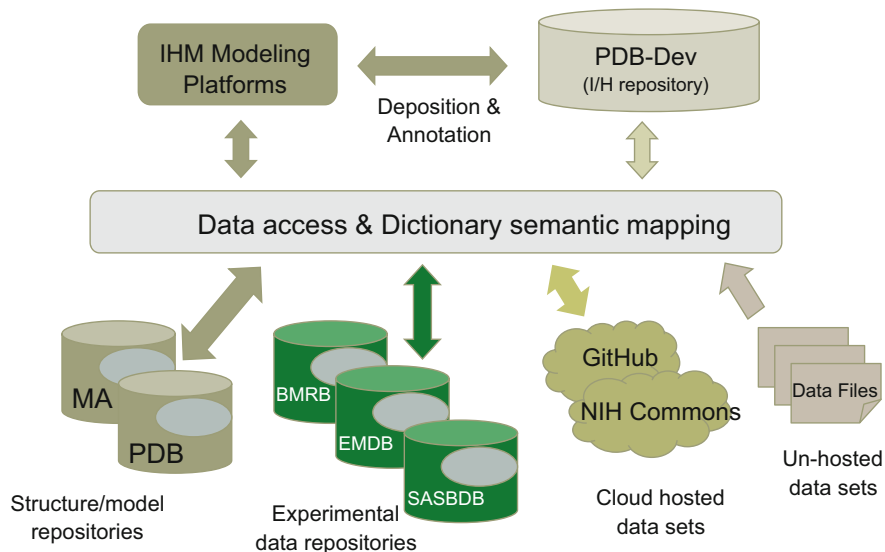
## 16.4 Methods for Data Exchange

The proposed federation comprises a network of information resources that contribute to the field of structural biology. The creation of a federation will greatly streamline the process of data preservation and access. The basis of such a federation is the establishment of mechanisms for exchanging information among its various members. This important process requires extensive participation and consensus building among the communities involved.

Experience suggests that the organization of the structural biology federation be based on autonomous repositories networked via a set of mutually agreed communication and data exchange protocols. The diversity of archived data types and data validation protocols require the greatest local autonomy in establishing data formats and standards, and to build and maintain each individual repository. Mutually agreed mechanisms are then required to enable member repositories to interoperate with each other in an effective manner including efficient methods for communication and data exchange. The objective of seamless interoperability with the federation can be achieved in several ways, as proposed below, and these may be adopted based on community consensus.

References to data residing in other repositories will rely on high level identifiers such as Digital Object Identifiers (DOIs), stable accession codes and persistent URLs. While experimental data and structural models will reside in their respective repositories, the spatial restraints and associated information derived from the experimental data, required for validation of the structural model, will be shared among the repositories. The limited set of commonly shared information need to be identified and defined accordingly to avoid duplication and to enable semantically precise data exchange. Software tools need to be developed to facilitate seamless interoperability among the repositories in the federation. These tools include development of methods for data harvesting, format conversion, semantic mapping and alignment of data residing in different repositories as well as mechanisms for exchange of data using secure industry-standard web services. Figure 16.3 shows a schematic representation of different layers of interoperability among various structural model and experimental data repositories in the proposed federation as well as developers of I/H modeling software.

To account for refinements of the structural models arising from revisions to the underlying experimental data and/or modeling methods, data exchange mechanisms should support versioning and updates to data residing in a particular repository. Timely propagation of updated information to other repositories within the federation will also need to be supported. These objectives can be achieved through mutual agreements on maintaining explicitly versioned data files and unambiguous descriptions of accession codes and version numbers within the commonly shared data definitions. Furthermore, automated messaging and communication tools are required to enable downstream dissemination of data updates.



**Fig. 16.3** A schematic portrayal of the data exchange among the structural model and experimental data repositories in the proposed structural biology federation

## 16.5 Conclusion

The future of structural biology relies heavily on the development of integrative/hybrid methods that combine information from a variety of experimental data sources with computational methods to elucidate the structures of complex macromolecular assemblies. These I/H methods are evolving into techniques that provide spatiotemporal information regarding molecular events at the cellular level. From an archival perspective, it is important to capture every structural and functional detail so that the knowledge gained from I/H models can be available for other applications in biotechnology and medicine as well as to guide future research. The structural biology community and the worldwide PDB (wwPDB, (Berman et al. 2007)) have combined their efforts to enable the archiving and dissemination of I/H models and associated experimental data and computational protocols in a concerted manner. Although the long-term vision of a comprehensive structural biology federation is yet to be fully materialized, the first steps in this direction have been productive and basic building blocks have been developed. These steps include bringing together several research communities contributing to the field of structural biology and the development of preliminary data standards and a prototype archiving system for I/H models. Further progress towards the establishment of a unified, global and interoperable network of structural biology resources that provides rich content of curated and validated structural data to the users, is the focus of ongoing and future research and community building efforts.

**Acknowledgements** This work was supported by NSF EAGER grant DBI-1519158. We thank our collaborators Andrej Sali and Benjamin Webb for their contributions towards the development of the I/H methods data dictionary and the PDB-Dev prototype system. We thank all our colleagues in the I/H methods Task Force and the members of the wwPDB for their help and support with this project.

## References

- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242
- Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* 35:D301–D303
- Berman HM, Westbrook J, Vallat B, Webb B, Sali A (2016) A data dictionary for archiving integrative/hybrid models. 66th annual meeting of the American Crystallographic Association, Denver, CO, USA, pp. 85–SA
- BioMagResBank (2004) NMRSTAR: dictionary version 3.1.1.78 [Online]. The Board of Regents of the University of Wisconsin System. Available: <http://www.bmrwisc.edu/dictionary/>. Accessed 15 Dec 2015
- Burley SK, Kurisu G, Markley J, Nakamura H, Velankar S, Berman HM, Sali A, Schwede T, Trewthella J (2017) PDB-Dev: a prototype system for depositing integrative/hybrid structural models. *Structure* 25:1317–1318
- Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R (2006) the peptide atlas project. *Nucleic Acids Res* 34:D655–D658
- Ferrin T, Huang C, Peeterson E, Goddard T, Couch G, Meng E, Morris S (2017) UCSF ChimeraX [Online]. Available: <https://www.rbvi.ucsf.edu/chimeraX/>. Accessed 5 July 2017
- Fitzgerald P MD, Westbrook JD, Bourne PE, McMahon B, Watenpaugh KD, Berman HM (2005) 4.5 macromolecular dictionary (mmCIF). In: Hall SR, McMahon B (eds) *International tables for crystallography G. Definition and exchange of crystallographic data*. Springer, Dordrecht, pp. 295–443
- GitHub Inc (2007) GitHub: how people build software [Online]. Available <http://www.github.com/>. Accessed 1 Nov 2013
- Haas J, Schwede T (2013) Model archive [Online]. Available: <http://www.modelarchive.org/>. Accessed 12 Oct 2016
- Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T (2013) The protein model portal – a comprehensive resource for protein structure and model information. *Database (Oxford)*, 2013, bat031
- Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *elife* 3
- Kachala M, Westbrook J, Svergun D (2016) Extension of the sasCIF format and its applications for data processing and deposition. *J Appl Crystallogr* 49:302–310
- Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181:662–666
- Kim SJ, Fernandez-Martinez J, Nudelman I, Shi Y, Zhang W, Raveh B, Herricks T, Slaughter BD, Hogan JA, Upla P, Chemmama IE, Pellarin R, Echeverria I, Shivaraju M, Chaudhury AS, Wang J, Williams R, Unruh JR, Greenberg CH, Jacobs EY, Yu Z, de la Cruz MJ, Mironska R, Stokes DL, Aitchison JD, Jarrold MF, Gerton JL, Ludtke SJ, Akey CW, Chait BT, Sali A, Rout MP (2018) Integrative structure and functional anatomy of a nuclear pore complex. *Nature* 555(7697):475–482
- Lawson CL, Baker ML, Best C, Bi C, Dougherty M, Feng P, van Ginkel G, Devkota B, Lagerstedt I, Ludtke SJ, Newman RH, Oldfield TJ, Rees I, Sahni G, Sala R, Velankar S, Warren J, Westbrook JD, Henrick K, Kleywegt GJ, Berman HM, Chiu W (2011) EMDDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res* 39:D456–D464

- Leitner A, Faini M, Stengel F, Aebersold R (2016) Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. *Trends Biochem Sci* 41:20–32
- Loquet A, Sgourakis NG, Gupta R, Giller K, Riedel D, Goosmann C, Griesinger C, Kolbe M, Baker D, Becker S, Lange A (2012) Atomic model of the type III secretion system needle. *Nature* 486:276–279
- Malfois M, Svergun DI (2000) sasCIF: an extension of core crystallographic information file for SAS. *J Appl Crystallogr* 33:812–816
- Patwardhan A, Lawson CL (2016) Databases and archiving for CryoEM. *Methods Enzymol* 579:393–412
- Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature* 185:416–422
- Politis A, Stengel F, Hall Z, Hernandez H, Leitner A, Walzthoeni T, Robinson CV, Aebersold R (2014) A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nat Methods* 11:403–406
- Protein Data Bank (1971) Protein Data Bank. *Nature New Biol* 233:223
- Rambo RP, Tainer JA, Hura GL (2017) BIOISIS [Online]. Available: <http://www.bioisis.net/>. Accessed 7 Aug 2017
- Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* 10:e1001244
- Sali A, Berman HM, Schwede T, Trewhella J, Kleywegt G, Burley SK, Markley J, Nakamura H, Adams P, Bonvin AM, Chiu W, Peraro MD, Di Maio F, Ferrin TE, Grunewald K, Gutmanas A, Henderson R, Hummer G, Iwasaki K, Johnson G, Lawson CL, Meiler J, Marti-Renom MA, Montelione GT, Nilges M, Nussinov R, Patwardhan A, Rappsilber J, Read RJ, Saibil H, Schroder GF, Schwieters CD, Seidel CA, Svergun D, Topf M, Ulrich EL, Velankar S, Westbrook JD (2015) Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* 23:1156–1167
- Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. *Nucleic Acids Res* 36:D402–D408
- Valentini E, Kikhney AG, Previtali G, Jeffries CM, Svergun DI (2015) SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res* 43:D357–D363
- Vallat B, Webb B, Westbrook J, Sali A, Berman H (2016a) Integrative/hybrid methods PDBx/mmCIF dictionary extension [Online]. Available: <https://github.com/ihmwg/IHM-dictionary>. Accessed 9 June 2016
- Vallat B, Webb B, Westbrook J, Sali A, Berman H (2016b) Integrative/hybrid methods PDBx/mmCIF dictionary extension documentation [Online]. Available: [https://github.com/ihmwg/IHM-dictionary/tree/master/dictionary\\_documentation](https://github.com/ihmwg/IHM-dictionary/tree/master/dictionary_documentation). Accessed 9 June 2016
- Vallat B, Webb B, Westbrook J, Sali A, Berman HM (2016c) The PDB-Dev prototype deposition and archiving system [Online]. Available: <https://pdb-dev.wwpdb.org/>. Accessed 31 Aug 2016
- Vallat B, Webb B, Westbrook JD, Sali A, Berman HM (2018) Development of a prototype system for archiving integrative/hybrid structure models of biological macromolecules. *Structure* 26(6):894–904 e2
- Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz PA, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus HJ, Albar JP, Martinez-Bartolome S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H (2014) ProteomeX change provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 32:223–226
- Vizcaino JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu QW, Wang R, Hermjakob H (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* 44:D447–D456

- Ward AB, Sali A, Wilson IA (2013) Biochemistry. Integrative structural biology. *Science* 339: 913–915
- Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, Myers CA, Kamisetty H, Blair P, Wilson IA, Baker D (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol* 30:543–548
- Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67:235–242