Springer Protocols Handbooks

Toshiya Senda
Katsumi Maenaka *Editors*

Advanced Methods in Structural Biology

Springer

SPRINGER PROTOCOLS HANDBOOKS

For further volumes:
http://www.springer.com/series/8623

Advanced Methods in Structural Biology

Editors

Toshiya Senda

Institute of Materials Structure Science, High Energy Accelerator Research Organization (KEK), Tsukuba, Japan

Katsumi Maenaka

Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan

Springer

Editors

Toshiya Senda
Institute of Materials Structure Science
High Energy Accelerator Research
Organization (KEK)
Tsukuba, Japan

Katsumi Maenaka
Faculty of Pharmaceutical Sciences
Hokkaido University
Sapporo, Japan

ISSN 1949-2448 ISSN 1949-2456 (electronic)
Springer Protocols Handbooks
ISBN 978-4-431-56028-9 ISBN 978-4-431-56030-2 (eBook)
DOI 10.1007/978-4-431-56030-2

Library of Congress Control Number: 2016942541

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Japan KK

Preface

Structural genomics studies have advanced the development of standard protocols for structural biology, leading to several automated experimental methods. As a result, it has become much easier to determine the tertiary structure of proteins. Indeed, the number of atomic coordinates of protein molecules in the Protein Data Bank (PDB) has been rapidly increasing in the last decade, with the current total exceeding 110,000 coordinates. Nonetheless, many difficult problems remain to be overcome in structural biology. For example, the structure determination of large protein complexes and membrane proteins is still difficult and often hard to accomplish using the simple standard protocols that are currently available. In many cases, researchers require more complex protocols that are specific to the protein complexes they are targeting. However, there are few textbooks focusing on these technical challenges.

The purpose of this monograph is thus to provide information to address difficult problems in structural biology and to assist in the development of a new protocol. This book will present not only advanced protocols for structural biology but also their theoretical backgrounds, which are critical to making a new protocol. This book addresses five areas: (1) protein expression and purification, (2) purification and crystallization of membrane proteins, (3) crystallization and crystal engineering, (4) interaction analysis, and (5) advanced methods for structural analyses. We hope these topics will support the many challenges faced by readers in the field of structural biology.

We greatly appreciate the contributors of these chapters as well as our colleagues who devoted their time and effort to make this book meaningful. We would also like to express our special thanks to the publisher, Springer Japan, for their generous assistance and expertise in producing this monograph.

Tsukuba, Japan — *Toshiya Senda*
Sapporo, Japan — *Katsumi Maenaka*

Contents

PART I PROTEIN EXPRESSION

PART II PURIFICATION AND CRYSTALLIZATION OF MEMBRANE PROTEINS

PART III CRYSTALLIZATION AND CRYSTAL ENGINEERING

PART IV INTERACTION ANALYSIS

PART V ADVANCED METHODS FOR STRUCTURAL ANALYSES

Contributors

Hiroki Akiba • *Department of Bioengineering, School of Engineering, The University of Tokyo, Tokyo, Japan*

Gwyndaf Evans • *Diamond Light Source, Harwell Science and Innovation Campus, Oxfordshire, UK; Membrane Protein Laboratory, Diamond Light Source, Harwell Science and Innovation Campus, Oxfordshire, UK*

James Foadi • *Diamond Light Source, Harwell Science and Innovation Campus, Oxfordshire, UK; Membrane Protein Laboratory, Diamond Light Source, Harwell Science and Innovation Campus, Oxfordshire, UK*

Shuya Fukai • *Synchrotron Radiation Research Organization and Institute of Molecular and Cellular Biosciences, The University of Tokyo, Tokyo, Japan; CREST, JST, Saitama, Japan*

Natsuki Fukuda • *Department of Analytical and Biophysical Chemistry, Graduate School of Pharmaceutical Sciences, Kumamoto University, Kumamoto, Japan*

Kiichi Fukui • *Graduate School of Engineering, Osaka University, Suita, Japan*

Atsushi Furukawa • *Laboratory of Biomolecular Science, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan*

Toshio Hakoshima • *Structural Biology Laboratory, Nara Institute of Science and Technology, Ikoma, Nara, Japan*

Kazuya Hasegawa • *JASRI/SPring-8, Sayo-gun, Hyogo, Japan*

Kunio Hirata • *RIKEN/SPring-8 Center, Sayo-gun, Hyogo, Japan; JST/PRESTO, Kawaguchi, Saitama, Japan*

Toshiaki Hosaka • *RIKEN Systems and Structural Biology Center, Yokohama, Japan; Division of Structural and Synthetic Biology, RIKEN Center for Life Science Technologies, Yokohama, Japan*

Fuyuhiko Inagaki • *Department of Structural Biology, Faculty of Advanced Life Science, Hokkaido University, Sapporo, Japan*

Tetsuo Ishida • *Department of Chemistry, Biology & Marine Science, University of the Ryukyus, Nishihara, Okinawa, Japan*

Tomomi Kimura-Someya • *RIKEN Systems and Structural Biology Center, Yokohama, Japan; Division of Structural and Synthetic Biology, RIKEN Center for Life Science Technologies, Yokohama, Japan*

Shunsuke Kita • *Center for Research and Education on Drug Discovery, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan*

Yoshihiro Kobashigawa • *Department of Analytical and Biophysical Chemistry, Graduate School of Pharmaceutical Sciences, Kumamoto University, Kumamoto, Japan*

Elena Krayukhina • *Graduate School of Engineering, Osaka University, Suita, Japan*

Seisuke Kusano • *RIKEN Systems and Structural Biology Center, Yokohama, Japan; RIKEN Structural Biology Laboratory, Yokohama, Japan*

Katsumi Maenaka • *Laboratory of Biomolecular Science and Center for Research and Education on Drug Discovery, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan*

Takayoshi Matsuda • *RIKEN Systems and Structural Biology Center, Yokohama, Japan; Center Director's Strategic Program, RIKEN Center for Life Science Technologies, Yokohama, Japan*

KAZUHIRO MIO • *Biomedical Research Institute and Molecular Profiling Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan*

HIROSHI MORIOKA • *Department of Analytical and Biophysical Chemistry, Graduate School of Pharmaceutical Sciences, Kumamoto University, Kumamoto, Japan*

YUSUKE NAKAHARA • *Department of Analytical and Biophysical Chemistry, Graduate School of Pharmaceutical Sciences, Kumamoto University, Kumamoto, Japan*

MASANORI NODA • *Graduate School of Engineering, Osaka University, Suita, Japan*

TAKAO NOMURA • *Center for Research and Education on Drug Discovery, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan*

TOMOHIDE SAIO • *Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo, Japan; Department of Structural Biology, Faculty of Advanced Life Science, Hokkaido University, Sapporo, Japan*

CHIKARA SATO • *Biomedical Research Institute and Molecular Profiling Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan*

MASAHIKO SATO • *Biomedical Research Institute and Molecular Profiling Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan*

MIKI SENDA • *Structural Biology Research Center, Photon Factory, Institute of Materials Structure Science, High Energy Accelerator Research Organization (KEK), Tsukuba, Japan*

TOSHIYA SENDA • *Structural Biology Research Center, Photon Factory, Institute of Materials Structure Science, High Energy Accelerator Research Organization (KEK), Tsukuba, Japan*

TATSURO SHIMAMURA • *Faculty of Medicine, Kyoto University, Kyoto, Japan*

KAZUMI SHIMONO • *RIKEN Systems and Structural Biology Center, Yokohama, Japan; Graduate School of Pharmaceutical Sciences, Toho University, Funabashi, Chiba, Japan*

TAKEHIRO SHINODA • *RIKEN Systems and Structural Biology Center, Yokohama, Japan; Division of Structural and Synthetic Biology, RIKEN Center for Life Science Technologies, Yokohama, Japan*

MIKAKO SHIROUZU • *RIKEN Systems and Structural Biology Center, Yokohama, Japan; Division of Structural and Synthetic Biology, RIKEN Center for Life Science Technologies, Yokohama, Japan*

TAKASHI TADOKORO • *Center for Research and Education on Drug Discovery, Faculty of Pharmaceutical Sciences, Hokkaido University, Sapporo, Japan*

TAKAHO TERADA • *RIKEN Systems and Structural Biology Center, Yokohama, Japan; RIKEN Structural Biology Laboratory, Yokohama, Japan*

KOUHEI TSUMOTO • *Department of Bioengineering, School of Engineering, The University of Tokyo, Tokyo, Japan; Medical Proteomics Laboratory, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan*

SUSUMU UCHIYAMA • *Graduate School of Engineering, Osaka University, Suita, Japan*

MIN YAO • *Faculty of Advanced Life Science, Hokkaido University, Sapporo, Japan*

SHIGEYUKI YOKOYAMA • *RIKEN Systems and Structural Biology Center, Yokohama, Japan; RIKEN Structural Biology Laboratory, Yokohama, Japan*

OLIVER B. ZELDIN • *Department of Molecular and Cellular Physiology, Stanford University, Stanford, CA, USA*

Part I

Protein Expression

Chapter 1

Expression in Bacteria and Refolding

Hiroki Akiba and Kouhei Tsumoto

Abstract

Production of proteins by bacterial expression is common due to straightforwardness and inexpensiveness. In this chapter, we focus on the expression system using *Escherichia coli* and refolding of inclusion bodies formed with insoluble protein materials. In the first half, several steps required to produce soluble proteins as much amount as possible, in *E. coli*, are described. Here, the choice of either vector or bacterial strains, induction, extraction, fusion of solubilizing tags, and strategies to facilitate disulfide bond formation are included. In the second half, strategies to get soluble proteins from inclusion bodies are described. Here, the mechanism of protein refolding, the isolation of inclusion bodies, the choice of solubilizing materials, the refolding step, and the effects of additives while refolding are included. The selection of various strategies in these steps is discussed.

Keywords Recombinant protein, *Escherichia coli*, Chaperone, Solubilization, Protein tag, Inclusion body, Denaturant, Detergent, Arginine

1 Introduction

Several bacterial expression systems are known to date, such as *Escherichia coli*, *Bacillus subtilis*, and *Bacillus megaterium* (reviewed in [1]). Here we focus on the methods for recombinant protein production in *E. coli*, including protein refolding. Expression in *E. coli* is one of the most straightforward and inexpensive strategies for production of recombinant proteins. Many plasmid vectors and competent cells of various strains optimized for high-level controlled protein production are commercially available. The production is reproducible and can be easily scaled up. Although expression in *E. coli* is often the "first choice," many considerations should be taken into account for efficient production of soluble recombinant proteins.

Proteins with low solubility or proteins not folded correctly form inclusion bodies (insoluble aggregates) in bacterial cells. In solution, proteins always exist in equilibrium between denatured,

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks, DOI 10.1007/978-4-431-56030-2_1, © Springer Japan 2016

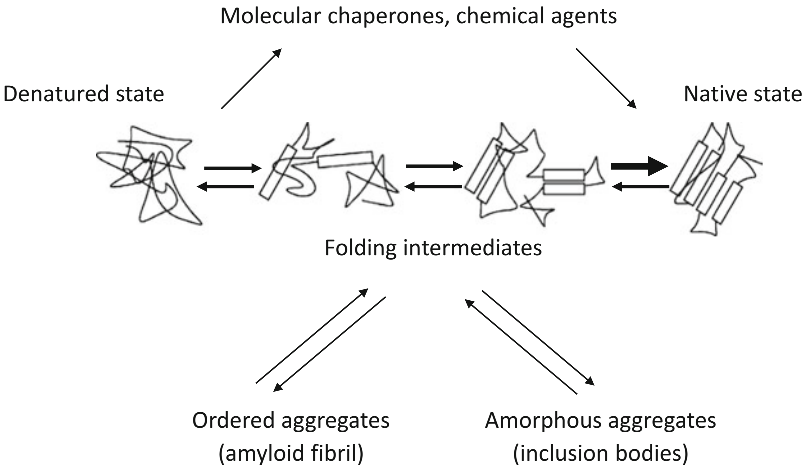


Fig. 1 Protein folding and its regulation. Proteins in solution are in equilibrium between denatured, intermediate, and native states. Folding intermediates are prone to aggregation. Molecular chaperones in vivo and chemical agents in vitro facilitate folding into the native state

native, intermediate, and aggregated states (Fig. 1). The native state is usually the most stable, but if the transition from the intermediate to the native state (bold arrow) is not facilitated, the reaction proceeds toward aggregation and inclusion bodies are formed. In this case, genetic engineering or protein refolding from inclusion bodies is required to shift the equilibrium toward the native state for efficient production of soluble proteins.

Refolding is a common in vitro method to obtain natively folded soluble proteins from inclusion bodies. Briefly, the inclusion bodies are solubilized with a buffer containing a high concentration of denaturant or detergent, and the denatured protein is folded by decreasing the concentration of the solubilizing agents.

This chapter consists of two main sections: expression in *E. coli* (Sect. 2) and refolding (Sect. 3). Both sections are intended as a guide on how to maximize the yield of soluble protein in the native state. In Sect. 2, we discuss expression strategies, including the choice of vectors and *E. coli* strains, which are the key steps for high-level production of soluble proteins. We also discuss periplasmic expression and alternative strains for protein production in the oxidative environment, because most therapeutic proteins are extracellular in origin and require disulfide bond formation in the oxidative environment. In Sect. 3, we summarize various strategies for inclusion body solubilization and protein refolding, and we discuss the effects of additives and other refolding strategies.

2 Expression in *E. coli*

2.1 Expression Strategies

Expression in *E. coli* includes four steps (Fig. 2): (1) the choice and construction of the expression vector, (2) its transformation into an appropriate *E. coli* strain, (3) induction of expression, and (4) extraction of the recombinant protein from either the medium or the lysate. Although the high-yield production of many proteins can be achieved by using generic strategies, optimization of each step may be necessary, especially for proteins from other organisms [2].

2.1.1 The Choice of the Expression Vector

This is one of the most important steps, which determines the protein yield and fate. Several parameters are crucial for the outcome, including (1) the origin of replication, (2) the promoter system, and (3) signal sequences or tag fusion sequences.

1. The plasmid copy number per cell depends on the origin of replication (*ori*) and may vary from low (2–15) to high (>15) [3]. The ColE1 *ori* ensures high expression levels. One of the most common plasmids with high copy number is pET; it has the pBR322 *ori* (similar to the ColE1 *ori*). Low-copy-number plasmids are used to reduce the expression during bacterial growth, because overexpression imposes a metabolic burden by using proliferation and survival resources. Growth inhibition may be particularly severe when the protein of interest is toxic.

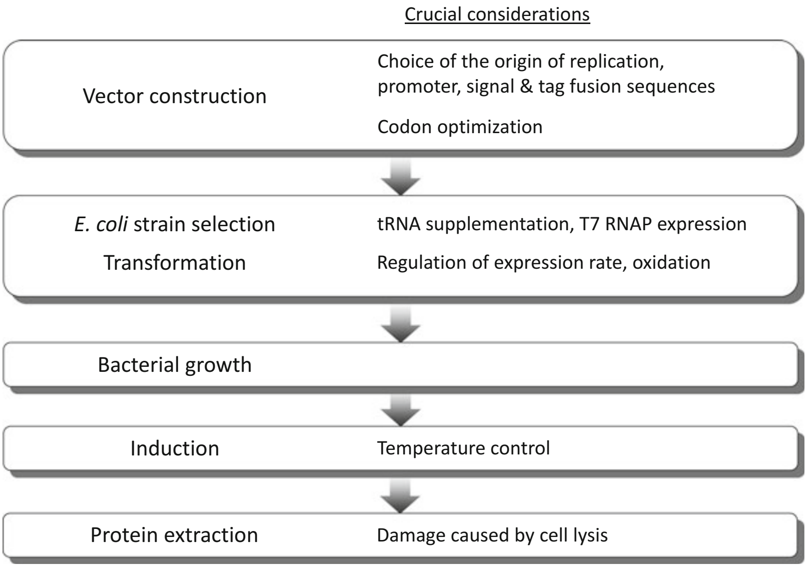

Crucial considerations
Vector construction
Choice of the origin of replication, promoter, signal & tag fusion sequences
Codon optimization
E. coli strain selection
tRNA supplementation, T7 RNAP expression
Transformation
Regulation of expression rate, oxidation
Bacterial growth
Induction
Temperature control
Protein extraction
Damage caused by cell lysis

Fig. 2 Experimental procedures and considerations for soluble protein production in *E. coli*

2. Inducible promoters allow the suppression of recombinant protein production in the early log phase of bacterial growth. Among inducible promoter cassettes in commercially available vectors, the T7 promoter system is the most popular. In this system, the mRNA for T7 RNA polymerase (T7 RNAP) is transcribed from the *lacUV5* promoter, which is activated by isopropyl β-D-thiogalactopyranoside, or generally abbreviated as IPTG [4], and the newly synthesized polymerase transcribes the gene under the control of the T7 promoter. This double regulation mechanism enables high levels of IPTG-induced expression. Other inducible promoters directly control the expression of genes of interest. The *araBAD* promoter is induced by arabinose. In comparison with the T7 promoter, the *araBAD* promoter allows dose-dependent expression upon induction and lower background in the absence of induction [3]. Some promoters can be activated by temperature changes. Cold-shock promoters (such as *cspA*) are induced by cooling below 15 °C and drive low-level expression [5].
3. Many vectors allow fusing a signal peptide or a tag sequence to either the N- or the C-terminus of the recombinant protein. In *E. coli*, N-terminal signal peptides of membrane proteins and secreted proteins are required for translocation into the periplasm (see Sect. 2.3). Peptide tags, such as polyhistidine (His-tag) and GST (glutathione S-transferase), are usually fused to the protein of interest for purification. Protein tags also increase recombinant protein solubility and yield (see Sect. 2.2).

Reduced protein levels may result from poor translation of several codons that lack corresponding tRNAs in *E. coli* (referred to as rare codons: see Table 1) [1]. In these cases, codon optimization may be required for efficient production of recombinant proteins.

2.1.2 The Choice of E. coli Strain

Host cells are engineered to maximize the yield of recombinant proteins [6]. Frequently used strains mentioned in this section are summarized in Table 2. The choice of the strain depends on the protein of interest and the vector. One of the most common strains is BL21, which is deficient in two cytoplasmic proteases. Note that hosts with the T7 RNA polymerase-encoding gene DE3, such as BL21(DE3), must be used with the T7 promoter system [4].

tRNA-supplemented strains are effective for the production of heterologous (e.g., mammalian) recombinant proteins without codon optimization [7]. BL21-CodonPlus and Rosetta 2 are BL21-derived strains designed for this purpose, with multiple supplemented tRNA genes for rare codons (Table 1).

Table 1
Rare codons in *E. coli*

Amino acid	Codon	*E. coli* K12	*E. coli* B	BL21-CodonPlus	Rosetta 2
Arg	AGG	1.6	2.1	*	*
Arg	AGA	1.4	2.4	*	*
Arg	CGA	4.3	2.4		
Arg	CGG	4.1	5.0		*
Pro	CCC	6.4	2.4	*(RP, RIPL)	*
Leu	CUA	5.3	3.0	*(RIL, RIPL)	*
Ile	AUA	3.7	5.0	*(RIL, RIPL)	*
Gly	GGA	9.2	8.2		*

Codon usage in *E. coli* strains is presented as frequencies per 1000 codons. BL21 and Rosetta 2 are derived from *E. coli* B strain. *Codons recognized by supplemented tRNA in BL21-CodonPlus and Rosetta 2. BL21-CodonPlus includes three strains: RIL (for AT-rich sequences), RP (for GC-rich sequences), and RIPL (for both). Codon usage data are from http://www.kazusa.or.jp/codon/

Table 2
Summary of *E. coli* strains described in this chapter

E. coli strain	Characteristics
BL21	Deficient in the *lon* and *ompT* protease genes
BL21-CodonPlus	tRNA supplemented to BL21
Rosetta 2	tRNA supplemented to BL21
BL21(DE3)pLysS/E	T7 RNAP inhibited by T7 lysozyme
C41, C43	BL21 derivatives for expression of membrane proteins
Lemo21(DE3)	Dose-dependent control of T7 lysozyme production
Origami 2	K12-derived strain for disulfide bond formation in the cytoplasm
Origami B	BL21-derived strain for disulfide bond formation in the cytoplasm
Rosetta-gami	tRNA supplemented to Origami strains
SHuffle	K12-derived strain for disulfide bond formation in the cytoplasm

Regulation of protein yield is particularly important if the protein of interest is toxic or if fast protein production does not allow proper folding or translocation. BL21(DE3)pLysS and BL21 (DE3)pLysE strains encode T7 lysozyme, which inhibits T7 RNAP and suppresses the expression of the gene of interest from the T7 promoter in the absence of induction [8]. The production of potentially harmful recombinant proteins is prevented and thus promotes efficient growth of the host before entering the log phase and eventually higher yield of recombinant protein. The C41 (DE3) and C43(DE3) strains, derived from BL21(DE3), are optimized for production of membrane proteins [9]. Even after induction, these strains produce proteins slower than BL21(DE3), which prevents cell lysis and increases protein yield [10]. In the recently developed Lemo21(DE3) strain, protein production can be controlled by the addition of rhamnose [10, 11]. This strain encodes T7 lysozyme (LysY) under the control of the *rhaBAD* promoter. Dose-dependent inhibition of T7 RNAP allows easy control and optimization of production of target proteins. Lemo21(DE3) is advantageous for production of toxic or membrane proteins and for protein accumulation in the periplasm.

2.1.3 Induction

Usually, expression is induced in the log phase of the bacterial culture (O.D. 600 = 0.4–0.6). Earlier induction may inhibit cell division (and thus the number of cells able to produce the recombinant protein is not sufficient), whereas later induction may decrease protein production because of the lack of resources. *E. coli* is cultured at 25–37 °C. Higher temperatures may increase the background expression; this is particularly undesirable if the protein is toxic. Lowering the temperature may be beneficial, especially after induction. A temperature below 30 °C allows sufficient time for protein folding and prevents unfolding followed by heat-shock protein-dependent proteolysis. Cultivation even below 10 °C may result in high protein yield [12].

2.1.4 Recombinant Protein Extraction

Secreted proteins with an affinity tag can be collected by affinity chromatography or salting out with ammonium sulfate, whereas proteins that are not secreted are collected upon cell disruption. Sonication is one of the most popular methods to disrupt bacterial cells; however, heat and air bubbles produced during sonication may lead to protein damage, in particular for unstable and membrane proteins. In such cases, French pressure cell press or other more gentle methods (e.g., the use of detergents) may be useful. Soluble proteins are prone to degradation by proteases; thus, the use of protease inhibitors or EDTA (ethylenediaminetetraacetic acid) is recommended to avoid proteolysis.

2.2 Increasing the Yield of Soluble Proteins

As discussed above, the choices of vectors, *E. coli* strains, and cultivation temperature affect the yield of soluble proteins (Fig. 2). For many difficult-to-express proteins, slower expression results in higher yield. Oxidative conditions promote proper folding of disulfide-rich proteins (see below). Some engineering strategies are possible.

Tag fusion (mainly N-terminal, but sometimes C-terminal) often greatly increases the amount of soluble protein. GST, NusA (N-utilizing substance A), MBP (maltose-binding protein), Trx (thioredoxin), and SUMO (small ubiquitin-like modifier) are major fusion tags used for this purpose. Many vectors designed for soluble tag fusion are commercially available. Solubility has been studied and screened with many tags [13–16]. Each fusion tag has advantages and disadvantages, and the tag should be selected according to the objectives [17]. The well-studied MBP and GST tags enhance the solubility of many proteins and also enable subsequent purification. For these two tags, GST tags are useful because the affinity against the ligand (glutathione or GSH) is sufficiently strong. However, the native property of GST to dimerize is sometimes problematic, and fused GST should be removed in case when this perturbs the subsequent uses. The SUMO tag, a small N-terminal tag, can be completely cleaved with SUMO proteases [17] to restore the original N-terminus.

Co-overexpression with molecular chaperones (enzymes that help appropriate protein folding) is another strategy to enhance the solubility of the protein of interest [18, 19]. The DnaK-DnaJ-GrpE chaperone system assists folding, whereas the GroEL-GroES system refolds misfolded proteins in the presence of ATP (Fig. 3a). Although the endogenous molecular chaperones function in bacteria, the expression level induced by the T7 promoter often overwhelms these enzymes. In such cases, chaperone overexpression may overcome this problem. Multi-domain proteins (such as cytosolic kinases), produced in the cells containing overexpressed DnaK-DnaJ-GrpE or GroEL-GroES, are soluble [18–20]. Vectors for co-expression with chaperones and suitable BL21-derived strains are commercially available.

2.3 Expression Under Oxidative Conditions

Bacterial GSH and Trx systems maintain a reducing environment in the bacterial cytoplasm [21], thus limiting the formation of disulfide bonds. Heterologous secreted proteins with intramolecular disulfide bonds, such as antibodies or growth factors, often fail to fold properly in the bacterial cytoplasm. Translocation of such proteins into the periplasm is desirable for appropriate disulfide bond formation. The periplasm, the region between bacterial inner and outer membranes, contains molecular chaperones (including disulfide bond isomerases DsbA and DsbC; Fig. 3b) suitable for disulfide bond-rich proteins and provides oxidative conditions advantageous for protein folding [21, 22]. In this respect, the periplasm is similar to the endoplasmic reticulum of the higher organisms.

Periplasmic translocation of recombinant proteins requires an N-terminal signal peptide and is performed by several systems. The most popular is the SecYEG machinery (Fig. 3b) [23]. The length of signal peptides recognized by SecYEG varies from 5 to 40 amino acids. Such sequences contain positive charges at the N-terminus, followed by a hydrophobic core and a less hydrophobic C-terminal region. Apart from being hydrophobic, signal peptides

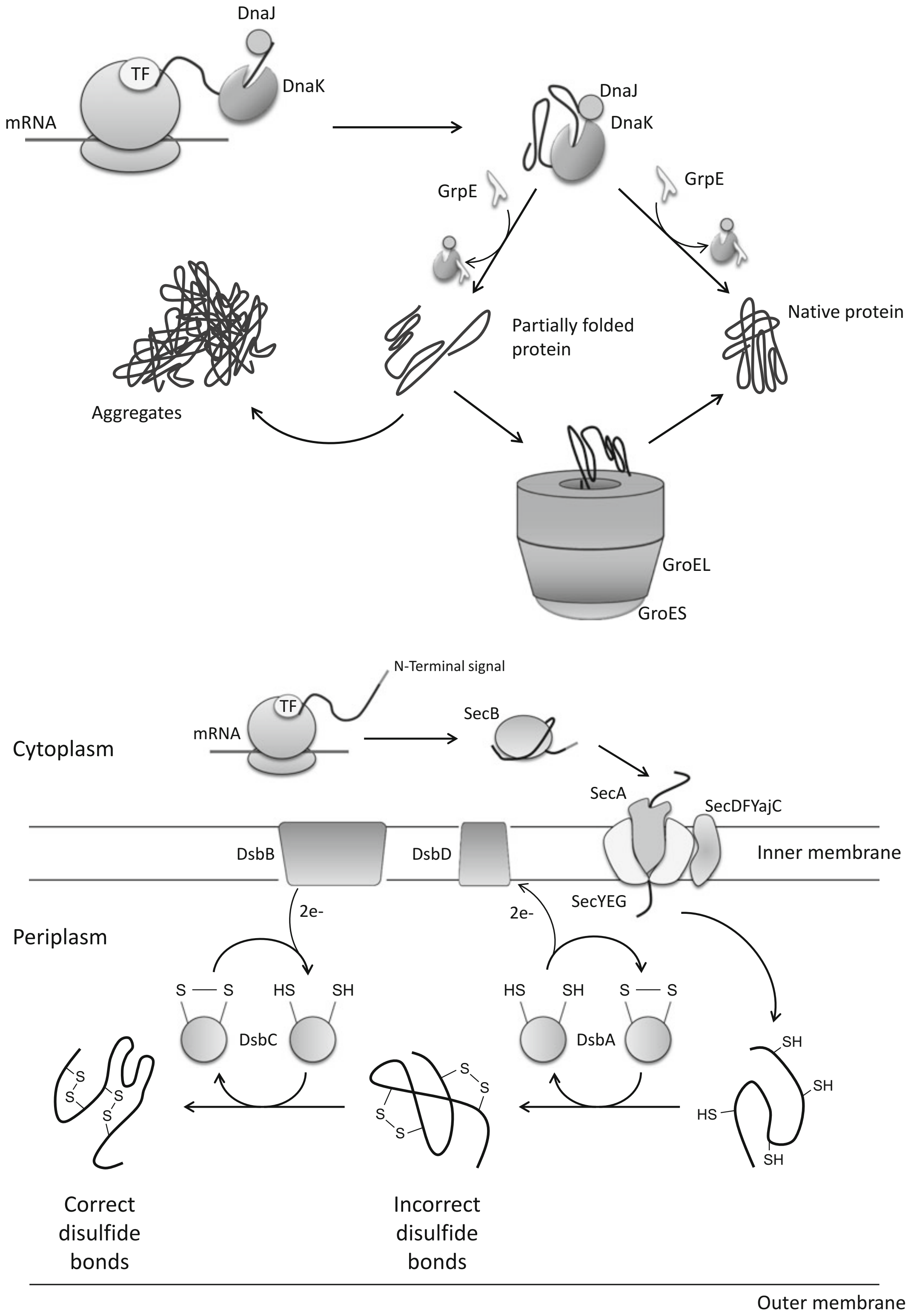


Fig. 3 (**a**) Molecular chaperones for protein folding in bacteria. Trigger factor (TF) and the DnaK-DnaJ-GrpE system facilitate folding of de novo synthesized proteins, whereas the GroEL-GroES system refolds partially folded intermediates. (**b**) Translocation of synthesized proteins to the periplasm. Synthesized protein with an

display little sequence similarity to each other. Signal peptides used in recombinant proteins are derived from bacterial periplasmic or membrane proteins or from proteins secreted into the medium, for example, *pelB*, *phoA*, *dsbA*, and *ompA* [22]. A recent study suggests that codon usage biased toward rare codons promotes periplasmic accumulation [24]. The periplasm contains lower amounts of proteases than the cytoplasm, and periplasmic proteins can be extracted by osmotic shock, a mild method that damages the bacterial cell wall and allows isolation of proteins of high purity.

Strains that enable accumulation of disulfide-rich proteins in the cytoplasm are available [21, 22]. In Origami and its derivatives (including tRNA-supplemented Rosetta-gami), thioredoxin reductase and glutathione reductase are mutated, resulting in a suppression of reducing agents in the cytoplasm [25]. This facilitates disulfide bond formation in the cytoplasmic proteins. The SHuffle strain additionally produces the disulfide bond isomerase DsbC in the cytoplasm [26]. Thus, in the latter two strains, disulfide-rich proteins can fold properly in the cytoplasm.

3 Refolding

3.1 Mechanism of Refolding

Refolding is a method by which a properly folded protein is obtained by solubilization and denaturation of an incorrectly folded protein, followed by folding into the native or native-like states (Fig. 1).

In many cases, the level of soluble protein is very low even though the protein is produced at a high level. Many recombinant proteins produced in bacteria do not fold correctly because of the insufficient level of chaperones, too fast production, or the absence of post-translational modifications [27], and these proteins form insoluble, amorphous aggregates (inclusion bodies, Fig. 1). These aggregates are different from ordered aggregates, or amyloid fibrils, in that the aggregated state is energetically disfavored in nature. As recombinant proteins are the main components of the inclusion bodies and are free from cytoplasmic proteases, proteins with high purity can be isolated from inclusion bodies. However, such proteins need to be refolded.

Each protein folds into an energetically favored three-dimensional structure, which is determined by its primary structure (Anfinsen's dogma) [28]. Proper folding gives −10 kcal/mol of stability (marginal stability), gained by the formation of hydrogen

◄

Fig. 3 (continued) N-terminal signal peptide is translocated by the transmembrane SecYEG system with the help of SecB and SecA. After signal peptide cleavage, the protein is folded into the native state with the help of periplasmic protein disulfide isomerases DsbA and DsbC, which facilitate correct disulfide bond formation with the help of two membrane proteins, DsbB and DsbD bonds and salt bridges in the hydrophobic core inside the protein molecule [29]. From the viewpoint of thermodynamics, marginal stability is the product of a balance between the enthalpy change upon non-covalent bond formation and the entropy change. This energetically favored event can be artificially controlled through solvents by choosing appropriate buffer for solubilization and refolding [30–32].

Folding is the equilibrium between the native structure and nonnative structure. This equilibrium depends strongly on the characteristics of the intermediate state, and thus the optimal refolding method for each protein is differed. Aggregation occurs mainly because of the interaction between partly folded or denatured protein molecules, and such interaction should be interrupted during refolding [33].

The structure of protein molecules in the inclusion bodies and the mechanism of inclusion body formation have been discussed in detail [34–37]. These studies discovered that proteins in the inclusion bodies are in equilibrium with those in the solution. Spectroscopic characteristics of the inclusion bodies of single-chain Fv (scFv), green fluorescent protein (GFP), β2-microglobulin, and hyperthermophilic proteins suggested that secondary structures of α-helix-rich hyperthermophilic proteins contained high levels of native-like structure, while the level was lower for the others [35]. Therefore, the choice between complete denaturation and maintaining the already existing secondary structure should be considered during protein extraction.

3.2 Recovery from Inclusion Bodies

The procedure includes three steps: (1) isolation of inclusion bodies, (2) solubilization, and (3) protein refolding (Fig. 4).

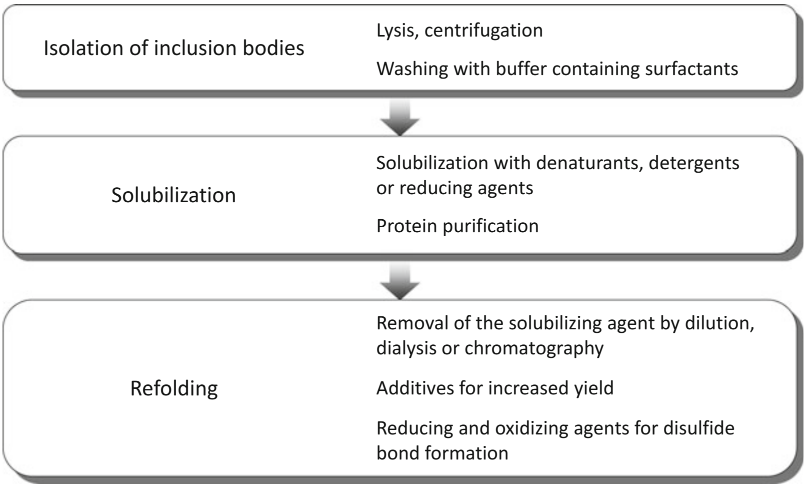

Isolation of inclusion bodies
Lysis, centrifugation
Washing with buffer containing surfactants
Solubilization
Solubilization with denaturants, detergents or reducing agents
Protein purification
Refolding
Removal of the solubilizing agent by dilution, dialysis or chromatography
Additives for increased yield
Reducing and oxidizing agents for disulfide bond formation

Fig. 4 Experimental procedures in refolding

3.2.1 Isolation of Inclusion Bodies

Inclusion bodies have high density and can be collected by relatively slow centrifugation (e.g., 6000 × *g*, 30 min). When the recombinant protein is mainly present in inclusion bodies and its solubility is very low, contaminants can be removed by washing the inclusion bodies with a surfactant (e.g., 1–4 % Triton-X 100), followed by surfactant removal before solubilization. If the protein is present in both soluble and insoluble fractions, the aggregates are relatively loose and soluble protein can be extracted from the inclusion bodies with L-arginine (vide infra).

3.2.2 Solubilization

Protein aggregates can be solubilized by a denaturant or detergent. Denaturants, or chaotropic agents, such as guanidine hydrochloride (GdnHCl) or urea, are most commonly used. These agents interact with proteins in a concentration-dependent manner and form hydrogen bonds and salt bridges, resulting in a disordered structure. A high concentration of GdnHCl (6 M) or urea (8 M) is necessary for complete denaturation, and prolonged incubation (from several hours to overnight) is required for the solution to become homogeneous. Affinity chromatography such as His-tag purification can be subsequently performed even in the presence of denaturants.

Detergents such as sodium dodecyl sulfate (SDS), cetyltrimethylammonium bromide (CTAB), or sodium lauroyl sarcosinate (sarkosyl) are used above their critical micelle concentrations for protein solubilization, and the partially ordered structure of the extracted proteins is preserved by their interaction with micelles. The detergents may be difficult to remove completely, which is a major drawback for full recovery of protein activity. Recently, a detergent lauroyl-L-glutamate (C12-Glu) was found to be highly efficient in solubilizing inclusion bodies and to be easily removed from the protein after refolding by dilution [38]. Refolding was achieved for various proteins, in particular those consisting of multiple domains, by optimizing C12-Glu concentration, the dilution procedure, additives, and temperature.

Reducing agents such as 2-mercaptoethanol or dithiothreitol would be used in this solubilizing step for cleaving incorrectly formed disulfide bonds and may help folding into the native state subsequently. Formation of disulfide bonds is necessary during refolding.

We have reported that soluble proteins with native or native-like structure could be extracted from insoluble fractions by using arginine [39, 40]. Arginine is not a chaotropic agent and cannot denature proteins, but it successfully extracts protein molecules with native-like structures, particularly when the protein is present in both soluble and insoluble fractions.

3.2.3 Refolding

Refolding is achieved by removing denaturants or detergents from the protein solution by dilution or dialysis. During the removal of these agents, equilibrium between correct folding and aggregation should be controlled by choosing appropriate methodologies [40].

Dilution is the simplest and the most frequently applied methodology. One of the four possible dilution strategies is used (Fig. 5a–c). In the conventional dilution method, the denatured protein is injected into the refolding buffer (Fig. 5a). In the reverse dilution method, refolding buffer is added to the solution of denatured protein. Dilution can also be performed by mixing the two solutions in a constant rate with an automated mixer (Fig. 5b). Time-dependent changes in denaturant or detergent concentrations encountered by the protein in the final solution are different in these methods (Fig. 5c). In the conventional dilution method (bold solid line), the concentration rapidly increases

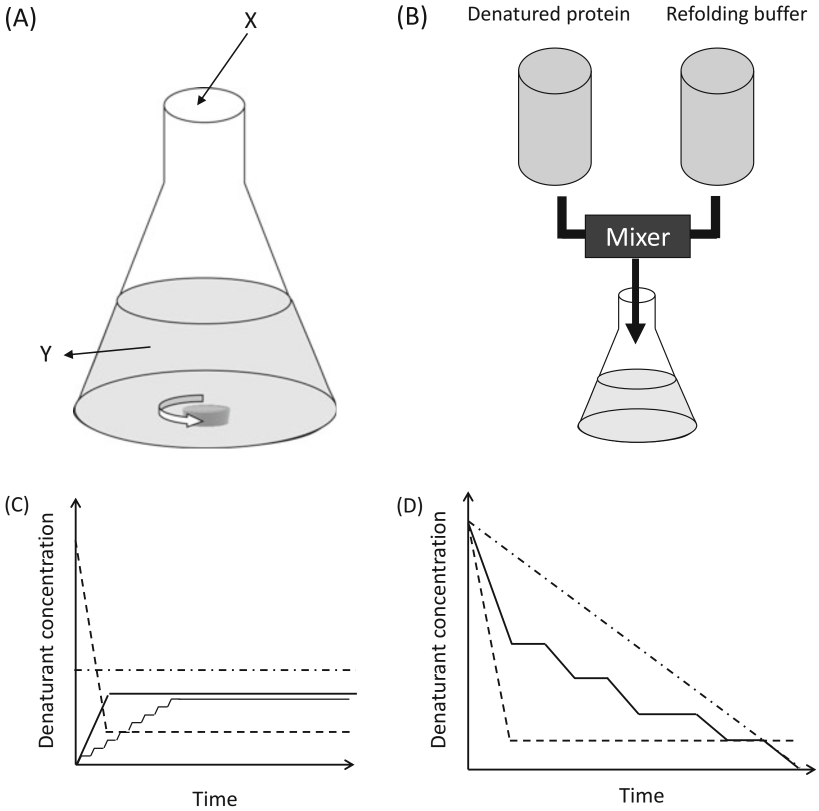


Fig. 5 Refolding strategies. (**a**) Conventional dilution. X, protein in a denaturant solution; Y, dilution buffer. In reverse dilution, X would be dilution buffer and Y would be protein. Pulsed dilution is similar to the conventional dilution except that protein is added in pulses with intervals. (**b**) Dilution with the use of a mixer. (**c**) Changes in the denaturant concentration faced by the protein in the final solution in different dilution methods. *Bold solid line*, conventional dilution; *dashed line*, reverse dilution; *dot-and-dash line*, dilution with a mixer; *thin solid line*, pulsed dilution. (**d**) Reduction in the denaturant concentration by dialysis. *Solid line*, stepwise dialysis; *dashed line*, single dialysis; *dot-and-dash line*, continuous dialysis. For details, see the main text

from zero to the final concentration, whereas reverse dilution allows high concentration at first (dashed line). When a mixer is used, the concentration remains constant (dot-and-dash line). These differences affect the solubility and stability of the folding intermediate. For example, conventional dilution is not recommended if the low concentration of denaturant at the first contact of the unfolded protein with the dilution buffer leads to instability. In such cases, addition of a relatively low concentration of denaturant (e.g., 0.5–1 M GdnHCl or 1–2 M urea) to the buffer may be useful.

Pulsed dilution is a method in which multiple aliquots of the protein in denaturant are diluted consecutively [32, 41]. A small aliquot is first diluted with the refolding buffer. After allowing some time for folding, the next aliquot is added and so on (Fig. 5a; thin solid line in Fig. 5c). Pulsed addition of denaturant facilitates aggregate solubilization in the final solution, since aggregates with hydrophobic surface are more sensitive to denaturants than natively folded proteins. Repetitive pulses of denaturant and dilution enable efficient refolding with high yield.

Dialysis is frequently used as well. Because of the relatively slow solute exchange between the inside and outside of the dialysis bag, the change in the environment around protein molecules is slow, which facilitates refolding at intermediate denaturant concentrations. Using diluted protein solution (≤7.5 μM) is recommended. Changes in denaturant concentration in different dialysis strategies are presented in Fig. 5d. In the single dialysis method, a dialysis bag with the denatured protein is placed into the final refolding buffer (dashed line). In continuous dialysis, the denaturant concentration outside the dialysis bag is continuously lowered using a pump (dot-and-dash line). Stepwise dialysis starts from a high concentration of denaturant in the refolding buffer, and the concentration is lowered in several steps, with stops at intermediate denaturant concentrations (solid line).

In stepwise dialysis, timing of the addition of additives can be controlled depending on the protein characteristics, in particular the rate of folding and disulfide bond formation. As disulfide bonds cannot form after, the native-like structure has been reached, disulfide bond formation should be controlled at the intermediate steps [42]. A typical example of the effectiveness of stepwise dialysis is refolding of scFv through appropriate control of additive concentrations in the intermediate state [43]. The inclusion bodies were first denatured by 6 M GdnHCl, followed by stepwise dialysis. After the GdnHCl concentration reached 1 M, 0.4 M L-arginine and 1 mM glutathione disulfide (GSSG) were added. This procedure yielded more than 80 % of folded protein [43]. This example illustrates the importance of the control of aggregation and oxidation during intermediate steps of refolding.

The choice between dilution and dialysis, and the exact procedure within each method, should aim at optimizing the denaturant concentration for initiation of refolding and disulfide bond formation and maintaining the stability of intermediates. The optimized ones are screened by several trials. When dialysis enables high yield for proteins, it is indicated that their folding is slow and the intermediate is relatively stable. When dilution is superior, the intermediate is unstable and the protein easily aggregates. The choice between these methods, selection of the concentration of the denatured protein, and the volume ratio of the protein to the refolding buffer are important for successful refolding.

Refolding on a resin, or the stationary phase of chromatography, is an alternative approach reported by many groups [44–49]. Denatured proteins are first captured on a resin (equilibrated with either denaturant or refolding buffer). Immobilized metal affinity chromatography, ion-exchange chromatography, hydrophobic exchange chromatography, or size-exclusion chromatography can be used for this purpose. Denaturants are removed by elution with refolding buffer. Because each protein molecule is captured separately, interactions between folding intermediates and denatured proteins are minimized, and aggregation is therefore suppressed. However, interaction of multiple molecules on the resin is not always prevented and thus inhibits refolding. When steric hindrance by the resin occurs or residues crucial for refolding are associated in the interaction with the resin, a weaker interaction between the resin and protein molecules may be beneficial, for example, changing pH or ionic strength in case of ion-exchange chromatography.

3.3 Small-Molecule Additives

Small-molecule additives are used to stabilize the native structure, destabilize misfolded proteins, and enhance the solubility of folding intermediates or denatured proteins. Additives are roughly divided into two groups: folding enhancers and aggregation suppressors (Fig. 6 and Table 3) [40]. Folding enhancers, such as sugars, polyols, ammonium sulfate, glycine, and alanine, stabilize the native structure of the protein without directly interacting with it; however, they also enhance protein aggregation. Polyethylene glycol (PEG), cyclodextrin, proline, and arginine form hydrophobic or polar interactions with proteins and suppress their aggregation but do not promote folding. Chaotropic agents such as GdnHCl and urea can be used as additives at reduced concentrations. Ionic liquids are also applied [50]. Mild detergents such as lauryl maltoside (*n*-dodecyl-β-D-maltopyranoside) have been reported to increase the yield of active proteins after refolding [51].

Among small-molecule additives, L-arginine is the most widely used (Fig. 7). Similar to GdnHCl, arginine has a guanidinium group, but instead of denaturing proteins, it enhances solubility of folding intermediates [52]. Our research suggests that the

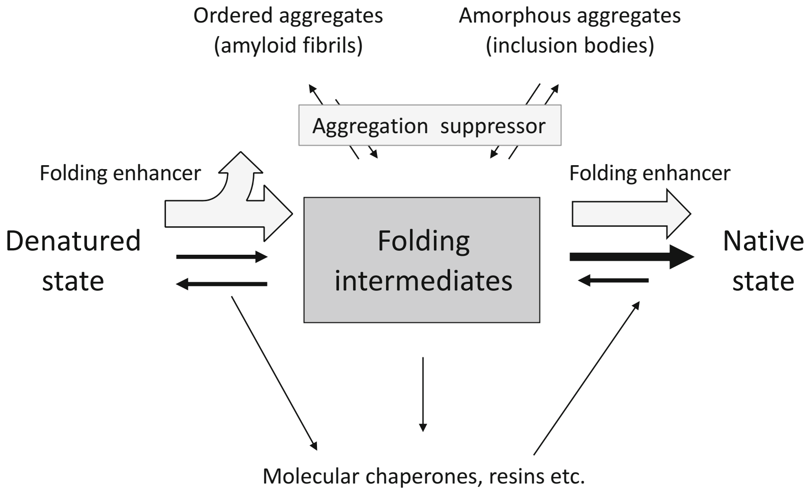


Fig. 6 Folding enhancers and aggregation suppressors. Folding enhancers facilitate intra- and intermolecular interactions and thus promote both folding into the native structure and aggregation. Aggregation suppressors only suppress aggregation and do not actively promote folding

Table 3
Classification of chemical agents affecting protein stability and protein–-protein interactions

Classification	Examples	Effect on protein stability	Effect on protein–protein interactions
Folding enhancers	Sugars (sucrose, glucose, etc.)	Stabilization	Enhancement
	Ammonium sulfate		
	Polyols (glycerol, sorbitol, etc.)		
	Glycine, alanine, serine		
Aggregation suppressors	Arginine, NDSBs[a]	Neutral	Reduction
	Proline		
	Cyclodextrin, PEG[b]		
	Mild detergents		
Denaturing agents	Urea, GdnHCl[c]	Destabilization	Reduction
	Strong detergents		

[a] Non-detergent sulfobetaines
[b] Polyethylene glycol
[c] Guanidine hydrochloride

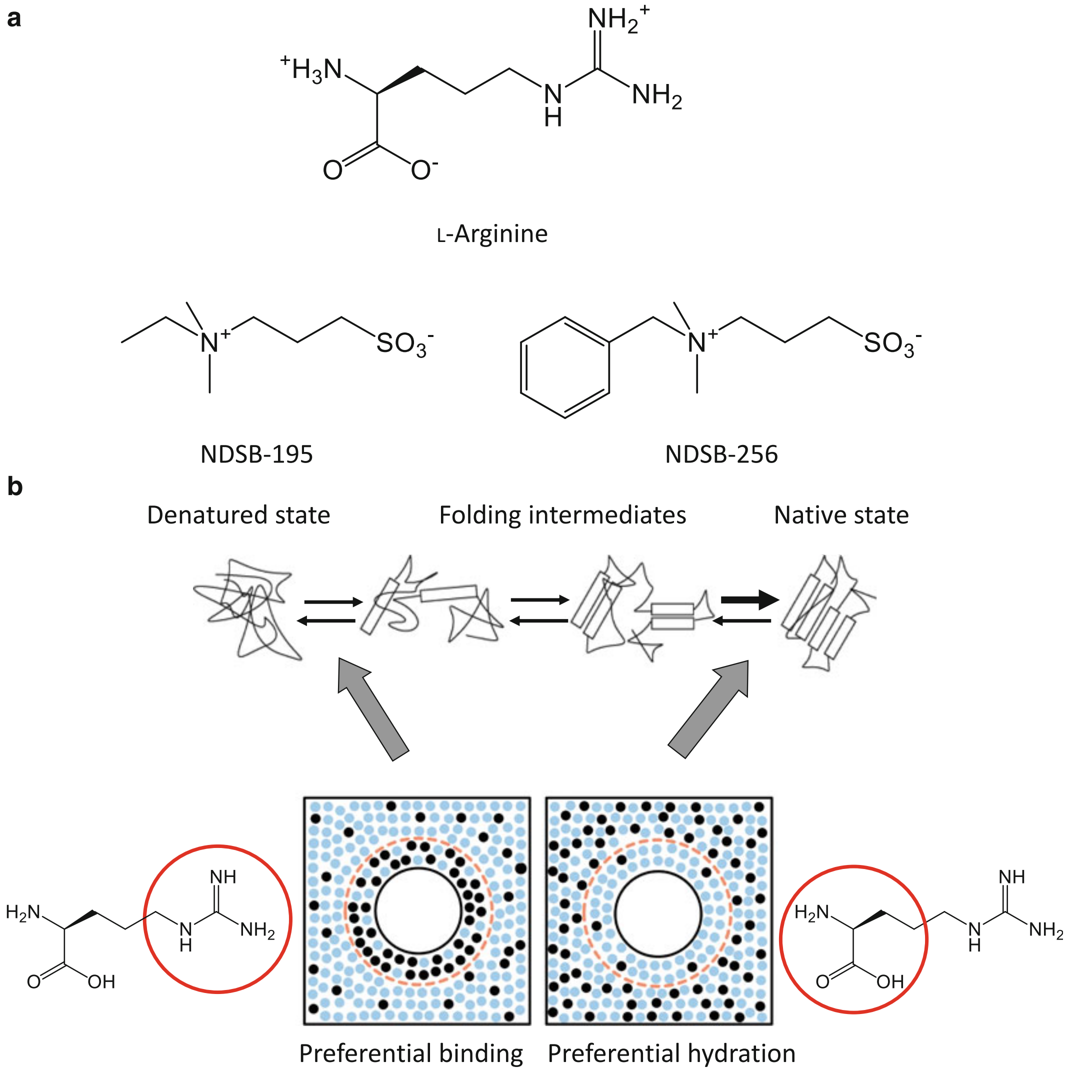


Fig. 7 Arginine and NDSBs. (**a**) Structure of L-arginine and NDSBs. (**b**) Arginine has two functional groups, the guanidinium ion and the α-amino acid moiety. The two groups act cooperatively by different mechanisms depending on the folding state of the protein. The guanidinium group preferentially binds to the denatured protein (*arrow* on the *left-hand side*), whereas the α-amino acid moiety assists preferential hydration of the protein in native(-like) states to enhance folding (*arrow* on the *right-hand side*). In the *lower* figure, *white circle* represents a protein, *red dashed line* represents its hydration shell, *blue dots* represent water molecules, and *black dots* represent arginine molecules

guanidinium group suppresses aggregation, whereas the α-amino acid moiety enhances folding by maintaining the appropriate hydration state (Fig. 7b). Cooperation of these two groups results in a chaperone-like function that is different from the simple effect of solvents and cosolutes [52]. Arginine is added at a relatively high concentration. In our studies, arginine is used in 0.4 M. A similar mechanism has been suggested for non-detergent zwitterionic sulfobetaines (NDSBs) [53], composed of a hydrophilic sulfobetaine moiety and a small hydrophobic group, such as phenyl or ethyl. NDSBs are aggregation suppressors also used at high concentrations (ca. 0.5 M).

Other common additives are chelators and redox-controlling agents. Chelators such as EDTA prevent peptide bond cleavage by contaminating proteases and metal-catalyzed oxidation of sulfur atoms in cysteine and methionine. Redox control using a pair of reducing and oxidizing agents, such as GSH and GSSG, is crucial for appropriate disulfide bond formation in cysteine-containing proteins. pH and the GSH to GSSG ratio are the major factors that control accurate disulfide bond formation. An excess of GSSG over cysteine side chains and alkaline pH would lead to the formation of incorrect disulfide bonds. An excess of GSH over GSSG is desirable (up to GSH:GSSG = 10:1), with the GSH concentration of 5–25 mM [54]. The optimal concentrations of GSH and GSSG depend on the amount of the protein of interest.

3.4 Molecular Chaperones and Alternative Refolding Strategies

Molecular chaperones (e.g., GroEL-GroES; see Fig. 3 [20]) help proper folding and function as protein additives. Immobilized enzymes are studied as alternatives [45, 55–58]. Protein proline *cis/trans* isomerase, protein disulfide isomerases DsbA and DsbC, and GroEL-GroES have been immobilized and used to aid on-column refolding. However, enzymes are more expensive than small molecules.

Artificial systems mimicking molecular chaperones have received much attention. A denatured protein is first refolded in a buffer containing detergents, so that aggregation is prevented and a native-like structure is formed. Next, the detergents are removed with cyclodextrin or polymers (such as PEG). This stepwise refolding system mimics the function of folding chaperones in vivo [50, 51, 59, 60]. Nanogels and zeolites can be used as substitutes for detergents [61, 62]. In another system, an insoluble protein was extracted with reversed micelles and refolded in an organic solution [63]. Further development of these new strategies would expand the applicability of refolding to diverse proteins.

Recently, refolding under high hydrostatic pressure (HHP) has been reported [64]. HHP (100–200 MPa) fosters high recovery of native proteins from aggregates in many cases, even at high protein concentrations. Under appropriate pressure, undesirable hydrophobic and electrostatic interactions are suppressed, so that misfolded intermediates are destabilized and native protein conformation is favored. Cosolutes such as chaotropic agents (GdnHCl, urea), arginine, and osmolytes (sugars and polyols) assist effective refolding under HHP [65]. Further studies are needed to overcome challenges, in particular application of HHP to refold multi-domain proteins.

3.5 Selection of Strategies

There is no "golden standard" refolding strategy, and the best choice depends on the characteristics of the protein, especially of its folding intermediate. The folding rate is positively correlated with the chain length and the contact order. Thus, proteins with a β-sheet-based structure generally fold more slowly than α-helix-rich proteins; the former are also more prone to aggregation during the intermediate stage [66], which is characterized by the formation of secondary structure, disulfide bonds, and macromolecular complexes and which varies among proteins. In contrast, the early stage of folding (collapse of a small part) and the late stage (packing of the side chains) are similar in all proteins [67].

The use of L-arginine or NDSB as an additive and the use of C12-Glu as a solubilizing agent enable robust screening for the optimal refolding conditions. The efficiency of these agents may stem from the mechanism of their action during refolding, in particular their effects at the intermediate stage. Many tools are available for selection of refolding methods [41], including an open-access database of refolding strategies (http://refold.med.monash.edu.au/) [68] and commercial high-throughput refolding screening kits [69].

The effectiveness of novel refolding methods and reagents is analyzed mainly by spectroscopic methods such as light scattering [69, 70]. The analyses are conducted by refolding small spherical proteins (such as lysozyme, carbonic anhydrase, citric acid synthase, or luciferase) which are known to be refolded successfully in conventional strategies. However, the newly developed methods are not always applicable to other proteins, especially to large or multidomain proteins. Therefore, the refolding methods should be selected carefully in each case.

4 Summary

Both expression in *E. coli* and refolding require careful selection of strategies. The vector and the *E. coli* strain are key factors for efficient expression of soluble recombinant proteins. More challenging proteins, for example, toxic and membrane proteins, as well as drug targets and therapeutic proteins, are gaining much attention. Combining an appropriate vector and an engineered *E. coli* strain that enables slow expression is effective in many cases. Strain engineering is ongoing, and further studies will enable easy and high-yield production of various proteins.

Refolding is crucial for obtaining functional proteins from inclusion bodies produced in bacteria. Optimization of the refolding strategy is often difficult because of the large number of parameters to be controlled. Recent development of additives such as arginine and NDSB has facilitated this task. High-throughput screening systems and a database helpful in selection of the first step in optimization are also available. Further studies on the agents used in refolding and better understanding of the mechanisms involved will lead to the development of novel and more efficient agents and strategies for refolding.

References

1. Terpe K (2006) Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biotechnol 72:211–222
2. Zerbs S, Frank AM, Collart FR (2009) Bacterial systems for production of heterologous proteins. Methods Enzymol 463:149–168
3. Sørensen HP, Mortensen KK (2005) Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. J Biotechnol 115:113–128
4. Studier FW, Moffatt BA (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol 189:113–130
5. Correa A, Oppezzo P (2011) Tuning different expression parameters to achieve soluble recombinant proteins in *E. coli*: advantages of high-throughput screening. Biotechnol J 6:715–730
6. Waegeman H, Soetaert W (2011) Increasing recombinant protein production in *Escherichia coli* through metabolic and genetic engineering. J Ind Microbiol Biotechnol 38:1891–1910
7. Baca AM, Hol WG (2000) Overcoming codon bias: a method for high-level overexpression of plasmodium and other AT-rich parasite genes in *Escherichia coli*. Int J Parasitol 30:113–118
8. Studier FW (1991) Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. J Mol Biol 219:37–44
9. Miroux B, Walker JE (1996) Over-production of proteins in *Escherichia coli*: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. J Mol Biol 260:289–298
10. Wagner S, Klepsch MM, Schlegel S et al (2008) Tuning *Escherichia coli* for membrane protein overexpression. Proc Natl Acad Sci U S A 105:14371–14376
11. Schlegel S, Löfblom J, Lee C et al (2012) Optimizing membrane protein overexpression in the *Escherichia coli* strain Lemo21(DE3). J Mol Biol 423:648–659
12. Song JM, An YJ, Kang MH et al (2012) Cultivation at 6–10 °C is an effective strategy to overcome the insolubility of recombinant proteins in *Escherichia coli*. Protein Expr Purif 82:297–301
13. Shih Y, Kung W, Chen J et al (2002) High-throughput screening of soluble recombinant proteins. Protein Sci 11:1714–1719
14. Hammarstrom M, Hellgren N, van den Berg S et al (2002) Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*. Protein Sci 11:313–321
15. Bird LE (2011) High throughput construction and small scale expression screening of multi-tag vectors in *Escherichia coli*. Methods 55:29–37
16. Vincentelli R, Cimino A, Geerlof A et al (2011) High-throughput protein expression screening and purification in *Escherichia coli*. Methods 55:65–72
17. Bell MR, Engleka MJ, Malik A, Strickler JE (2013) To fuse or not to fuse: what is your purpose? Protein Sci 22:1466–1477
18. Wang D, Huang XY, Cole PA (2001) Molecular determinants for Csk-catalyzed tyrosine phosphorylation of the Src tail. Biochemistry 40:2004–2010
19. Haacke A, Fendrich G, Ramage P, Geiser M (2009) Chaperone over-expression in *Escherichia coli*: apparent increased yields of soluble recombinant protein kinases are due mainly to soluble aggregates. Protein Expr Purif 64:185–193
20. Thomas JG, Ayling A, Baneyx F (1997) Molecular chaperones, folding catalysts, and the recovery of active recombinant proteins from *E. coli*. To fold or to refold. Appl Biochem Biotechnol 66:197–238
21. Salinas G, Pellizza L, Margenat M et al (2011) Tuned *Escherichia coli* as a host for the expression of disulfide-rich proteins. Biotechnol J 6:686–699
22. De Marco A (2009) Strategies for successful recombinant expression of disulfide bond-dependent proteins in *Escherichia coli*. Microbiol Cell Fact 8:26
23. Mergulhão FJM, Summers DK, Monteiro GA (2005) Recombinant protein secretion in *Escherichia coli*. Biotechnol Adv 23:177–202

24. Zalucki YM, Beacham IR, Jennings MP (2011) Coupling between codon usage, translation and protein export in *Escherichia coli*. Biotechnol J 6:660–667

25. Prinz WA, Åslund F, Holmgren A, Beckwith J (1997) The role of the thioredoxin and glutaredoxin pathways in reducing protein disulfide bonds in the *Escherichia coli* cytoplasm. J Biol Chem 272:15661–15667

26. Bessette PH, Åslund F, Beckwith J, Georgiou G (1999) Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. Proc Natl Acad Sci U S A 96:13703–13708

27. Baneyx F, Mujacic M (2004) Recombinant protein folding and misfolding in *Escherichia coli*. Nat Biotechnol 22:1399–1408

28. Anfinsen CB, Scheraga HA (1975) Experimental and theoretical aspects of protein folding. Adv Protein Chem 29:205–300

29. Tanford C (1997) How protein chemists learned about the hydrophobic factor. Protein Sci 6:1358–1366

30. Rudolph R, Lilie H (1996) In vitro folding of inclusion body proteins. FASEB J 10:49–56

31. Clark E (1998) Refolding of recombinant proteins. Curr Opin Biotechnol 9:157–163

32. Singh SM, Panda AK (2005) Solubilization and refolding of bacterial inclusion body proteins. J Biosci Bioeng 99:303–310

33. Bhavesh NS, Panchal SC, Mittal R, Hosur RV (2001) NMR identification of local structural preferences in HIV-1 protease tethered heterodimer in 6 M guanidine hydrochloride. FEBS Lett 509:218–224

34. Markossian KA, Kurganov BI (2004) Protein folding, misfolding, and aggregation. Formation of inclusion bodies and aggresomes. Biochem Mosc 69:971–984

35. Umetsu M, Tsumoto K, Ashish K et al (2004) Structural characteristics and refolding of in vivo aggregated hyperthermophilic archaeon proteins. FEBS Lett 557:49–56

36. Kopito RR (2000) Aggresomes, inclusion bodies and protein aggregation. Trends Cell Biol 68:524–530

37. Fink AL (1998) Protein aggregation: folding aggregates, inclusion bodies and amyloid. Fold Des 3:R9–R23

38. Kudou M, Yumioka R, Ejima D et al (2011) A novel protein refolding system using lauroyl-L-glutamate as a solubilizing detergent and arginine as a folding assisting agent. Protein Expr Purif 75:46–54

39. Tsumoto K, Umetsu M, Kumagai I et al (2003) Solubilization of active green fluorescent protein from insoluble particles by guanidine and arginine. Biochem Biophys Res Commun 312:1383–1386

40. Tsumoto K, Ejima D, Kumagai I, Arakawa T (2003) Practical considerations in refolding proteins from inclusion bodies. Protein Expr Purif 28:1–8

41. Burgess RR (2009) Refolding solubilized inclusion body proteins. Methods Enzymol 463:259–282

42. Welker E, Wedemeyer WJ, Narayan M, Scheraga HA (2001) Coupling of conformational folding and disulfide-bond reactions in oxidative folding of proteins. Biochemistry 40:9059–9064

43. Tsumoto K, Shinoki K, Kondo H et al (1998) Highly efficient recovery of functional single-chain Fv fragments from inclusion bodies overexpressed in *Escherichia coli* by controlled introduction of oxidizing reagent—application to a human single-chain Fv fragment. J Immunol Methods 219:119–129

44. Li M, Su Z-G, Janson J-C (2004) In vitro protein refolding by chromatographic procedures. Protein Expr Purif 33:1–10

45. Jungbauer A, Kaar W, Schlegl R (2004) Folding and refolding of proteins in chromatographic beds. Curr Opin Biotechnol 15:487–494

46. Geng X, Wang C (2007) Protein folding liquid chromatography and its recent developments. J Chromatogr B 849:69–80

47. Freydell EJ, van der Wielen L, Eppink M, Ottens M (2010) Ion-exchange chromatographic protein refolding. J Chromatogr A 1217:7265–7274

48. Freydell EJ, van der Wielen LAM, Eppink MHM, Ottens M (2010) Size-exclusion chromatographic protein refolding: fundamentals, modeling and operation. J Chromatogr A 1217:7723–7737

49. Matsumoto M, Misawa S, Tsumoto K et al (2003) On-column refolding and characterization of soluble human interleukin-15 receptor α-chain produced in *Escherichia coli*. Protein Expr Purif 31:64–71

50. Yamaguchi S, Yamamoto E, Mannen T, Nagamune T (2013) Protein refolding using chemical refolding additives. Biotechnol J 8:17–31

51. Zardeneta G, Horowitz PM (1994) Detergent, liposome, and micelle-assisted protein refolding. Anal Biochem 223:1–6

52. Arakawa T, Ejima D, Tsumoto K et al (2007) Suppression of protein interactions by arginine: a proposed mechanism of the arginine effects. Biophys Chem 127:1–8

53. Expert-Bezançon N, Rabilloud T, Vuillard L, Goldberg ME (2003) Physical-chemical features of non-detergent sulfobetaines active as protein-folding helpers. Biophys Chem 100:469–479
54. De Bernardez Clark E, Hevehan D, Szela S, Maachupalli-Reddy J (1998) Oxidative renaturation of hen egg-white lysozyme. Folding vs aggregation. Biotechnol Prog 14:47–54
55. Altamirano MM, García C, Possani LD, Fersht AR (1999) Oxidative refolding chromatography: folding of the scorpion toxin Cn5. Nat Biotechnol 17:187–191
56. Teshima T, Kohda J, Kondo A et al (2000) Preparation of *Thermus thermophilus* microspheres with high ability to facilitate protein refolding. Biotechnol Bioeng 68:184–190
57. Tsumoto K, Umetsu M, Yamada H et al (2003) Immobilized oxidoreductase as an additive for refolding inclusion bodies: application to antibody fragments. Protein Eng 16:535–541
58. Preston NS, Baker DJ, Bottomley SP, Gore MG (1999) The production and characterisation of an immobilised chaperonin system. Biochim Biophys Acta 1426:99–109
59. Machida S, Ogawa S, Xiaohua S et al (2000) Cycloamylose as an efficient artificial chaperone for protein refolding. FEBS Lett 486:131–135
60. Daugherty DL, Rozema D, Hanson PE, Gellman SH (1998) Artificial chaperone-assisted refolding of citrate synthase. J Biol Chem 273:33961–33971
61. Nomura Y, Ikeda M, Yamaguchi N et al (2003) Protein refolding assisted by self-assembled nanogels as novel artificial molecular chaperone. FEBS Lett 553:271–276
62. Chiku H, Kawai A, Ishibashi T et al (2006) A novel protein refolding method using a zeolite. Anal Biochem 348:307–314
63. Sakono M, Kawashima Y, Ichinose H et al (2004) Direct refolding of inclusion bodies using reversed micelles. Biotechnol Prog 20:1783–1787
64. Kim Y-S, Randolph TW, Seefeldt MB, Carpenter JF (2006) High-pressure studies on protein aggregates and amyloid fibrils. Methods Enzymol 413:237–253
65. Lee S, Carpenter JF, Chang BS et al (2006) Effects of solutes on solubilization and refolding of proteins from inclusion bodies with high hydrostatic pressure. Protein Sci 15:304–313
66. Baker D (2000) A surprising simplicity to protein folding. Nature 405:39–42
67. Ferguson N, Fersht AR (2003) Early events in protein folding. Curr Opin Struct Biol 13:75–81
68. Chow MKM, Amin AA, Fulton KF et al (2006) REFOLD: an analytical database of protein refolding methods. Protein Expr Purif 46:166–171
69. Rathore AS, Bade P, Joshi V et al (2013) Refolding of biotech therapeutic proteins expressed in bacteria: review. J Chem Technol Biotechnol 88:1794–1806
70. Basu A, Li X, Leong SSJ (2011) Refolding of proteins from inclusion bodies: rational design and recipes. Appl Microbiol Biotechnol 92:241–251

Chapter 2

Expression of Proteins in Insect and Mammalian Cells

Shunsuke Kita, Katsumi Maenaka, and Takashi Tadokoro

Abstract

Producing recombinant proteins in native conformations and functions is one of the most important aspects of structural biology. Because demand for tertiary structure information of mammalian proteins such as membrane proteins and multi-subunit complexes has been increasing in the field of biological and pharmaceutical sciences, target proteins for structural studies have shifted from prokaryote to mammalian proteins. However, since mammalian proteins are often unstable and require posttranslational modifications, it is difficult to prepare sufficient amounts of functional proteins for structural studies using bacteria expression system. Nowadays, insect and mammalian cell expression systems are widely used for overcoming such problems. In this chapter, we explain the basic concepts of these expression systems and provide examples of advanced new techniques including baculovirus-silkworm expression and HEK293 GnTI-cell expression to confer uniformed *N*-glycans suitable for structural studies.

Keywords Protein expression, Insect cells, Mammalian cells, Baculovirus, Silkworm, Constitutive expression, Transient expression

1 Expression of Protein in Insect Cells

1.1 Introduction

1.1.1 Overview

The production of mammalian proteins such as membrane proteins and multi-subunit complexes is often difficult in bacterial expression systems, which do not exhibit mammalian posttranslational modification. Thus, insect cell expression systems are widely used because their machineries for translation, intracellular transport, and posttranslational modification (such as glycosylation, phosphorylation, methylation, acylation, and acetylation) are similar to those in mammalian cells. In addition, insect cells can often express milligram amounts of proteins and are inexpensive compared with mammalian cells [1]. Conversely, the posttranslational modifications provided by insect cells are similar to, but partially different from, those in mammalian cells. Mammalian N-linked glycoproteins contain complex branched sugars, which are composed of mannose, galactose, N-acetylglucosamine, and neuraminic acid. In insect cells, N-linked glycosylation generally results in

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks, DOI 10.1007/978-4-431-56030-2_2, © Springer Japan 2016

oligosaccharides or high-mannose-type oligosaccharides. To introduce mammalian-type glycosylation, customized cell lines known as Mimic™ Sf9 cells (Invitrogen), which express mammalian glycosyltransferases, are commercially available (see Sect. 1.1.2) [2]. Furthermore, the MultiBac system can be used for simultaneous expression of large multi-subunit complexes (see Chap. 3). Therefore, nowadays, mammalian proteins are expressed in insect cells for structural analyses [3–5].

Insect cell expression systems are mainly divided into two systems, viral and nonviral. The viral system uses baculovirus, which has strong infectivity against Lepidoptera (butterflies and moths). The polyhedrin promoter of baculovirus strongly induces expression of downstream genes; thus, the polyhedrin promoter is mostly used in baculovirus systems. At present, various insect cell expression systems are commercially available, such as Bac-to-Bac® (Invitrogen), BaculoDirect™ (Invitrogen), and flashBAC™ (OET). The details of these kits are described in Sect. 1.1.3 [6–8].

The nonviral systems are subdivided into transient and constitutive expression. For transient expression, the foreign gene is introduced only into the host cells and is not integrated into the host genome; thus, the expression does not proceed continuously. In contrast, for constitutive expression, the foreign gene is integrated into the genome so that the gene is replicated and maintained after cell division. Because the transient expression does not need the preparation of a stable clone carrying the gene of interest, it is suitable for rapid expression of the target gene. Although the constitutive expression system requires a much longer time to select a stable highly expressing clone, once the stable clone has been established, the gene of interest is continuously obtained as long as the stable clone survives. Among various constitutive expression systems, the *Drosophila* expression system (Invitrogen) is a well-known system. Details of this system are described in Sect. 1.1.3.4.

1.1.2 Cell Lines and Virus Types

Sf9, Mimic Sf9, Sf21, SF+, and High Five cells are basic cell lines used for recombinant protein production. Sf9, Mimic Sf9, Sf21, and SF+ cells are derived from *Spodoptera frugiperda*. Sf9 cells are the standard cell lines, but Sf21 cells usually show higher protein production than do Sf9 cells. SF+ cells are suitable for suspension cultures and are thus used for large-scale protein production. Mimic Sf9 cells are a derivative of Sf9 cells that stably express mammalian glycosyltransferases. Proteins produced in Mimic Sf9 cells contain terminally sialylated N-glycans different from high-mannose-type N-glycans modified in insect cells. High Five cells, derived from *Trichoplusia ni*, are suited for secretion of recombinant proteins. S2 cells are derived from *Drosophila melanogaster* and are used for both transient and constitutive protein expressions [9].

Two baculoviruses, *Autographa californica* multiple nuclear polyhedrosis virus (AcMNPV) and *Bombyx mori* nuclear polyhedrosis virus (BmNPV), are commonly used to infect the cell lines. AcMNPV can infect Sf9, Mimic Sf9, Sf21, SF+, and High Five cells, while BmNPV is used to infect Bm5 and BmN4 cells derived from silkworms. Silkworms are an attractive protein-producing factory because they have the ability to express abundant amounts of silk proteins. Because BmNPV (and not AcMNPV) directly infects the silkworm *Bombyx mori*, BmNPV expression in larvae and pupae of silkworms is now used for large-scale production of mammalian proteins [10–12].

1.1.3 Commercially Available Kits

1.1.3.1 Bac-to-Bac Expression System

The Bac-to-Bac expression system has an advantage in the production of recombinant virus [6] because it does not require recombination in insect cells. The genome of recombinant virus can be prepared using *E. coli* DH10Bac cells harboring both baculovirus DNA called "bacmid" and helper plasmid coding transposase. In the *E. coli* DH10Bac cells, a gene of interest is transferred from the original transfer vector (known as pFastBac™ series (Invitrogen)) to bacmid by transposases. The successful transposition is verified by the blue-white selection method. Inserting the mini-Tn7 sequence into the mini-*att*Tn7 attachment site on the bacmid disrupts the expression of the LacZα peptide. Thus, the colonies containing the recombinant bacmid turn white, whereas the colonies harboring the unaltered bacmid remain blue. The recombinant bacmid and unaltered bacmid are further distinguished using the PCR method (see Note 1). Once recombinant bacmid is transfected into insect cells, recombinant viruses are generated in a week. The type of virus used in this system is not limited only to AcMNPV. *E. coli* cells harboring the DNA genome of BmNPV were also established and used for silkworm expression. Another unique aspect of the Bac-to-Bac system is the pFastBac™ Dual plasmid harboring a p10 promoter in addition to the conventional polyhedrin promoter, enabling co-expression. In summary, the Bac-to-Bac system establishes a fast and simple method for producing recombinant virus and avoids the need to isolate the recombinant virus from the parent, nonrecombinant virus. The simplified schematic representation of this system is illustrated in Fig. 1.

1.1.3.2 BaculoDirect Expression System

The BaculoDirect expression system introduces Gateway® Technology to facilitate direct transfer of the gene of interest into the baculovirus genome by site-specific recombination of lambda phage [7]. The protocol of the system is as follows: The Gateway® entry clone containing the gene of interest is mixed with BaculoDirect linearized DNA. The gene of interest is transferred from the entry clone to the virus DNA by the recombinase. The success of the recombination can be confirmed by PCR. The resulting

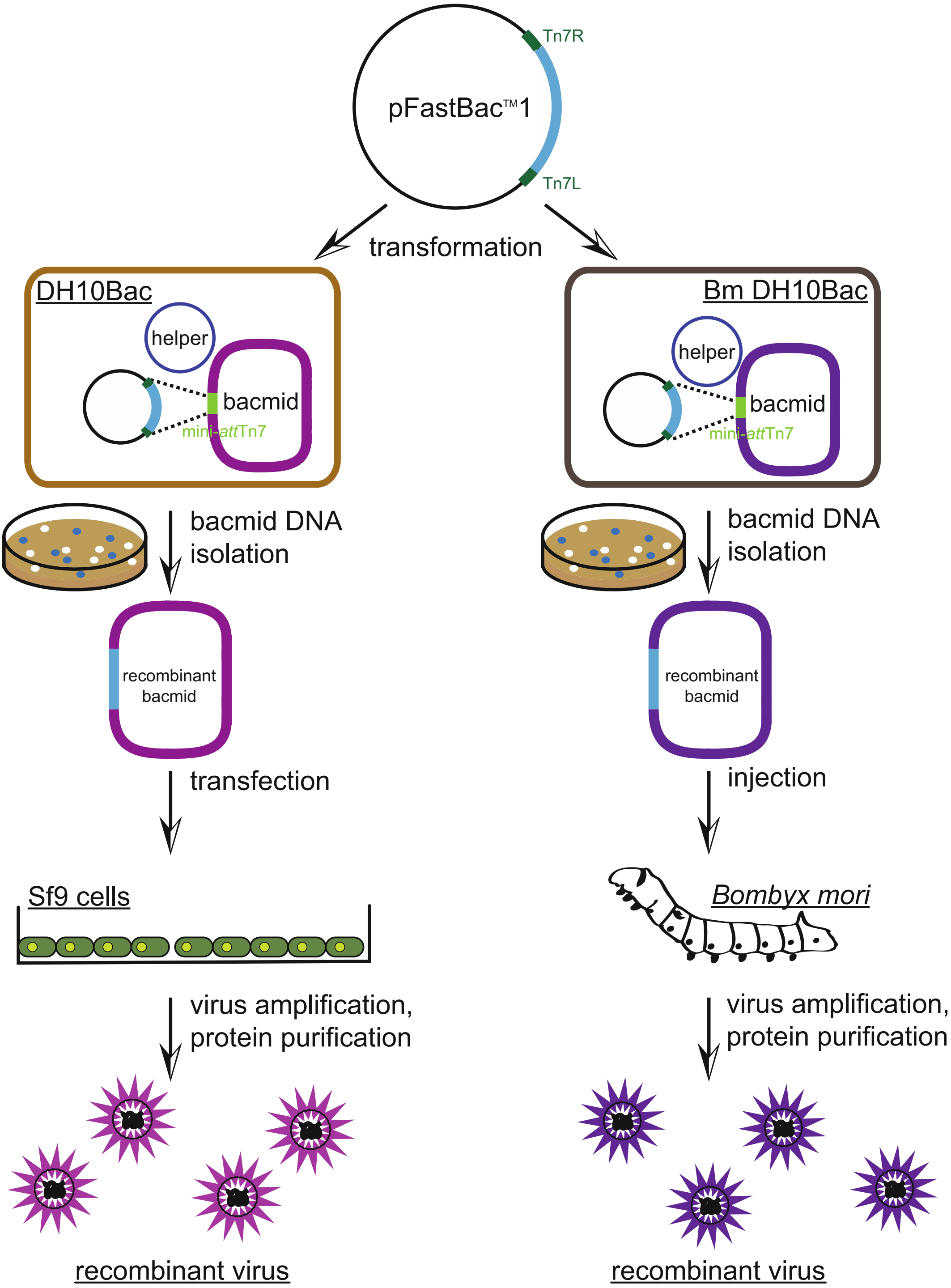


Fig. 1 Schematic representation of the Bac-to-Bac expression system. The flowchart for producing recombinant AcMNPV is shown on the *left* and BmNPV is shown on the right. The pFastBac™ plasmid DNA is introduced into DH10Bac host and BmDH10Bac host. The gene of interest is colored as cyan and the transposition sequences (Tn7R and Tn7L) are colored as green. The recombinant bacmid DNA is extracted from the white colony and is transfected into Sf9 cells and *B. mori* larvae (pupae), to generate the recombinant virus

recombinant virus DNA is transfected into insect cells, usually Sf9 cells, and then the recombinant virus containing the gene of interest is generated. The recombinant virus can be selectively isolated by ganciclovir selection, using the following method, the nonrecombinant virus expresses herpes simplex virus type 1 thymidine kinase (HSV1 tk), which phosphorylates ganciclovir. In Sf9 cells possessing nonrecombinant viruses, ganciclovir is phosphorylated and incorporated into the DNA to inhibit gene replication. Hence, recombinant virus can be selectively replicated after transfection using a BaculoDirect system. The simplified schematic representation of this system is illustrated in Fig. 2.

1.1.3.3 *flash*BAC Expression System

The *flash*BAC system is a simplified method to produce recombinant baculovirus [8]. The advantage of this method is that the recombination occurs by co-transfection of insect cells with the baculovirus genome and the transfer vector containing the gene of interest. The baculovirus genome used in the *flash*BAC system lacks a part of an essential gene (ORF 1629) and contains a bacterial artificial chromosome (BAC) at the polyhedron gene locus, replacing the gene of interest and ORF 1629 downstream of the polyhedrin promoter of the transfer vector if the recombination is successful. Because the deletion of ORF 1629 prevents virus replication in insect cells, the parent virus does not propagate, and the recombinant virus can be obtained without plaque purification. The other advantage of this system is that viruses lacking nonessential genes for replication are commercially available. The *flash*BAC GOLD lacks chitinase A (*chiA*) and V-cathepsin (*v-cath*), which contributes to the productivity and secretion efficiency of the target protein. The *flash*BAC ULTRA lacks p10, p74, and p26 in addition to *chiA* and *v-cath*. Deletion of these nonessential genes from *flash*BAC improves secretion efficiency and protein yield [13, 14]. The simplified schematic representation of this system is illustrated in Fig. 3.

1.1.3.4 *Drosophila* Expression System

The *Drosophila* expression system uses *Drosophila melanogaster* Schneider (S2) cells and combines the advantages of both the mammalian and baculovirus systems [15, 19]. The S2 cells grow rapidly (doubling time, 24 h) and reach high density (5.0×10^7 cells/mL) in a low-cost medium without serum and CO_2. The *Drosophila* expression system achieves expression of mg amount of eukaryotic proteins with low cost. Although the proteins produced in S2 cells do not contain exactly the same posttranslational modifications as mammalian proteins, their activity is often maintained and sometimes higher than those in mammalian proteins. The *Drosophila* expression system can be applied to transient and constitutive expression. The transient expression is used for small-scale and rapid production of a target gene. In transient expression, the

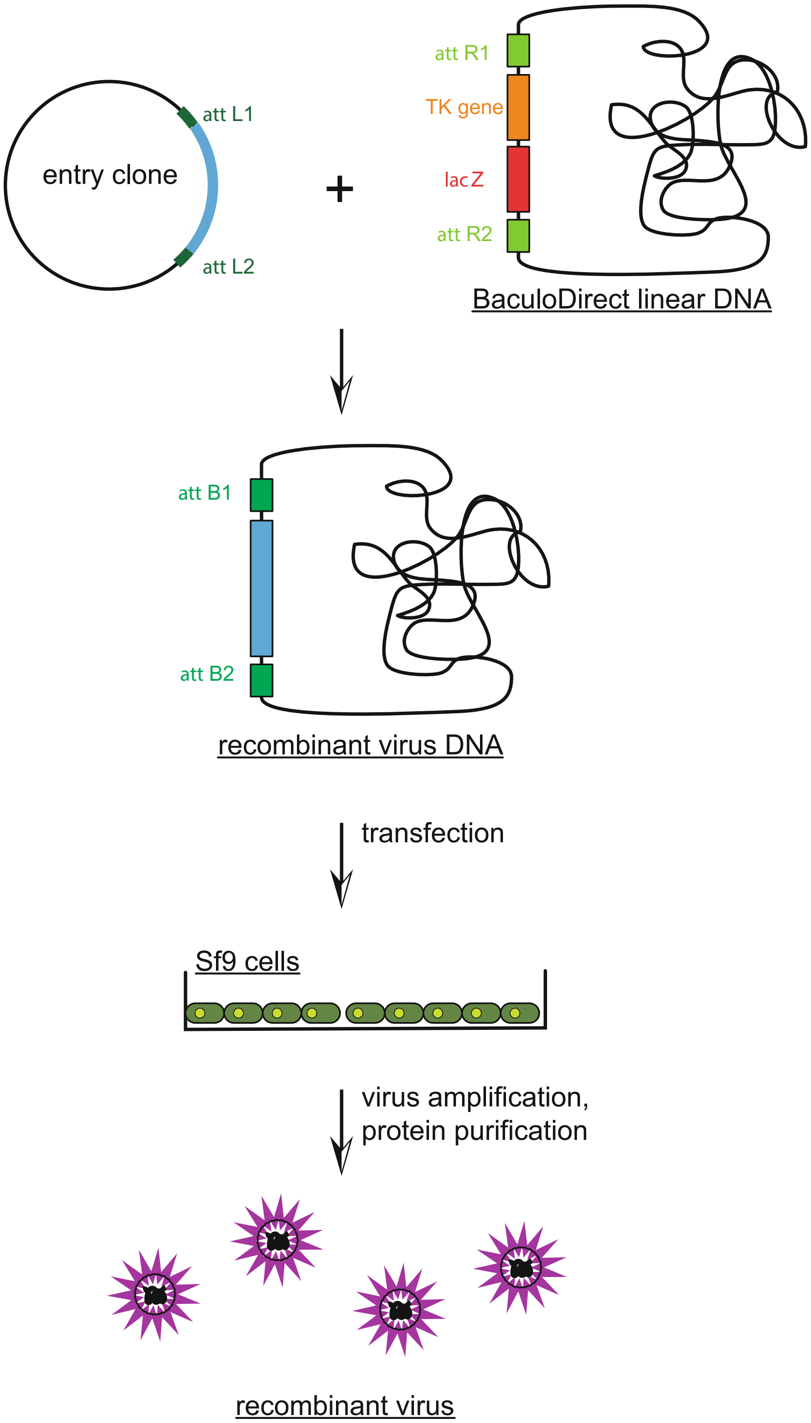


Fig. 2 Schematic representation of the BaculoDirect expression system. The flowchart for producing recombinant AcMNPV is shown. The gene of interest is introduced to BaculoDirect linear DNA using Gateway® Technology. The resulting recombinant virus DNA is transfected into Sf9 cells and the recombinant virus is generated. The gene of interest is cyan and the TK gene is *orange*

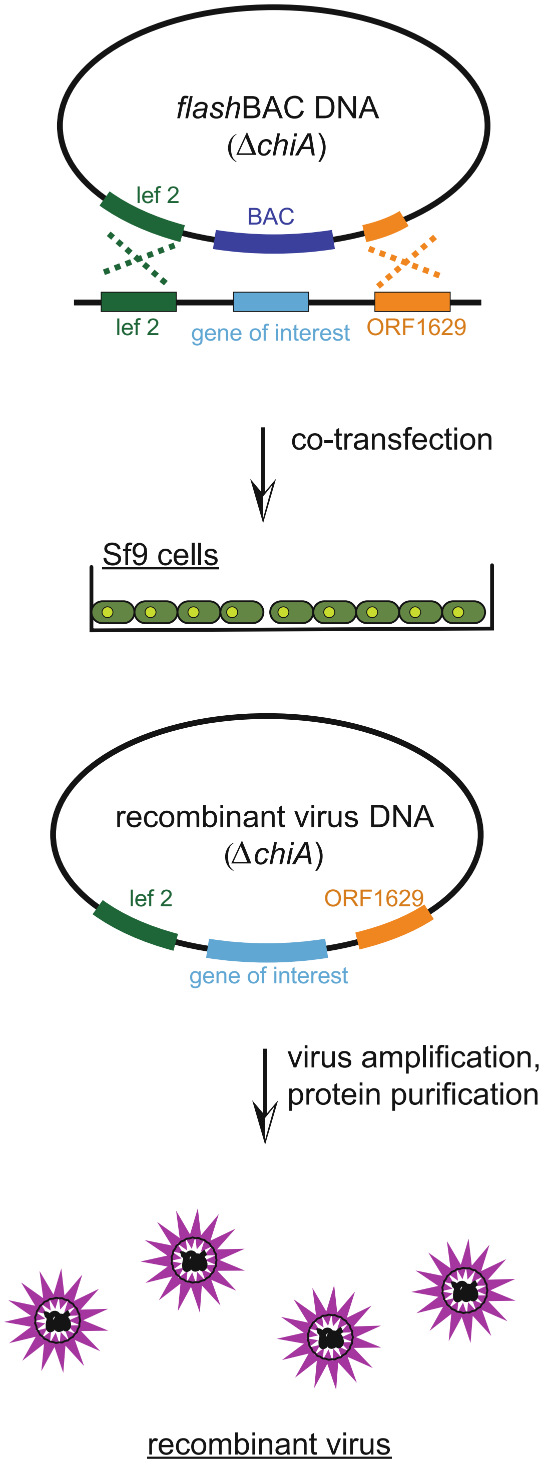


Fig. 3 Schematic representation of the *flash*BAC expression system. The transfer vector containing the gene of interest and *flash*BAC DNA are co-transfected into Sf9 cells. In Sf9 cells, the gene of interest is introduced to *flash*BAC DNA by homologous recombination, and the recombinant virus is generated from it

target gene is transfected into S2 cells and is induced by copper sulfate. Once the appropriate construction is determined, the stable expression clone should be established for large-scale expression. The establishment of the stable clone is achieved by co-transfection of the expression vector harboring the gene of interest and the selection vector harboring the antibiotic resistance gene, such as hygromycin, blasticidin, and puromycin. Optimization of the ratio of the expression and selection vectors leads to the establishment of a clone harboring high-copy number of the gene of interest.

1.2 Materials

In this and the next section, the materials and methods for the production of a recombinant virus using the Bac-to-Bac system are described in detail. Expression in insect cell lines using AcMNPV and in silkworms using BmNPV is described for practical use. A simple flowchart is illustrated in Fig. 1. Green fluorescent protein (GFP) expression in *B. mori* larvae, pupae, and Sf9 cells is shown in Fig. 4.

1.2.1 Preparation of Recombinant Bacmids

Escherichia coli strains: DH10Bac, BmDH10Bac (WT, CP^-, CP^--Chi^-).

Transfer vectors: pFastBac™1, pFastBac™HT, pFastBac™Dual.

Antibiotics: 50 g/L kanamycin, 10 g/L tetracycline, 10 g/L gentamicin, 50 g/L ampicillin.

Media: 2 × yeast extract tryptophan (2 × YT) media, Luria broth (LB) media.

Chemicals: isopropyl β-D-thiogalactopyranoside (IPTG), X-gal.

1.2.2 Expression in Insect Cells

Insect cell lines: Sf9, Sf21, High Five.

Media: Sf900 II SFM (Invitrogen), Fetal Bovine Serum (HyClone), Express Five SFM.

Transfection reagents: X-tremeGENE HP Transfection Reagent (Roche).

Flasks: Erlenmeyer flasks.

Safety cabinet: biological safety cabinet (Sanyo).

Shaker incubator: Bioshaker (Taitec).

1.2.3 Expression in Silkworm

Silkworms: *Bombyx mori* (Kinshu × Showa race) (Ehime-Sanshu).

Feed: synthetic diet (Ehime-Sanshu).

Rearing cage: plastic Tupperware.

Transfection reagents: DMRIE-C (Invitrogen).

10 × phosphate-buffered saline (PBS): 1.37 M NaCl, 27 mM KCl, 100 mM Na_2HPO_4, 18 mM KH_2PO_4. Adjust the pH to 7.4 with HCl.

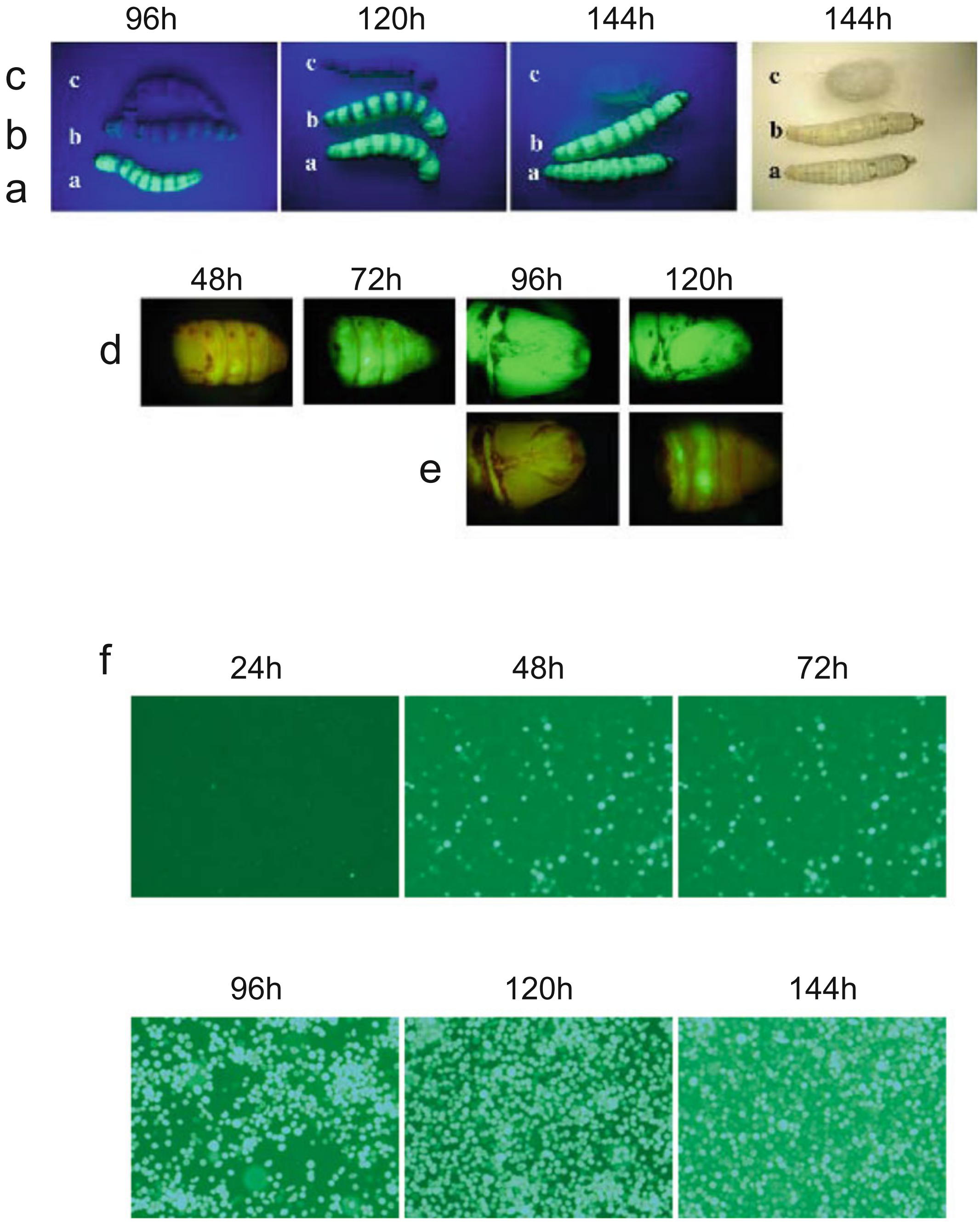


Fig. 4 GFP expression in *B. mori* larvae, pupae, and Sf9 cells. *B. mori* larvae were infected by needlepoint immersed in the BmNPV/GFP virus (*a*), by direct syringe injection of BmNPV bacmid/GFP DNA (*b*) and mock (*c*). *B. mori* pupae were infected by needlepoint immersed in the BmNPV/GFP virus (*d*) and by direct injection of BmNPV bacmid/GFP DNA using a pipette (*e*). The photographs of the larvae were taken at 96, 120, and 144 h after infection using a UV illuminator in complete darkness. (*f*) Sf9 cells were transfected with GFP/pFastBac1 plasmid using X-Treme GENE. The photographs of the cells were taken at 24, 48, 72, 96, 120, and 144 h after infection [10] (Reprinted from Ref. [10], with permission from Elsevier)

1.3 Methods

1.3.1 Expression of Proteins in Insect Cells

1.3.1.1 Preparation of Recombinant Bacmids

1. Introduce 1–50 ng recombinant plasmid DNA into DH10Bac chemically competent cells.
2. Heat the cells at 42 °C for 45 s. Incubate on ice for 2 min.
3. Add 1 mL room temperature LB medium.
4. Shake tubes at 37 °C for 1 h.
5. Add tetracycline (final concentration, 10 μg/mL) and shake tubes at 37 °C for 12–15 h.
6. Add gentamicin (final concentration, 7 μg/mL) and shake tubes at 37 °C for 2 h.
7. Prepare three tenfold serial dilutions of the transformed cells (10^{-1}, 10^{-2}, 10^{-3}) with LB medium, and plate them on LB agar plates containing appropriate antibiotics and chemicals (50 μg/mL kanamycin, 7 μg/mL gentamicin, 200 μM IPTG, 40 μg/mL X-gal).
8. Incubate the LB plates at 37 °C for 24 h.
9. Pick the white colonies using sterilized chips and re-streak them on fresh LB agar plates. After re-streaking, dip the chips into a tube containing PCR reaction mixture. The reaction mixture is subjected to agarose gel electrophoresis (see Note 2).
10. Isolate the bacmids from the positive clones by using a commercially available mini-prep kit (see Note 3).

1.3.1.2 Preparation of Recombinant Virus (Insect Cells)

1. Inoculate Sf9 cells in six wells and culture the cells until semi-confluent.
2. Prepare bacmid reagents by mixing 2 μg bacmid with 100 μL Sf900 II SFM.
3. Prepare transfection reagents by mixing 8 μL X-tremeGENE with 100 μL Sf900 II SFM.
4. Mix bacmid reagents with transfection reagents.
5. Incubate bacmid and transfection mixture at room temperature for 30 min.
6. Remove the culture medium from each well and add 800 μL Sf900 II SFM.
7. Add 200 μL transfection reagent as droplets.
8. Incubate at 27 °C for 5 h.
9. Add 1 L Sf900 II SFM supplemented with 10 % fetal bovine serum (FBS).
10. Incubate the plate at 27 °C for 5 days.

11. Collect the supernatant in sterilized tubes.
12. Centrifuge the tubes for 5 min at 1000 × g to remove extra cells.
13. Collect the supernatant and store at 4 °C until use (P1 virus).

1.3.1.3 Amplification of Recombinant Virus

1. Inoculate Sf9 cells in six wells and culture the cells until semi-confluent.
2. Remove the culture medium and add the P1 virus stock.
3. Incubate the plate at 27 °C for 5–7 days.
4. Collect the supernatant from the wells in sterilized tubes.
5. Centrifuge the tubes for 5 min at 1000 × g to remove extra cells.
6. Collect the supernatant and store at 4 °C until use (P2 virus).

1.3.1.4 Expression Check

1. Inoculate Sf9/Sf21/High Five cells in six wells and culture the cells until semi-confluent.
2. Remove the culture medium and add the P2 virus stock.
3. Incubate the plate at 27 °C for 3–5 days.
4. Collect both the supernatant and the cells from the wells in the same sterilized tubes.
5. Centrifuge the tubes for 5 min at 1000 × g and collect the supernatant and cells separately.
6. Check for protein expression in the medium.
   (a) Mix the supernatant with SDS-PAGE sample buffer.
   (b) If the expression of the protein is low, the supernatant should be concentrated using trichloroacetic acid (TCA).
7. Check for protein expression in the cell lysate.
   (a) Add 100 μl appropriate buffer to the collected cell (e.g., 50 mM HEPES-NaOH pH 7.5, 250 mM NaCl, 10 % glycerol, 1 mM β-ME, 1 % NP40, 1 × protease inhibitor cocktail).
   (b) Vortex briefly and incubate at 4 °C for 10 min.
   (c) Repeat step (b) for two times.
   (d) Centrifuge the cells for 30 min at 20,000 × g at 4 °C.
   (e) Collect both the supernatant (soluble fraction) and the precipitant (insoluble fraction), and mix them with SDS-PAGE sample buffer.

1.3.2 Expression in Silkworm

1.3.2.1 Preparation of Recombinant Bacmid Using BmDH10Bac Cells

Same as the method described above (see Sect. 1.3.1.1).

1.3.2.2 Injection of Recombinant Bacmid into Silkworm *Bombyx mori*

1. Mix 50 μL recombinant bacmid (20 ng/μL) with 3 μL DMRIE-C (Invitrogen).
2. Incubate the bacmid mixture at room temperature for 45 min.
3. Inject 50 μl bacmid mixture on the first day of the fifth instar larvae using a syringe with a 26-gauge needle and on the fourth day of pupae using a pipette.
4. Cultivate the infected silkworm larvae and pupae at 25 °C for 5–7 days (see Note 4).

1.3.2.3 Collection of the Hemolymph and Fat Body from Infected Silkworm Larvae and Pupae

Isolation of the Hemolymph Fraction from Silkworms

1. Transfer the infected silkworm onto ice for 2 min to stop their activities.
2. Make a small hole at the center of the larvae using a syringe with a 26-gauge needle.
3. Isolate the hemolymph to tubes containing 50 μl 5 % (w/v) sodium thiosulfate per one larva (see Note 5).
4. Centrifuge the tubes for 10 min at 20,000 × g at 4 °C.
5. Transfer the supernatant to new tubes and freeze them using liquid nitrogen and store them at −80 °C until use.

Isolation of the Fat Body Fraction from Silkworm

1. Collect the hemolymph fraction using the same method as above (see section "Isolation of the hemolymph fraction from silkworms").
2. Secure the larvae onto the rubber plate using pins to prepare them for dissection.
3. Slice the larvae down its back using scissors and pin the skin of larvae onto the rubber plate.
4. Discard the internal organs and wash the skin using 1 × PBS with 0.5 % (w/v) sodium thiosulfate.
5. Scrape off the fat bodies from the skin and collect them into tubes containing 0.5 % (w/v) sodium thiosulfate.
6. Centrifuge the tubes for 10 min at 10,000 × g at 4 °C.
7. Discard the supernatant and freeze the fat bodies using liquid nitrogen and store them at −80 °C until use.

2 Expression in Mammalian Cells

2.1 Introduction

We have described the recombinant protein expression in bacteria and in insect cells (Chaps. 1 and 2). At present, both of the expression systems are generally more productive and less costly than mammalian cell systems. However, using these systems, we sometimes face difficulties in producing recombinant proteins such as secreted and/or membrane proteins, which require posttranslational modifications and usually result in accumulation of inclusion bodies. In these cases, mammalian cells would be another option to consider, because mammalian cells can produce high molecular weight proteins, membrane proteins, and posttranslationally modified proteins derived from eukaryotes. Here we provide guidance for overexpression of recombinant proteins in mammalian cells.

One of the biggest advantages of using mammalian expression systems is the appropriate posttranslational modifications including glycosylation, disulfide formation, phosphorylation, and acetylation, which provide appropriate biological activities to the proteins of interest—one of the main concerns for biological research and medical use. Another advantage is that the proteins can be secreted in the culture supernatant by adding a mammalian signal sequence to the gene of interest. This approach ensures the quality of the proteins because each export step checks whether they are properly folded and modified. It is also helpful for reducing the number of steps required for protein purification, because the culture supernatants, which have relatively low contaminants, can be directly applied to tag-based affinity chromatography. Finally, mammalian expression systems are adopted to produce high molecular weight proteins, which are often expressed in insoluble forms in different systems such as *E. coli*.

Similar to the other expression systems, the selection of vectors and host mammalian cell lines is important for producing recombinant proteins. At present, several mammalian promoters are available; CMV (human cytomegalovirus), CAG (chicken β-actin promoter coupled with CMV), SV40 (simian virus 40), human EF-1α (elongation factor-1 α), and human UbC (human ubiquitin C) are utilized for the constitutive expression, and tTA/Tet (a tetracycline-controlled transactivator protein (tTA); the expression of the transactivator can be regulated with the different concentrations of tetracycline (Tc)), GLVP/TAXI (GAL4-PR-LBD) (GLVP is a regulator controlled by the GAL4 DNA binding, consisting of a human progesterone receptor ligand-binding domain (PR-LBD) fused to the yeast GAL4 DNA-binding domain), and GAL4-E1b (consisting of the binding sites for the yeast GAL4 DNA-binding domain followed by the adenovirus E1b promoter TATA box) are utilized for inducible expression. Constitutive promoters are usually chosen for overproduction of recombinant proteins unless strict control of its expression is necessary because of the toxicity to cell growth. There are several cell lines commonly used for protein expression, including COS7 (derived from monkey kidney tissue), CHO (derived from Chinese hamster ovary), and HEK293 (derived from human embryonic kidney). The heterogeneity of the posttranslational modification that occurred in mammalian cells, especially glycosylation, sometimes affects the quality of the protein crystals. To overcome this problem, the glycosylation modified cell lines are also available, such as CHO Lec1$^-$ and HEK293S GnTI$^-$, both of which are unable to synthesize complex N-glycans due to the deficiency in N-acetylglucosaminyltransferase I (GnTI) [16, 17]. Recently, further modified cell lines, such as Expi293F™ cell line (Thermo Fisher), are available to achieve greater protein yield. Usually, the expression level in mammalian cells is not high, and the addition of affinity tags is necessary to resolve some issues in protein preparation by increasing the yield, enhancing folding, and reducing the number of purification steps. The affinity tag technology is described in a later chapter (see Chap. 4).

Transfection of mammalian cells with plasmid vectors is quite different from transformation of *E. coli*. In the case of *E. coli*, a single-plasmid DNA enters into a single cell, and the number of copies of the DNA increases as a result of the host cell enzymes. In principle, each *E. coli* clone transformed with the same plasmid vector is identical. In contrast, transfection with popular transfection reagents typically introduces multiple-plasmid DNAs into a cell. Unlike typical plasmids for *E. coli*, plasmids for mammalian cells do not replicate in host cells (unless they have a viral replication origin). Among transfectants, a small percentage of cells, usually less than 10 % or even lower, possess the gene of interest randomly integrated into the host cell genome. The number of the integrated genes and the locus (loci) of the integrated gene(s) differs between individual cells. Therefore, each clone is not identical after selection with drugs and varies significantly in protein productivity.

Protein expression in mammalian cells can be performed using two procedures, transient and constitutive expression, as described earlier for insect cell expression. For transient expression, the transfection efficiency directly alters the yield. We usually harvest cells expressing recombinant protein from a few days up to a week after the transfection because the yield decreases after longer periods. To overcome this problem, constitutive expression with stable cell lines is required. To establish stable cell lines, transfection is carried out in the same way for transient expression, and the cells are grown without a selection drug for a few days. As a next step, cells were cultured and selected in the presence of appropriate drugs. For this purpose, the plasmid for transfection must contain a drug-resistant gene. Within 1–7 days, most of the cells die in the presence of the drug and floating cells may be visible. This is either because the cells are not transfected or the drug-resistant gene is not integrated into the host genomes. For drug selection, the medium containing the drug should be changed to a fresh medium every 3–5 days until certain cells survive and form a colony(ies). This drug selection process usually takes a month or more to obtain a single colony. These selected clones have potential for high productivity. Many factors that determine the productivity, such as the genome location and the number of genes integrated, are unpredictable. Thus, it is recommended to select several clones. However, this would lead to consistent overproduction of the target protein once a stable cell line is established.

Another choice of gene delivery into mammalian cells is viruses: lentiviruses, adenoviruses, and insect viruses (baculoviruses). The lentiviral vectors are derived from the human immunodeficiency virus (HIV-1). The advantage of the lentiviral system is the capability to mediate transduction and constitutive expression of the gene of interest into dividing and nondividing cells. Thus, it is useful to establish the stable cell lines. However, generation and transduction of the lentiviruses require biosafety level 3 (BSL3). The adenoviruses are also useful for the recombinant protein production since it can be prepared at high titer. Moreover, the viruses are not usually integrated into the host genome so that it is relatively safe for the researchers.

Finally, we describe baculoviruses modified to be used in mammalian cells. The baculoviruses modified by engineering of a mammalian expression cassette for transgene expression in mammalian cells are commonly referred to as BacMam viruses. There is concern that mammalian viruses could be harmful to not only mammalian cells but also the researchers themselves. Conversely, insect viruses are much safer because of their inability to replicate in mammalian cells. Various mammalian cell lines have been reported to be transducible with baculovirus [18]. The procedure for virus generation and amplification is the same as those for the baculovirus insect cell system (see Sect. 1).

Here we provide an example of transient protein expression in mammalian cells using human signaling lymphocytic activation molecule (hSLAM), a cellular receptor of measles virus that mediates important regulatory signals in immune cells [19]. Overproduction and purification procedures were slightly modified from the reported ones [19].

2.2 Materials

2.2.1 Transfection and Cell Harvesting

Materials: *PBS (phosphate-buffered saline), incomplete Dulbecco's Modified Eagle's Medium (DMEM) "serum-free medium," "complete DMEM," DMEM supplemented with 10 % fetal bovine serum (FBS).*

Cells: *HEK293T cells, grown in complete DMEM. HEK293T cells contain the SV40 large T antigen, which allows episomal replication of transfected plasmids containing the SV40 origin of replication. This allows amplification of transfected plasmids and extended temporal expression of the desired gene products.*

Plasmids: *The pCA7-hSLAM plasmid, for the overproduction of hSLAM protein extracellular domain [19]. The plasmid was constructed by ligating PCR-amplified fragments of the hSLAM extracellular domain into the expression vector pCA7. This vector contains the signal sequence derived from the pHLsec vector [20], which is located upstream of the protein-coding sequence. Therefore, the* *hSLAM is expected to be secreted to the culture medium because of the signal sequence.*

Reagents: *Polyethylenimine (PEI) max (Polysciences) is used for transfection.*

Equipment: *Basic tissue culture facilities, e.g., tissue culture hood, 37 °C, 5 % CO_2 tissue culture incubator, cell culture dish and/or flask, centrifuge for harvesting cells, aspirator, 0.45-μm pore size filter unit.*

Buffers: *10 × wash buffer, 500 mM Na_2HPO_4, 1.5 M NaCl, 100 mM imidazole.*

2.2.2 His-Tag Purification

Colum: *HisTrap™ FF 5 mL (GE Healthcare)*

Buffers: *Wash buffer (1 × wash buffer), 50 mM Na_2HPO_4, 150 mM NaCl, 10 mM imidazole. Elution buffer, 50 mM NaH_2PO_4, 150 mM NaCl, 500 mM imidazole.*

Make sure that all buffers for column purification are degassed before use.

Equipment: *Peristaltic pump, stand to hold a column, SDS-PAGE, Western blot apparatus, and power supplies.*

2.3 Methods

2.3.1 Transfection

Transfection is carried out using PEI-mediated gene delivery strategy (Fig. 5). PEI is a synthetic polycation that can bind tightly to DNA. Thus, PEI forms a stable complex with the plasmid DNA, which enters the nucleus via endocytosis without significant cellular

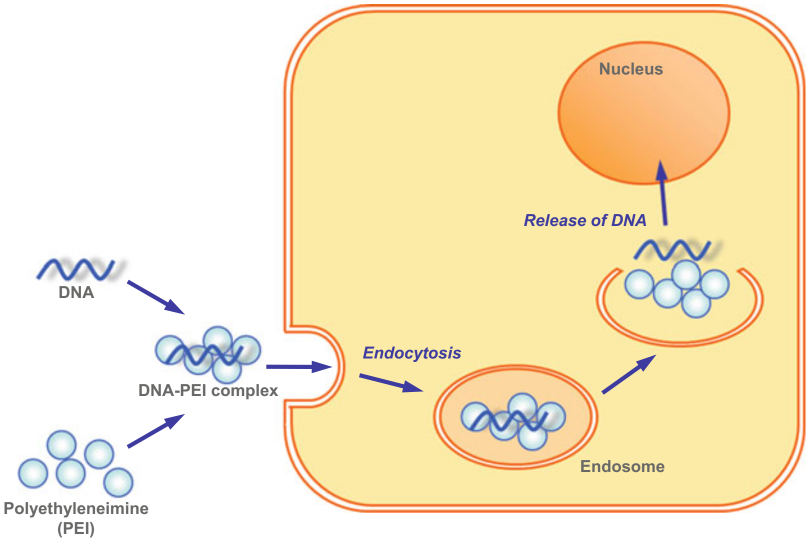


Fig. 5 PEI-mediated transfection

obstacles. The typical protocol for PEI-mediated transfection is described below and should be optimized depending on the purpose (see Notes 6, 7, 8, and 9).

1. Seed the HEK293T cells in DMEM supplemented with 10 % FBS, L-glutamine, and nonessential amino acids 1 day before transfection. Grow the cells in a 15-cm dish until they were approximately 80 % confluent.
2. Add 1 mL each of DMEM without serum into two tubes. Add 50 μg plasmid encoding the hSLAM extracellular domain (Fig. 6a) to one tube containing medium and PEI Max with twice the amount of the plasmid (100 μg in this case) to

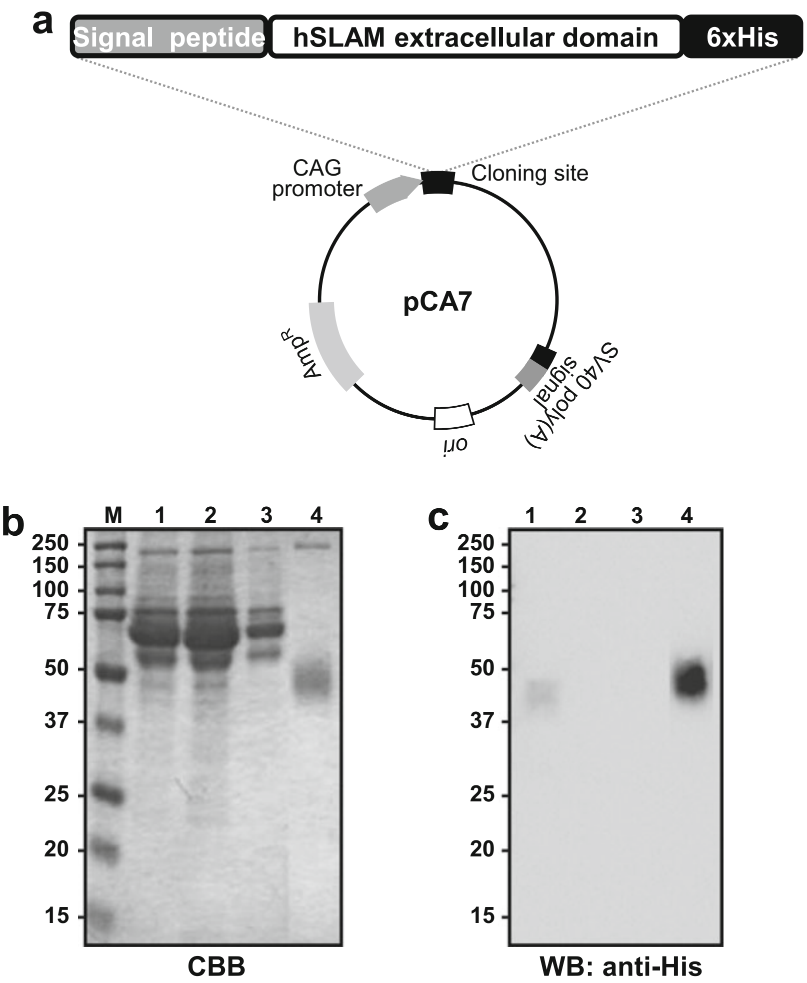


Fig. 6 Expression of the hSLAM in mammalian cells. (**a**) The map of the hSLAM expression plasmid, pCA7-hSLAM. (**b**) SDS-PAGE of the protein of interest. *M* Protein size marker, *1* culture supernatant, *2* flow-through fraction, *3* wash fraction, *4* elution fraction. (**c**) Western blot probed with anti-His antibody. The loaded samples are identical to (**b**)

another tube. The amount of medium and DNA described above is for one 15-cm dish. Mix thoroughly by tapping tubes.

3. Combine DNA mixture and PEI Max mixture in an appropriate tube, and mix well.
4. Incubate for 15 min at room temperature. Discard the medium and replace it with fresh DMEM medium containing 2 % (v/v) FBS. DMEM containing 2 % FBS is prepared by mixing complete DMEM containing 10 % FBS and serum-free DMEM.
5. Add the plasmid DNA-PEI Max mixture to the cells. After adding the reagent mixture, mix well by swirling gently.
6. Incubate overnight in 5 % CO_2 incubator, and change the medium to complete DMEM the next day.
7. Incubate in 5 % CO_2 incubator for a further 3–5 days. The protein of interest is secreted to the medium via signal sequencing.

2.3.2 Harvesting of Secreted Protein Containing the Culture Supernatant

1. Harvest the culture supernatant.
2. Add 10× wash buffer, about 1/10 volume of the supernatant, and incubate overnight at 4 °C.
3. Centrifuge at 5000 × g for 10 min.
4. Filter the medium using a 0.45-μm pore size filter. The filtration is performed by aspirator if the total volume is more than 100 mL or using a syringe with an attached filter for smaller volumes.

2.3.3 His-Tag Purification

Purify the protein of interest from the culture supernatant using His-tag affinity chromatography. The details of affinity chromatography are described in Chap. 4.

1. Centrifuge the cultured medium at 5000 × g for 10 min to remove cell debris.
2. Filter the harvested supernatant with a 0.45-μm membrane filter.
3. Attach a HisTrap™ FF column to a peristaltic pump and equilibrate it with wash buffer (more than three times the column volume).
4. Inject the culture supernatant into the column, and collect the liquid fraction.
5. After injection, equilibrate the column with wash buffer (more than five times the column volume). Collect the wash fraction.
6. Elute the protein using the elution buffer (one to five column volume). Pool the fractions containing the protein.

7. Analyze the protein samples using SDS-PAGE with Coomassie Brilliant Blue (CBB) staining and Western blotting (Fig. 6b, c) (see Note 10). Further purification steps should be considered if good quality is required for structural studies including X-ray crystallography.

3 Notes

1. Recombinant bacmid DNA is extracted from the white colonies and is further analyzed by PCR using M13 Forward (−40) and M13 Reverse primers. As these primers are complementary to either end of the transposition site, the success of the transposition is verified by PCR analysis.
2. PCR is performed using GoTaq® Master Mix (Promega). The length of the amplified DNA from a positive clone becomes 2300 base pairs (bp) plus the length of the target gene. The length of amplified DNA from a negative clone becomes 300 bp.
3. Bacmid DNA is sometimes damaged by freezing and thawing. Bacmid DNA should be stored at 4 °C.
4. The silkworm is cultivated in the incubator at 25 °C. A beaker containing water should be placed on the top shelf of the incubator to maintain an appropriate humidity in the incubator. Feed the synthetic diet twice a day.
5. The collected hemolymph immediately turns black without sodium thiosulfate. 1-phenyl-2-thiourea could be used as a substitute for sodium thiosulfate.
6. We would suggest that the optimization of the transfection condition at a small scale using a 24- or 6-well plate is better initially. Several factors can be considered: the amount of plasmid DNA, the ratio of DNA-PEI, the incubation time of the DNA-PEI mixture, the medium (serum percentage) and host cells, and the cultivation time after transfection (see Fig. 7).
7. Other transfection reagents, such as FuGENE HD (Promega), can be used instead of polyethylenimine (PEI) when the transfection efficiency is low.
8. It is advisable to select the plasmid containing a drug-resistant gene if drug selection and/or establishment of stable cell lines is important. G418, also known under the trade name Geneticin, is often used for this purpose. The drug is inactivated by the neomycin phosphotransferase encoded in NeoR or KanR, originally isolated from a bacterial transposon, Tn5. To be resistant to G418, the gene must be expressed by the appropriate promoter and have the appropriate expression for the host cells.

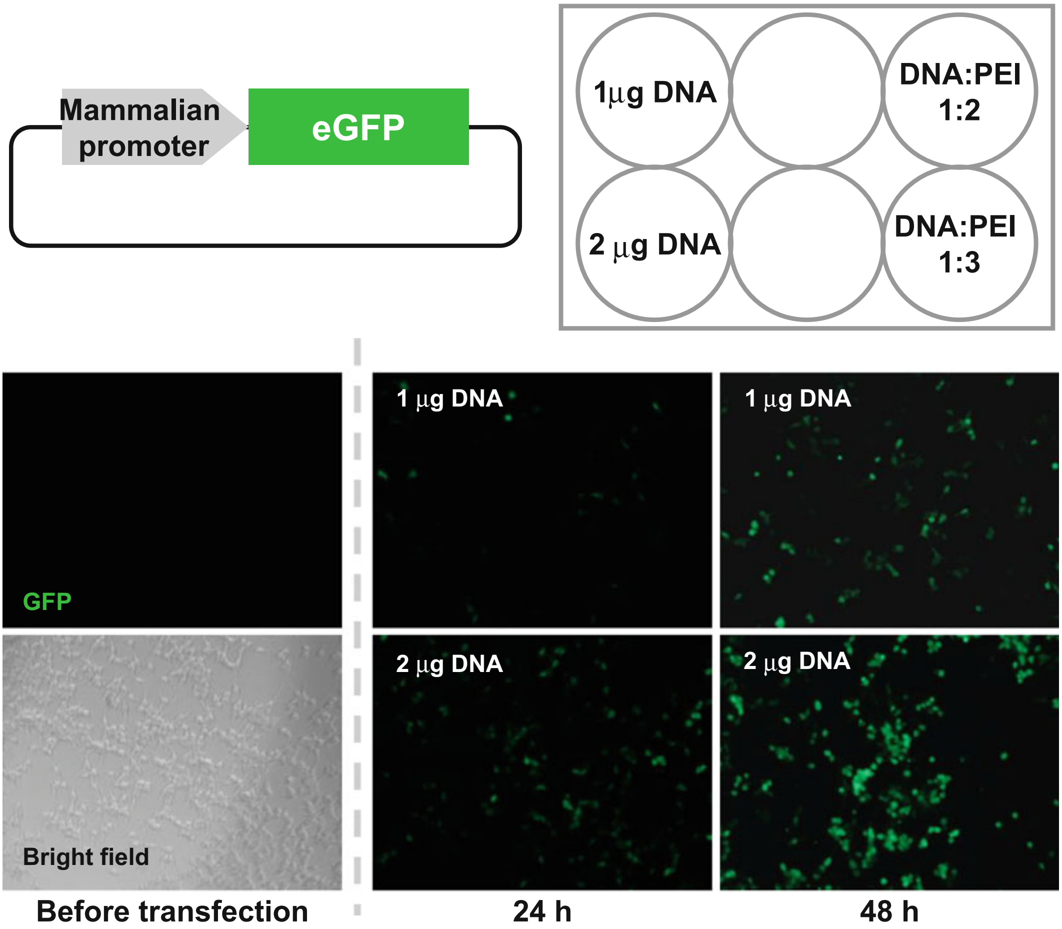


Fig. 7 Example of the optimization of transfection condition. HEK293T cells were transfected with the plasmid encoding mammalian promoter and eGFP as a reporter. Images used to examine the amount of DNA and DNA-PEI ratio are shown

9. If the secretion of the protein is not sufficient, the signal peptide sequence for secretion can be replaced with another signal peptide. While we provided an example using the pHLsec signal sequence (MGILPSPGMPALLSLVSLLSVLL-MGCVAE), other signal sequences are commercially available vectors, such as pDISPLAY (Invitrogen) or pSecTag2 (Invitrogen).

10. It is strongly recommended, especially for a preliminary experiment, to keep aliquots utilized at various stages of expression and purification. If there is a problem in the yield or/and purity of the target protein, the aliquoted samples should be analyzed, for example, the cell pellets and cell debris in step 1, Sect. 2.3.2, and the flow-through fraction in step 4, Sect. 2.3.3.

References

1. Kollewe C, Vilcinskas A (2013) Production of recombinant proteins in insect cells. Am J Biochem Biotechnol 9:255–271. doi:10.3844/ajbbsp.2013.255.271
2. Growth and maintenance of mimic™ Insect cells. Invitrogen, Thermo Fisher Scientific
3. Massotte D (2003) G protein-coupled receptor overexpression with the baculovirus–insect cell system: a tool for structural and functional studies. Biochim Biophys Acta Biomembr 1610:77–89. doi:10.1016/S0005-2736(02)00720-4
4. Bieniossek C, Imasaki T, Takagi Y, Berger I (2012) MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem Sci 37:49–57. doi:10.1016/j.tibs.2011.10.005
5. Berger I, Fitzgerald DJ, Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. Nat Biotechnol 22:1583–1587. doi:10.1038/nbt1036
6. Bac-to-Bac Baculovirus Expression System. Invitrogen, Thermo Fisher Scientific
7. BaculoDirect™ baculovirus expression system. Invitrogen, Thermo Fisher Scientific
8. flashBAC one-step baculovirus protein expression. Oxford Expression Technologies (2008)
9. Moraes AM, Jorge SAC, Astray RM et al (2012) Drosophila melanogaster S2 cells for expression of heterologous genes: from gene cloning to bioprocess development. Biotechnol Adv 30:613–628. doi:10.1016/j.biotechadv.2011.10.009
10. Motohashi T, Shimojima T, Fukagawa T et al (2005) Efficient large-scale protein production of larvae and pupae of silkworm by *Bombyx mori* nuclear polyhedrosis virus bacmid system. Biochem Biophys Res Commun 326:564–569. doi:10.1016/j.bbrc.2004.11.060
11. Kato T, Kajikawa M, Maenaka K, Park EY (2010) Silkworm expression system as a platform technology in life science. Appl Microbiol Biotechnol 85:459–470. doi:10.1007/s00253-009-2267-2
12. Kajikawa M, Sasaki K, Wakimoto Y et al (2009) Efficient silkworm expression of human GPCR (nociceptin receptor) by a *Bombyx mori* bacmid DNA system. Biochem Biophys Res Commun 385:375–379. doi:10.1016/j.bbrc.2009.05.063
13. Kaba SA, Salcedo AM, Wafula PO et al (2004) Development of a chitinase and v-cathepsin negative bacmid for improved integrity of secreted recombinant proteins. J Virol Methods 122:113–118. doi:10.1016/j.jviromet.2004.07.006
14. Hitchman RB, Possee RD, Siaterli E et al (2010) Improved expression of secreted and membrane-targeted proteins in insect cells. Biotechnol Appl Biochem 56:85–93. doi:10.1042/BA20090130
15. Drosophila expression system. Invitrogen, Thermo Fisher Scientific
16. Stanley P, Chaney W (1985) Control of carbohydrate processing: the lec1A CHO mutation results in partial loss of N-acetylglucosaminyltransferase I activity. Mol Cell Biol 5:1204–1211. doi:10.1128/MCB.5.6.1204
17. Reeves PJ, Callewaert N, Contreras R, Khorana HG (2002) Structure and function in rhodopsin: high-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible N-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. Proc Natl Acad Sci U S A 99:13419–13424. doi:10.1073/pnas.212519299
18. Kost TA, Condreay JP, Jarvis DL (2005) Baculovirus as versatile vectors for protein expression in insect and mammalian cells. Nat Biotechnol 23:567–575. doi:10.1038/nbt1095
19. Hashiguchi T, Ose T, Kubota M et al (2011) Structure of the measles virus hemagglutinin bound to its cellular receptor SLAM. Nat Struct Mol Biol 18:135–141. doi:10.1038/nsmb.1969
20. Aricescu AR, Assenberg R, Bill RM et al (2006) Eukaryotic expression: developments for structural proteomics. Acta Crystallogr D Biol Crystallogr 62:1114–1124

Chapter 3

Application of MultiBac System to Large Complexes

Shuya Fukai

Abstract

Multisubunit protein complexes regulate numerous biologically important processes. Elucidation of their functional mechanisms based on their three-dimensional structures allows us to understand biological events at the molecular level. Crystallography and electron microscopy are powerful tools for analyzing the structures of biological macromolecules. However, both techniques require large-scale preparation of pure and structurally homogenous samples, which is usually challenging for large multisubunit complexes, particularly from eukaryotes. In this chapter, we describe the principles and methods of producing multisubunit complexes in insect cells using the MultiBac system.

Keywords Multisubunit protein complexes, Baculovirus, Insect cell expression, MultiBac system

1 Introduction

Multisubunit protein complexes play critical roles in numerous cellular processes, such as gene regulation, protein degradation, and intracellular signaling. Some complexes are abundant in cells and can be obtained from natural sources for use in biochemical and biophysical experiments. However, expressions of many others are spatiotemporally restricted, particularly in eukaryotes; therefore, overproduction of the recombinant complexes is often required for biochemical and biophysical analyses. Multisubunit protein complexes can be prepared in two different ways: recombinant subunits are either produced separately in cells and reconstituted in vitro, or coexpressed and reconstituted in cells. The first approach is effective when individual subunits themselves are stable and soluble in cells. Otherwise, the latter, the coexpression approach, is the only method of obtaining sufficient amounts of protein complexes for analyses. Unfortunately, most protein complexes in eukaryotes have subunits that are unstable and/or insoluble in isolation.

Techniques for recombinant protein production in the bacteria *Escherichia coli* can be rapidly and easily carried out. Several coexpression systems using *E. coli* have been developed. Insertion of several

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks, DOI 10.1007/978-4-431-56030-2_3, © Springer Japan 2016

genes of interests with their upstream ribosome binding sites (i.e., the Shine-Dalgarno sequence) between one promoter in the 5′-end and one terminator in the 3′-end enables polycistronic expression of proteins. Furthermore, cotransformation with two or three expression plasmids carrying different antibiotic resistance markers can be used for the coexpression. However, many eukaryotic protein complexes have subunits that are difficult to produce in *E. coli*, possibly owing to problems on transcription, translation, folding, and/or posttranslational modifications (e.g., phosphorylation and glycosylation), resulting in the failure of efficient production in *E. coli*.

An excellent solution to this difficult protein production in *E. coli* is the use of a baculovirus expression system, which employs an engineered baculovirus genome derived from the *Autographa californica* nuclear polyhedrosis virus (AcNPV) and appropriate transfer vectors (Fig. 1). Polyhedrin and p10 are products from the

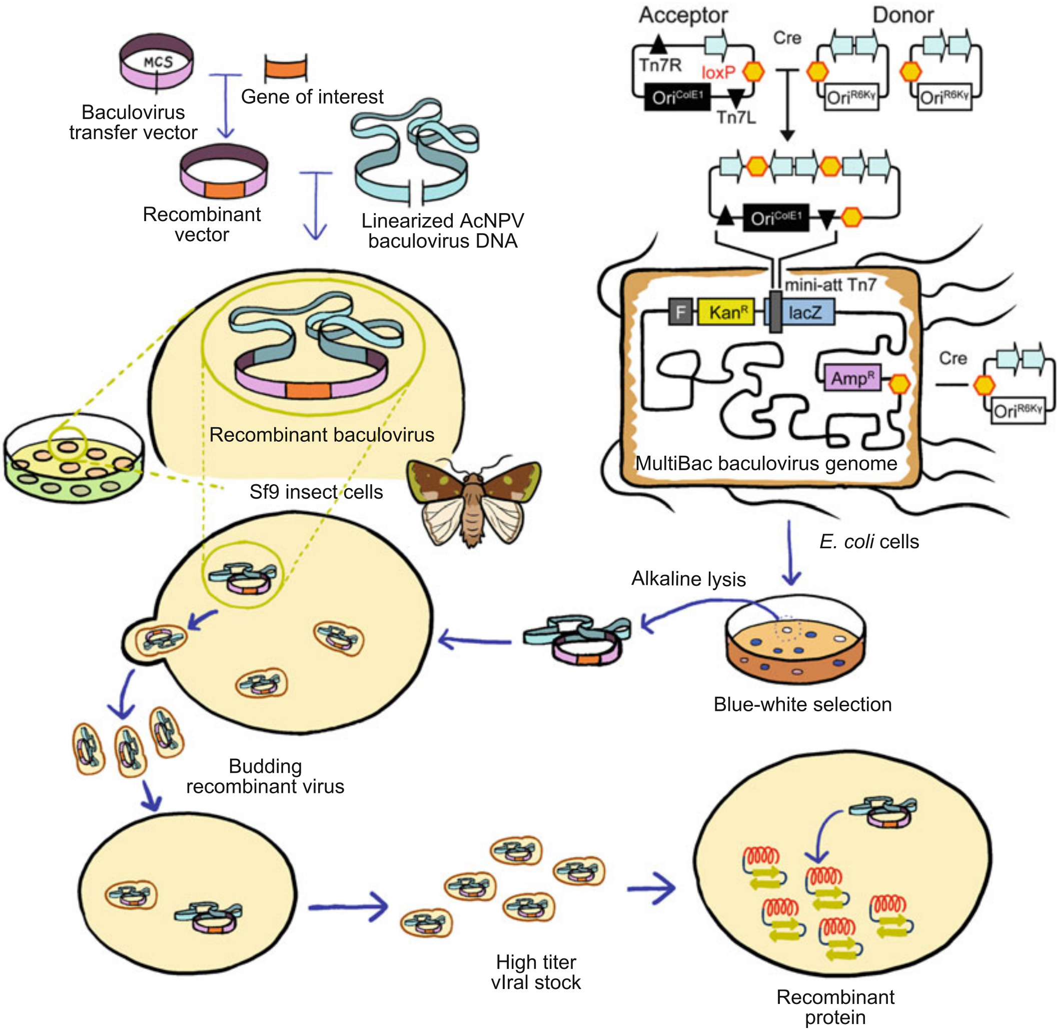


Fig. 1 Schematic drawing of MultiBac baculovirus-insect cell expression system

AcNPV genome and highly accumulated in the late stage of infection [1]. Their transcription promoters allow overproduction of proteins in insect cells. Genes of interest are first inserted between one of these strong transcription promoters (polyhedrin or p10 promoter) and a polyadenylation signal (SV40 or HSVtk polyadenylation signal) in the transfer vector and then integrated into the engineered AcNPV genome. The engineered AcNPV genome containing genes of interest is transfected to cells of the caterpillar *Spodoptera frugiperda* Sf9 or Sf21 cell line. The transfected Sf cells produce viruses that can infect insect cells to overexpress genes of interest. Suspension culture of the transfected Sf cells in shaker flasks is convenient and effective for large-scale protein production.

The genes of interest can be integrated to the engineered AcNPV genome in either insect cells or *E. coli* through the transfer vector. In the first case, both the linearized AcNPV genome containing a lethal deletion and the transfer vector are cotransfected into insect cells, and recombination occurs between the engineered AcNPV genome and the transfer vector. Because the transfer vector can rescue the lethal deletion of the engineered AcNPV genome, viruses that are generated from the genome fused with the transfer vector propagate. However, practically, this strategy might generate nonrecombinant viruses, which could apparently decrease virus titer (infection efficiency) and gene expression level. In such a case, a virus species with higher titer should be isolated from a single plaque by a plaque assay, which may require additional time and effort. This step can be bypassed when genes of interest are integrated to the engineered AcNPV genome in *E. coli* cells, applying the transposing reaction of the Tn7 transposon [2].

The Tn7 transposon is a relatively large DNA segment (14 kb) from the Tn7 phage, which can be integrated into a specific position called the Tn7 attachment site (attTn7) in the *E. coli* genome with high frequency. Tn7 transposon encodes a Tn7 transposase complex, which catalyzes the insertion of DNA elements containing the specific sequences Tn7L and Tn7R into the attTn7 site of another DNA molecule. As the transfer vector containing Tn7L and Tn7R is introduced to specific *E. coli* strains with a copy of the engineered AcNPV genome containing attTn7 and a helper plasmid encoding the Tn7 transposase complex, the transfer vector is transposed to the attTn7 site of the engineered AcNPV genome in the *E. coli* cells. Because the attTn7 site in the engineered AcNPV genome is located in the middle of *lacZα*, positive clones (i.e., *E. coli* cells containing the transposed AcNPV genome) are easily selected by standard blue-white selection in the presence of X-Gal and IPTG. The integrated AcNPV genome can then be isolated from the positive clones by a standard alkaline lysis protocol and used for the subsequent transfection to *S. frugiperda* cells.

In the baculovirus-insect cell system, coinfection by several viruses is one of two options for multigene expression: viruses expressing each subunit of the complex are first prepared and then simultaneously applied to infect *S. frugiperda* cells for coexpression of the appropriate subunit combination. However, this strategy can only be used under specific conditions because it is difficult to achieve a uniform expression of individual subunits owing to the differences in virus titer and gene expression level. Differences in the expression level among individual subunits may result in the failure of homogeneous complex formation. Another versatile option is multigene incorporation to a single AcNPV genome to generate a single species of virus that can produce multisubunit proteins. The MultiBac expression system was designed and developed for this purpose by Berger and colleagues [3–5].

Multigene incorporation to a single AcNPV genome requires a transfer vector containing a set of genes of interest. The MultiBac system enables two distinct methods for easier construction of the transfer vector for this purpose: one is the use of a multiplication module (Fig. 2) and the other is recombinase-mediated vector fusion (Fig. 3).

Transfer vectors for the “first- and second-generation (old)” MultiBac systems (i.e., pKL, pFL, pUCDM, pSPL, pFBDM, and pKDM) [3, 5] (Fig. 2a) have two multicloning sites, flanked by the restriction sites *Pme*I and *Avr*II. In addition, between these two multicloning sites, there is the multiplication module M to be digested by the restriction enzymes *Bst*Z17I and *Spe*I, which generate cohesive ends compatible with those generated from *Pme*I and *Avr*II, respectively. Therefore, a multigene expression cassette withdrawn from one transfer vector by *Pme*I-*Avr*II digestion can easily be incorporated into the *Bst*Z17I-*Spe*I-digested multiplication module M in another transfer vector (Fig. 3a). Iteration of the incorporation using the multiplication module can multiply expression cassettes in the transfer vector. Similarly, in the “third-

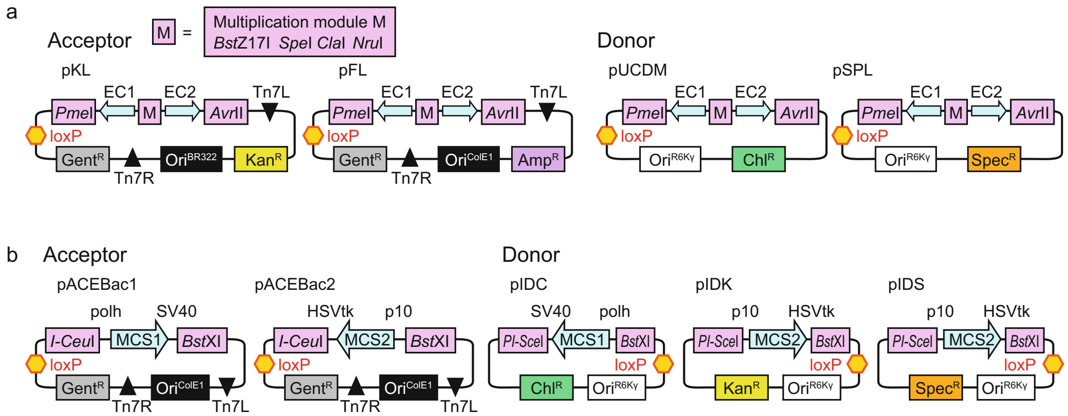


Fig. 2 Transfer vectors for MultiBac system. (**a**) Old set of transfer vectors. (**b**) New set of transfer vectors

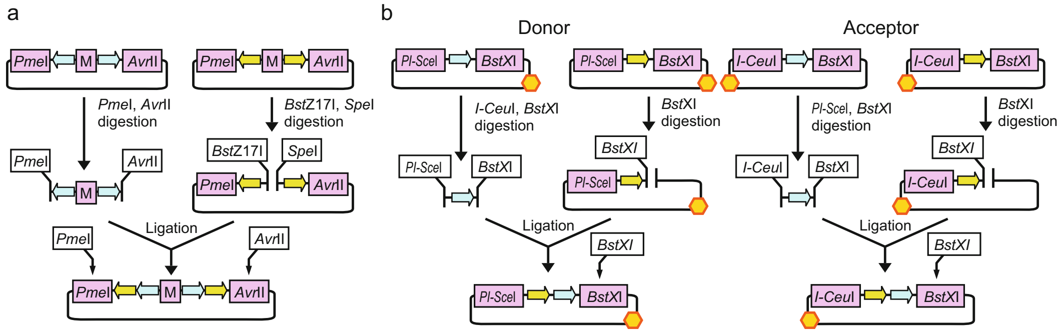


Fig. 3 Concept of multiplication module. The expression cassette can be excised from one vector by restriction digestion and transferred to another transfer vector. (**a**) Multiplication for old set of transfer vectors. (**b**) Multiplication for new set of transfer vectors

generation (new)" MultiBac system [4] (http://www.epigenesys.eu/images/stories/protocols/pdf/20120313121202_p54.pdf), transfer vectors (i.e., pACEBac1, pACEBac2, pIDC, pIDK, and pIDS) (Fig. 2b) have one multicloning site flanked by a homing endonuclease site (I-CeuI for pACEBac1 and pACEBac2 or PI-SceI for pIDC, pIDK, and pIDS) and a compatible *Bst*XI site, which enable iterative incorporation of expression cassettes to the transfer vector, generating a multigene transfer vector (Fig. 3b).

Multiplication of an expression cassette using the multiplication module should enable the incorporation of unlimited numbers of genes to the transfer vector in principle. However, practically, iterative incorporation of the expression cassette generates DNA plasmids that are too large to handle. To address this issue, for the MultiBac system, an additional concept is applied; that is, a number of transfer vectors can be fused by recombination to generate a multigene transfer vector that contains larger numbers of genes (Fig. 4).

Transfer vectors used for the MultiBac system are classified into two groups: "acceptors" and "donors" (Fig. 2). Both acceptors and donors have a short imperfect inverted repeat (LoxP) for the recombination reaction catalyzed by Cre recombinase [6]. Only acceptors contain Tn7L and Tn7R for integration to the engineered AcNPV genome. Acceptors have a ColE1 DNA replication origin and can propagate in standard *E. coli* cloning strains, whereas donors have an R6Kγ DNA replication origin and require a *pir* gene product for their propagation. Therefore, donors can be retained in *pir*-negative *E. coli* strains only when they are fused with acceptors. Propagation of donors in *pir*-negative *E. coli* strains indicates that the donors are appropriately fused with acceptors. This is an important feature because *E. coli* cells can retain two or more independent (i.e., unfused) plasmids with distinct antibiotic markers simultaneously in the presence of the antibiotics. One acceptor can be fused with one or two donors by in vitro Cre recombinase reaction

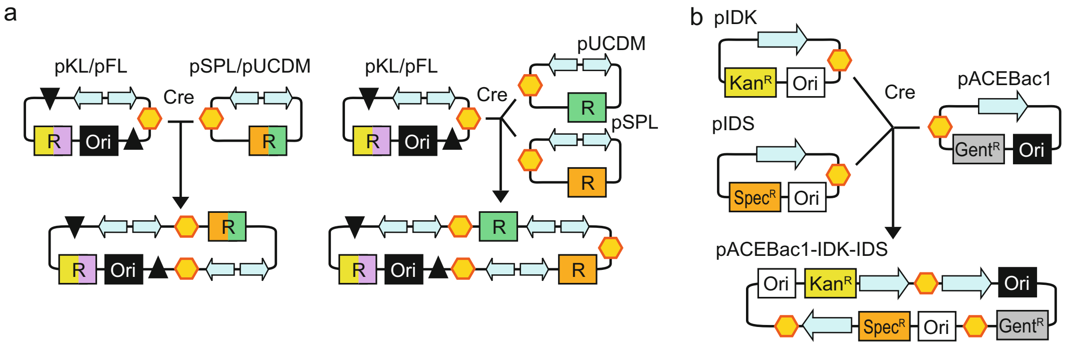


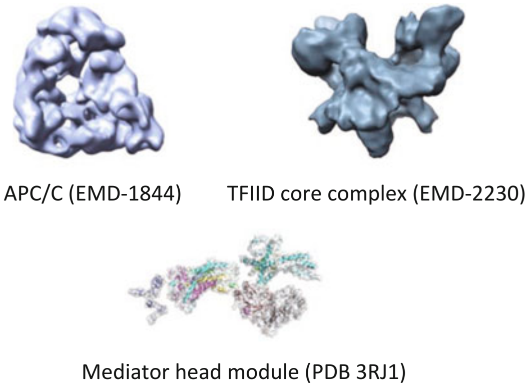
Fig. 4 Concept of recombination-mediated multiplication of expression modules. Cre-mediated recombination allows fusion of two to four transfer vectors. (**a**) Cre-mediated fusion for old set of transfer vectors. (**b**) Cre-mediated fusion for new set of transfer vectors

in one step, followed by transformation of standard *pir*-negative *E. coli* strains with the resultant multigene transfer vector in the presence of the appropriate combination of antibiotics. In the new MultiBac system, the Cre-mediated fusion with three donors can be practically performed in two steps.

The MultiBac system also applies new technologies for the engineered AcNPV genome, where genes encoding viral protease and apoptotic activities are deleted to avoid protein degradation and delay lysis of the infected insect cells. Furthermore, in addition to the attTn7 site, the MultiBac AcNPV genome contains the LoxP site, which is useful for additional functionalities. For example, when a yellow fluorescence protein (YFP) gene is integrated to the LoxP site of the MultiBac AcNPV genome, the produced YFP provides information about protein production and virus performance through its fluorescence almost in real time. This YFP-integrated MultiBac AcNPV genome (EMBacY) is included in the new MultiBac system kit.

Many of the recent structural studies of biologically important eukaryotic multisubunit complexes utilize the MultiBac expression system for their production [7–9] (Fig. 4). All of such complexes were challenging targets of structural studies because of their large size and/or complicated subunit composition.

The anaphase-promoting complex (APC/C) is a cell cycle regulator and composed of 13 subunits with a total molecular weight of ~1.1 MDa. The gene assembly encoding the complete APC/C was inserted into two MultiBac baculoviruses, one encoding eight subunits and the other five. Coinfection by two viruses enables the production of the entire 1.1 MDa APC/C complex. The purified complex was subjected to electron microscopy analysis, revealing the complicated subunit architecture of this huge complex [9] (Fig. 5).



APC/C (EMD-1844)

TFIID core complex (EMD-2230)

Mediator head module (PDB 3RJ1)

Fig. 5 Three-dimensional structures of large multisubunit complexes analyzed using MultiBac expression system. For APC/C, two subcomplexes (named TPR5 and SC8) were also analyzed besides the entire APC/C

The transcription mediator controls transcription through its interactions with RNA polymerase and transcriptional activators. The mediator consists of 25 or more subunits with a total molecular weight of ~1.2 MDa. The head module of the yeast mediator (seven subunits, Mw 223 kDa) was produced using the MultiBac system and subjected to cryo-EM analysis and X-ray crystallography, revealing the architecture and dynamics at the atomic level [8].

A general transcription factor, TFIID, binds gene promoters and regulates the initiation event of transcription. TFIID is composed of the TATA-box-binding protein (TBP) and 13 TBP-associated factors (TAFs) with a total molecular weight of over 1 MDa. The TFIID core complex (five subunits, Mw 650 kDa) was produced using the MultiBac system and subjected to cryo-EM analysis, revealing its subunit stoichiometry and architecture [7]. For the production of this TFIID core complex, an additional technology was applied for the uniform expression of the individual subunits. Coexpression of the TFIID core subunits by the original MultiBac system showed imbalanced production of the individual subunits, which hampered the purification of the complexes. This imbalanced expression problem was solved by a polyprotein strategy, where a number of proteins are encoded in one large open reading frame (ORF) and generated by proteolysis with a highly specific protease in a manner similar to protein production by RNA viruses such as the coronavirus. The TAF-encoding genes are concatenated into a single ORF, spaced by cleavage sites for a protease NIa from the tobacco etch virus (TEV). The TEV protease gene precedes the TAF genes in the ORF. A cyan fluorescent protein (CFP) gene is also inserted in the 3′ end of the ORF for checking the expression of all conjoined proteins.

These structural studies of large protein complexes including those with molecular weights of over 1 MDa demonstrate that the

MultiBac system is a powerful tool for the overproduction of challenging multisubunit protein complexes. Higher-resolution analysis of complex structures requires more homogeneous samples in terms of size, composition, and posttranslational modification and conformation, which typically require the removal of specific subunits and/or modification of the individual subunits by site-directed mutations and/or trimming regions that are predicted or experimentally shown to be flexible or disordered. The modular concept of the MultiBac system is highly compatible with this optimization process. It has been announced that the MultiBac system is still under development and that new technologies are planned to be included. This powerful tool for the production of multiprotein complexes will further accelerate the study of the structural biology of challenging targets to elucidate more complex biological processes.

2 Materials

2.1 Bacteria Work

Escherichia coli strains: DH10Bac, DH10MultiBac, a standard *pir*-negative strain (e.g., DH5α), *PirHC, *PirLC (see Note 1)

Transfer vectors: pKL, pFL, pUCDM, pSPL (old set), pACEBac1, pACEBac2, pIDC, pIDK, pIDS (new set)

Antibiotics (1000×): 50 g/L kanamycin, 10 g/L tetracycline, 10 g/L gentamicin, 50 g/L ampicillin, 30 g/L chloramphenicol

Enzymes: Cre recombinase (NEB), high-fidelity DNA polymerase (Toyobo KOD plus neo), in-fusion cloning kit (Clontech)

Medium: Luria broth (LB) medium (Nacalai Tesque)

Equipment: Shaker incubator (temperature controlled at 37 °C; INNOVA 44R, TAITEC BR-23FP•MR), electroporator (BioRad MicroPulser), toothpick

Chemicals: X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside), IPTG

2.2 Insect Cell Work

Insect cell lines: Sf9, Sf21

Antibiotics (200×): Penicillin-streptomycin mixed solution (Nacalai Tesque)

Medium: Sf-900II SFM serum-free medium (Invitrogen), fetal bovine serum (Gibco)

Transfection reagent: X-Treme GENE HP transfection reagent (Roche)

Flasks: Erlenmeyer flasks (125, 250, 500, 1000, 2000 mL; Corning)

Biological safety hood with UV illumination

Six-well (35-mm-diameter) tissue culture plates (Falcon)

Shaker incubator (temperature controlled at 27 °C; TAITEC G•BR200)

3 Methods

3.1 Cloning of Genes of Interest for Construction of MultiBac Transfer Vectors

The old set of MultiBac plasmids contains two multicloning regions, whereas the new set contains a single multicloning region. Conventional cloning methods using restriction enzymes and ligases are applicable. However, sequence- and ligation-independent cloning (SLIC) methods [10] are highly convenient for large DNA insertions, which typically have multiple restriction sites. Therefore, we use SLIC for the construction of MultiBac transfer vectors. In-fusion cloning kits (Clonetech) are commercially available. The Clonetech website (https://www.takara-bio.co.jp/infusion_primer/) is highly convenient for the design of PCR primers for SLIC.

1. Linearize 3–5 μg of vectors (for ten or less SLIC reactions) by either PCR or restriction enzyme digestion and purify them by agarose gel electrophoresis (see Note 2).
2. Set up SLIC reaction (10 μL for each), as shown in Table 1.
3. Incubate the reaction mixture at 50 °C for 15–20 min and stop the reaction by placing the mixture on ice.
4. Transform chemically competent *E. coli* cells with 1 μL of the reaction. Typical *E. coli* cloning strains (e.g., DH5α) are available for vectors harboring the ColE1 or pBR322 replication origin (i.e., pFL, pKL, pACEBac1, and pACEBac2), whereas a special strain (PirHC or PirLC) is required for vectors harboring the *pir* replication origin (i.e., pUCDM, pSPL, pIDC, pIDK, and pIDS) (see Table 2).
5. Plate the transformed cells on LB agar plates containing the appropriate antibiotics for selection (see Table 2) and incubate them at 37 °C for 12–15 h.

Table 1
SLIC reaction

5 X reaction mix (enzyme included)	2 μL
PCR-amplified insert	~50 ng
Linearized vector	~50 ng
Pure water	to 10 μL

Table 2
Antibiotics and host strains of transfer vectors for selection

Vectors	Antibiotics	Host strains
pFL	Ampicillin, gentamicin	Standard
pKL	Kanamycin, gentamicin	Standard
pACEBac1, pACEBac2	Gentamicin	Standard
pUCDM, pIDC	Chloramphenicol	PirHC, PirLC
pSPL, pIDS	Spectinomycin	PirHC, PirLC
pIDK	Kanamycin	PirHC, PirLC

6. Select positive colonies by PCR analysis. A premixed PCR solution (e.g., Promega GoTaq Master Mix) is convenient for this purpose. Each colony is picked with a sterilized toothpick, and the toothpick is briefly dipped into a tube containing a PCR reaction mix. The reaction products are separated by agarose gel electrophoresis. We use the 5′ and 3′ primers derived from the insert and vector sequences, respectively, to confirm that the insert is integrated into the vector. Typically, four to eight colonies are sufficient to obtain two or more positive clones.
7. Isolate vectors from positive clones using a commercially available mini-prep kit (e.g., Promega Wizard Plus SV Minipreps DNA Purification System) and verify the nucleotide sequences of the insert by DNA sequencing. Promoter and terminator regions should also be verified when vectors are linearized by PCR in Step 1.
8. Store the verified vectors at −20 °C until use.

3.2 Multiplication of Expression Cassettes

Multigenes can be assembled in a single vector by using a multiplication module (Fig. 3). Genes inserted into the old vectors (i.e., pKL, pFL, pUCDM, and pSPL) can be excised as a cassette by restriction enzyme digestion with *Pme*I and *Avr*II. This cassette can be transferred to another vector digested with *Bst*Z17I and *Spe*I. Similarly, genes inserted into the new vectors (i.e., pACEBac1, pACEBac2, pIDC, pIDK, and pIDS) can be excised by *Bst*XI digestion and either I-*Ceu*I or PI-*Sce*I digestion and transferred to another vector digested with *Bst*XI within either of the donor vectors or acceptor vectors.

1. Prepare the cassette and linearized vector by restriction enzyme digestion of 2–3 μg of vectors containing gene(s) of interest and purify them by agarose gel electrophoresis before ligation (see Note 3).

2. Set up the ligation reaction with the purified cassette and linearized vector.
3. Transform the appropriate *E. coli* strains with 1 μL of the ligated reaction product. The efficiency of transformation may decrease if the ligated vector is longer than 10 kb. Therefore, we use electrocompetent cells for longer vectors, instead of chemically competent cells. For electroporation, we use a MicroPulser (BioRad) with the conditions preset for *E. coli*.
4. Select positive clones as in Step 3.1.6, except that the 5′ and 3′ primers derived from the insert are used for colony PCR analysis because the upstream and downstream sequences are the same in the cassettes.
5. Isolate the vector from the positive clone using a commercially available mini-prep kit (e.g., Promega Wizard Plus SV Minipreps DNA Purification System) and store it at −20 °C until use.

3.3 Cre-mediated Fusion

Cre-mediated fusion is also available for generating multigene transfer vectors. In both the new and old systems, it is guaranteed that one acceptor vector can be fused with one or two donor vectors. Furthermore, fusion between one acceptor and three donors is possible in the new system.

1. Set up the Cre reaction (10 μL for each), as shown in Table 3, and incubate the reaction mixture at 37 °C for 30 min to 1 h (see Note 4). Optionally, the reaction can be stopped by heating at 65 °C for 5 min. To avoid the integration of more than one acceptor, the amount of the acceptor vector should be lower than those of the donor vectors.
2. Transform the *pir*-negative *E. coli* strain (e.g., DH5α) with 1 μL of the Cre reaction mixture. Cre fusion generates longer vectors, particularly in the old system. Therefore, we use electrocompetent cells for transformation with the Cre-fused vectors.
3. Select positive clones as in Step 3.1.6.

Table 3
Cre reaction

10 X Cre reaction buffer (NEB)	1 μL
Acceptor vector	~400 ng
Donor vector(s)	~500 ng each
Cre recombinase (NEB)	1 μL
Pure water	to 10 μL

4. Isolate the vector from the positive clones using a commercially available mini-prep kit (e.g., Promega Wizard Plus SV Minipreps DNA Purification System).
5. Confirm that all genes are present in the vector by PCR analysis after the vector isolation (see Note 5) and store the vector at −20 °C until use.

3.4 Preparation of Multigene-Integrated Engineered AcNPV Genome

Multigene transfer vectors constructed in Steps 3.2 and/or 3.3 are integrated into the engineered AcNPV genome retained in special *E. coli* strains such as DH10Bac and DH10MultiBac (see Note 6). DH10MultiBac is deficient in protease and chitinase to reduce proteolysis and extend cell viability. For the transformation, we use electrocompetent cells (see Note 7).

1. Prepare electrocompetent DH10Bac or DH10MultiBac cells that retain the engineered AcNPV genome and helper plasmid.
   (a) Inoculate 0.5 L of LB medium containing 50 mg/L kanamycin and 10 mg/L tetracycline in a 2 L flask with 5 mL of a fresh overnight culture in the same medium. Prepare 1 L of sterilized deionized water and 5 mL of sterilized 10 % glycerol and keep them in a refrigerator or a cold room.
   (b) Grow cells at 37 °C with shaking to OD_{600} of 0.5–0.8.
   (c) Chill the flask on ice for 15–30 min. Keep the cells as close to 0 °C as possible in the steps below.
   (d) Centrifuge the culture in a cold rotor at 4000 × *g* for 15 min.
   (e) Remove as much of the supernatant as possible. Do not be concerned about the loss of a few cells while removing the supernatant.
   (f) Gently suspend the obtained cell pellet in 500 mL of the cold sterilized deionized water from Step (a).
   (g) Centrifuge the suspension as in Step (d) and remove the supernatant as in Step (e).
   (h) Repeat Steps (f) and (g).
   (i) Gently suspend the obtained cell pellet in 10 mL of the cold sterilized 10 % glycerol from Step (a).
   (j) Repeat Step (g).
   (k) Gently suspend the cell pellet in 2 mL of the cold sterilized 10 % glycerol from Step (a).
   (l) Flash-freeze this suspension in 50–200 μL aliquots in liquid N_2 and store them at −80 °C until use.
2. Mix 1 μL of 100–500 ng/μL multigene transfer plasmid with 50 μL of electrocompetent cells and incubate it on ice for 5 min.

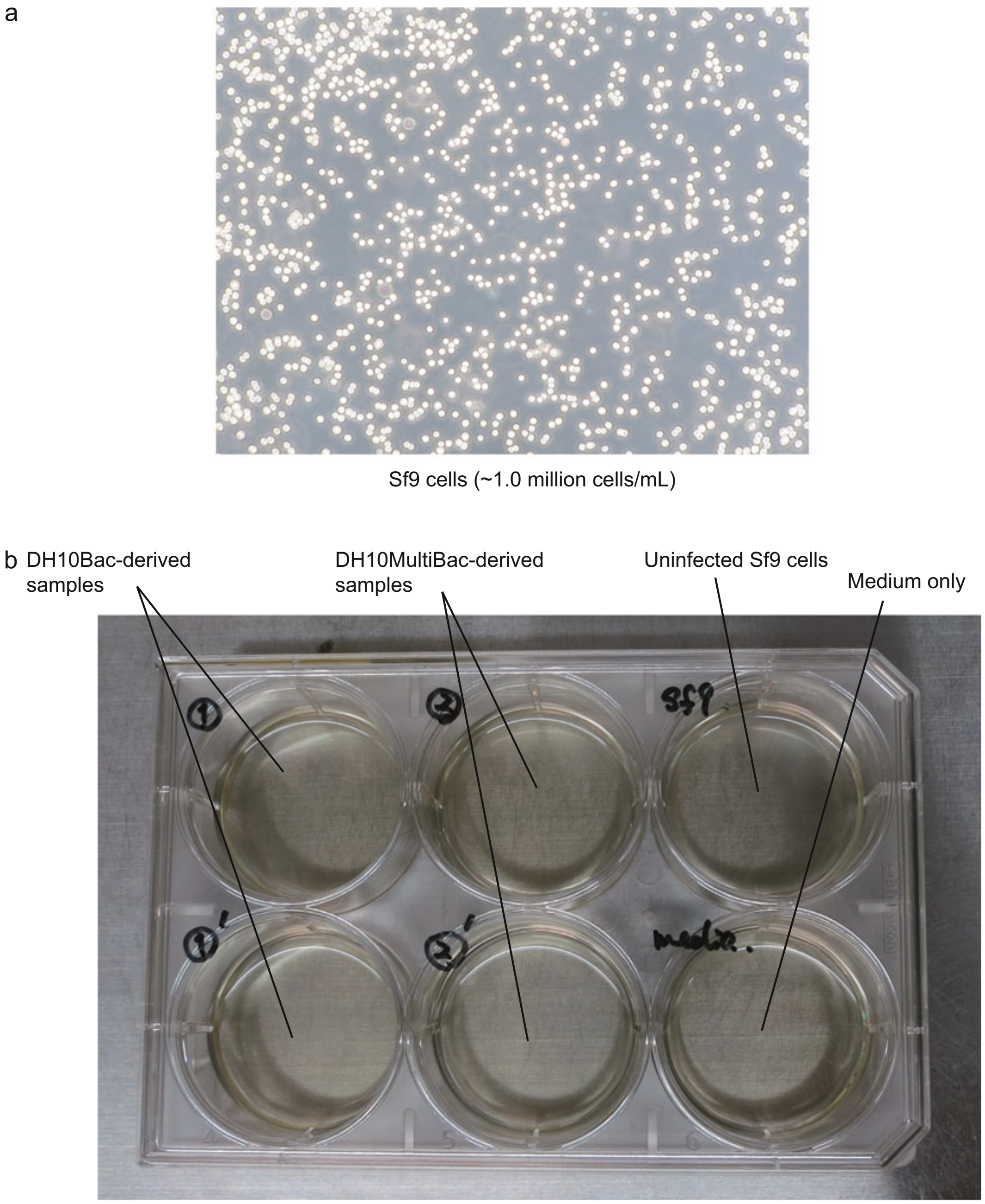
3. Transfer the cells to an electroporation cuvette and electroporate them using the appropriate electroporation equipment (e.g., MicroPulsor, BioRad).
4. Suspend the cells with 500 μL of LB medium (or richer medium such as SOC) in the cuvette and transfer them to a 1.5 mL tube.
5. Incubate the cells at 37 °C with shaking for 6 h.
6. Transfer 10 μL of the culture to 1 mL of LB medium in a new tube. Then, 10 μL of the diluted culture is transferred to 90 μL of LB medium in another new tube.
7. Streak 100 μL of these two diluted cultures on LB plates containing 50 mg/L kanamycin, 7 mg/L gentamicin, 10 mg/L tetracycline, X-Gal (or an equivalent indicator), and IPTG. We supply 50 μL of 2 % X-Gal and 25 μL of 0.2 M IPTG for each antibiotic-containing plate just before streaking the cells. The culture dilution is important for optimal colony separation.
8. After the incubation at 37 °C for 2 days, larger white colonies appear if the genes in the transfer plasmid are successfully integrated to the engineered AcNPV genome in *E. coli* cells. Restreaking on a fresh plate is recommended to confirm the white phenotype.
9. Pick a white colony for each construct and inoculate it to 5 mL of LB medium supplemented with 50 mg/L kanamycin, 7 mg/L gentamicin, and 10 mg/L tetracycline in a culture tube.
10. Incubate the culture with shaking at 37 °C for 15–17 h.
11. Collect the cells by centrifugation and suspend them in 300 μL of Solution 1 (15 mM Tris-Cl, pH 8.0, 10 mM EDTA, 100 mg/L RNase A, see Note 8).
12. Add 300 μL of Solution 2 (0.2 N NaOH, 1 % SDS, see Note 8) and gently mix the solution by inverting the tube upside down several times, followed by incubation at room temperature for 5 min.
13. Slowly add 300 μL of Solution 3 (3 M potassium acetate, pH 5.5, see Note 8) and gently mix the solution by inverting the tube upside down several times, followed by incubation on ice for 5 min. A thick white precipitate of *E. coli* proteins and genomic DNA appears.
14. Clear the solution by centrifugation at 14,000 × *g* for 10 min. During the centrifugation, label another fresh 2 mL tube and add 0.8 mL of 2-propanol to it.
15. Carefully transfer the supernatant to the 2 mL tube containing 2-propanol to avoid contamination of the white precipitate as much as possible and mix it gently by inverting the tube upside down a few times.

16. Place the tube on ice for 10 min and centrifuge it at 14,000 × *g* for 10 min at room temperature. The multigene-integrated bacmid is precipitated as a translucent pellet in this step.
17. Remove the supernatant and add 500 μL of 70 % ethanol.
18. Centrifuge the tube at 14,000 × *g* for 5 min at 4 °C or room temperature .
19. Remove the supernatant as much as possible and air-dry the pellet (the integrated AcNPV genome) in a sterile hood to avoid contamination of microorganisms in the transfected cells.
20. Store the dried pellet at −20 °C until use (Note 9).

3.5 Initial Virus Preparation (P1)

The following steps should be performed in a sterile hood to avoid contamination of insect cell culture:

1. Add 30 μL of sterile pure water to the tube containing the isolated AcNPV genome pellet.
2. Dissolve the pellet by gently tapping the tube and incubate the tube at room temperature for 5–10 min to completely dissolve the pellet.
3. Add 200 μL of an antibiotic-free insect cell medium (e.g., Grace's insect cell medium) to the tube containing the bacmid solution (Tube A).
4. Add 100 μL of an antibiotic-free insect cell medium to another fresh tube (Tube B).
5. Add 8 μL of the transfection reagent, X-Treme GENE HP (Roche), to Tube B.
6. Transfer the mixture containing the transfection reagent in Tube B to Tube A containing the isolated AcNPV genome, followed by incubation at room temperature for 15–30 min.
7. During the incubation, seed 0.5–1.0 million cells to each well of a 6-well tissue culture plate (see Note 10) and add a supplemented insect cell medium [containing antibiotics and/or 4 % fetal bovine serum (FBS)]) with a final volume of 3 mL for each well (Fig. 6).
8. Add one-half of the mixture containing the isolated AcNPV genome and transfection reagent dropwise to each well.
9. Incubate the plate at 27 °C for 60–72 h.
10. Collect the supernatant from the wells and transfer it to sterile 15 mL tubes (P1 virus).
11. Store the P1 virus at 4 °C (or −80 °C after flash-freezing in liquid N_2 for long-term storage) protected from light (see Note 11).


a
Sf9 cells (~1.0 million cells/mL)
b DH10Bac-derived samples
DH10MultiBac-derived samples
Uninfected Sf9 cells
Medium only
1
2
Sf9
1'
2'
medim.
4
5
6

Six-well plate for P1 virus preparation

Fig. 6 Sf9 cells and P1 virus preparation. (**a**) Sf9 cells at a density of ~1.0 million/mL. (**b**) Six-well plate for P1 virus preparation

3.6 Virus Amplification (P2) and Expression Check

1. Prepare 9 mL of cell culture (containing 4 % FBS and antibiotics) at a density of 0.5–1.0 million cells/mL in 125 or 250 mL Erlenmeyer flasks (see Note 12). For virus amplification, we use flasks whose volumes are 10- to 25-fold larger than the culture volume.
2. Inoculate 1 mL of the P1 virus to a 9 mL cell culture for infection.
3. Culture the infected cells at 27 °C for 60–72 h with shaking. We recommend the monitoring of cell growth every 12–24 h. We suggest that cell density should be lower than 1.5 million cells/mL to avoid insufficient aeration. When the density is over 1.5 million cells/mL, dilute the culture to a density of 0.5–1.0 million cells/mL, keeping the culture volume at 1/10–1/25 of the vessel.
4. Centrifuge the culture for 5 min at 2000 × *g* and collect the supernatant (P2 virus). The cell pellet after the centrifugation is used in Step 6.
5. Store the P2 virus at 4 °C (or −80 °C after flash-freezing in liquid N_2 for long-term storage) protected from light (see Note 13).
6. Check the protein expression using the cell pellet collected by centrifugation (see Note 14).
   (a) Suspend the cell pellet in 300 μL of an appropriate buffer for tag-affinity beads (e.g., 50 mM Tris-Cl, pH 8.0, containing 150 mM NaCl and 20 mM imidazole for Ni-chelating beads) and transfer it to a 1.5 mL tube.
   (b) Disrupt the cells by sonication (e.g., Branson Sonifier 450A equipped with a microtip) for 15 s on ice. Take 5 μL as the whole extract for SDS-PAGE analysis (Sample A).
   (c) Centrifuge the disrupted cell suspension at 20,000 × *g* for 15 min at 4 °C. During the centrifugation, pre-equilibrate 15 μL of affinity beads with a sonication buffer in a fresh 1.5 mL tube, following the manufacturer's instruction.
   (d) Take 5 μL of the supernatant as the soluble extract (Sample B) and pick a very small amount of the resulting pellet with a micropipette tip as the insoluble extract (Sample C) for SDS-PAGE analysis.
   (e) Transfer the remaining supernatant to the fresh 1.5 mL tube containing pre-equilibrated affinity beads.
   (f) Incubate the sample for 0.5–1 h at 4 °C.
   (g) Centrifuge the sample at 500 × *g* for 3 min and remove the supernatant.

(h) Wash the beads with 500–1000 μL of a sonication buffer (or a more stringent buffer of choice) three times by iteration of buffer addition, centrifugation, and buffer removal. Take 5 μL of each wash solution as the wash samples (Samples D–F) for SDS-PAGE analysis.

(i) Add 20 μL of an SDS-PAGE loading buffer as the purified sample (Sample G) for SDS-PAGE analysis.

(j) Analyze Samples A–G by SDS-PAGE with standard Coomassie brilliant blue and/or immunostaining.

3.7 Large-Scale Protein Production

For large-scale protein production, we use 2 L Erlenmeyer flasks containing 400–500 mL of cell culture (see Note 15). For example, two flasks are used for 0.8–1 L cell culture. The number of flasks depends on the expression level of your target protein. We use a fresh virus preparation propagated from the stock P2 or P3 virus for large-scale protein production.

1. Inoculate 250 μL of the stock P2 or P3 virus to 25 mL of cell culture (containing antibiotics and 4 % FBS) for infection at a density of 0.5–1 million cells/mL in 0.5 L flask to obtain fresh virus preparations for large-scale expression.
2. Culture the infected cells at 27 °C for 60–72 h.
3. Seed cells to 400–500 mL media containing antibiotics and 4 % FBS at a final density of 0.3 million cells/mL for each flask 1 day after Step 1.
4. Add 25 mL of the infected cell culture (from Steps 1–2) to a fresh 400–500 mL cell culture at a density of one to two million cells/mL for each flask (from Step 3).
5. Culture the infected cells from Step 4 at 27 °C for 60–72 h (see Note 16).
6. Collect the cells by centrifugation at 2000 × *g*.
7. Flash-freeze the cells in liquid N_2 and store them at −80 °C until use.

4 Notes

1. PirHC and PirLC are strains required for preparation of donor plasmids (i.e., pIDC, pIDK, pIDS, pUCDM, pSPL). These two strains contain the *pir* gene in their genomes and can propagate plasmids that have *pir* replication origins. LC and HC mean low copy and high copy, respectively. PirLC strains are used when inserted genes make the propagation of donor plasmids difficult in PirHC.
2. We favor restriction enzyme digestion for the linearization of vectors for SLIC in order to avoid PCR-associated mutations.

3. The uncut original circular vector sometimes migrates similarly to the cassette and cannot be removed by gel extraction. To avoid this problem, we recommend using vectors with two different antibiotic resistance makers for the cassette-mediated multiplication (e.g., the cassette from pUCDM is transferred to pFL).

4. Although we never tried fusion with more than two donors, an online manual for the new system describes in detail the protocol for fusion with more than two donors. Following such a protocol, the Cre reaction is performed at 30 °C for 1–2 h with ~500 ng of each donor and ~400 ng of one acceptor (http://www.epigenesys.eu/images/stories/protocols/pdf/20120313121202_p54.pdf).

5. According to the online manual (http://www.epigenesys.eu/images/stories/protocols/pdf/20120313121202_p54.pdf), restriction analysis using appropriate restriction enzymes is highly recommended for confirming the presence of all genes in the vector.

6. We try both DH10Bac and DH10MultiBac and compare them.

7. Protocols for chemically competent cells are described in the online manual for the new system (http://www.epigenesys.eu/images/stories/protocols/pdf/20120313121202_p54.pdf).

8. Solutions 1, 2, and 3 are our own preparations, but solutions from commercially available plasmid mini-prep kits can be used for the bacmid isolation.

9. We quickly move to Step 3.5 after isolating the integrated AcNPV genome.

10. A 35 mm culture dish is available for this purpose. However, we recommend a 6-well plate because media in the culture dish seem to evaporate much faster than those in the 6-well plate.

11. We quickly move to Step 3.6 after preparing P1 virus.

12. We reuse plastic Erlenmeyer flasks supplied by Corning after autoclaving.

13. Baculoviruses can be stored in liquid N_2 in the form of baculovirus-infected insect cells (BIICs), which is widely adopted in many laboratories [11] (http://www.epigenesys.eu/images/stories/protocols/pdf/20120313121202_p54.pdf). This BIIC storage is recommended in terms of space saving in refrigerators and almost no loss of titer. Some viruses lose their transfection activity within a month or less when stored in the form of a virus solution at 4 °C, while others retain the activity for more than a 2–3 years.

14. Protein expression level may depend on incubation time. You can determine when cells stop proliferation after infection and analyze time-dependent changes in expression level by sampling one million cells every 12 or 24 h after the proliferation arrest.
15. For protein production, High Five cells (derived from the caterpillar *Trichoplusia ni*) may be more suitable than Sf cells in some cases.
16. Optionally, 1 day after infection, the infected cells are cultured at 20 °C for 72 h. This change in culture temperature increases the expression level and/or protein solubility in some cases.

Acknowledgments

This work was supported by grants from MEXT and JST CREST.

References

1. Smith GE, Vlak JM, Summers MD (1983) Physical analysis of autographa californica nuclear polyhedrosis virus transcripts for polyhedrin and 10,000-molecular-weight protein. J Virol 45:215–225
2. Craig NL (1991) Tn7: a target site-specific transposon. Mol Microbiol 5:2569–2573
3. Berger I, Fitzgerald DJ, Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. Nat Biotechnol 22:1583–1587
4. Bieniossek C, Imasaki T, Takagi Y, Berger I (2012) MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem Sci 37:49–57
5. Fitzgerald DJ, Berger P, Schaffitzel C, Yamada K, Richmond TJ, Berger I (2006) Protein complex expression by using multigene baculoviral vectors. Nat Methods 3:1021–1032
6. Abremski K, Hoess R, Sternberg N (1983) Studies on the properties of P1 site-specific recombination: evidence for topologically unlinked products following recombination. Cell 32:1301–1311
7. Bieniossek C, Papai G, Schaffitzel C, Garzoni F, Chaillet M, Scheer E, Papadopoulos P, Tora L, Schultz P, Berger I (2013) The architecture of human general transcription factor TFIID core complex. Nature 493:699–702
8. Imasaki T, Calero G, Cai G, Tsai KL, Yamada K, Cardelli F, Erdjument-Bromage H, Tempst P, Berger I, Kornberg GL, Asturias FJ, Kornberg RD, Takagi Y (2011) Architecture of the mediator head module. Nature 475:240–243
9. Schreiber A, Stengel F, Zhang Z, Enchev RI, Kong EH, Morris EP, Robinson CV, da Fonseca PC, Barford D (2011) Structural basis for the subunit assembly of the anaphase-promoting complex. Nature 470:227–232
10. Haffke M, Viola C, Nie Y, Berger I (2013) Tandem recombineering by SLIC cloning and Cre-LoxP fusion to generate multigene expression constructs for protein complex research. Methods Mol Biol 1073:131–140
11. Wasilko DJ, Lee SE (2006) TIPS: titerless infected-cells preservation and scale-up. Bioproc J 5:29–32

Chapter 4

Purification Using Affinity Tag Technology

Atsushi Furukawa, Katsumi Maenaka, and Takao Nomura

Abstract

Affinity tag technology is a prerequisite for high and rapid purification of recombinant proteins in structural studies because of specific interactions of tags. Widely used tags are polyhistidine tags specific to metal-chelating ligands and glutathione S-transferase tag for glutathione-immobilized ligands. Furthermore, tags binding to antibodies, such as FLAG, Fc, and HA, are also popular for protein preparation and, in addition, are utilized for biological and biochemical analyses, e.g., western blotting, immunoprecipitation, immunofluorescence assay, and flow cytometry. Some tags improve the solubility of proteins. In this chapter, we introduce the features of these representative tags and show several practical examples.

Keywords Affinity purification, Affinity tags, Solubility enhancement, Biologics

1 Introduction

Recombinant DNA technology is essential for large-scale protein preparation in structural studies. In the early days, purification of recombinant proteins was performed using traditional ion-exchange chromatography and gel filtration. However, the introduction of the hexa-histidine tag used for immobilized metal affinity chromatography (IMAC) dramatically changed the strategy for purification, because this tag confers easy and rapid purification in good yields and with high purity. Nowadays, many tags are commercially available and have additional advantages such as the improvement of the characteristics of the target proteins (Table 1). A schematic image of affinity purification methods are shown in Fig. 1. In this chapter, we describe the basic concepts of representative tags and detail the procedures for purification using some tags.

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks,
DOI 10.1007/978-4-431-56030-2_4, © Springer Japan 2016

Table 1
General tags that are widely used for expression and purification of proteins. In the "sequence" column, the terms "protein" and "compound" indicate a tag that is relatively huge and chemical modification of a specific sequence, respectively

Purification methods	Tag	Residues	Sequence
By tag-ligand affinity	His	6	HHHHHH
	GST	262	Protein
	MBP	375	Protein
	Biotin		Compound
	Strep-Tag	8	WSHPQFEK
By antibody	FLAG	8	DYKDDDDK
	Myc	11	EQKLISEEDL
	Fc	ca. 250	Protein
	1D4	9	TETSQVAPA
	HA	9	YPYDVPDYA
	GFP	238	Protein
None	SUMO	75–100	Protein

2 Polyhistidine Tag

IMAC, introduced by Porath et al., is one of the most powerful methods for protein purification [36]. IMAC is based on the coordination interaction between transition metal ions immobilized on a resin and cationic amino acids in proteins. Transition metal ions, such as Cu^{2+}, Co^{2+}, Ni^{2+}, Zn^{2+}, and Fe^{3+}, are immobilized on an agarose, sepharose, or silica gel resin through spacer ligands, such as N,N,N′-tris-(carboxymethyl)-ethylenediamine, nitrilotriacetic acid (NTA), and iminodiacetic acid (IDA). Ni ions and NTA chelator are the most commonly used. The main amino acid interacting with metal ions is histidine. The imidazole ring in histidine serves as an electron donor and can form a coordination bond with a transition metal ion immobilized on the resin. Using genetic engineering, the polyhistidine tag can be added to the target protein to enable strong and specific binding to transition metal ions. The side-chain imidazole ring of the histidine residues interacting with the metal ions can be replaced with imidazole during the elution. The first use of the polyhistidine tag was developed by Hochuli et al. [16]. Terpe investigated the effect of the tag length, ranging from two to ten residues, and demonstrated that the hexahistidine tag

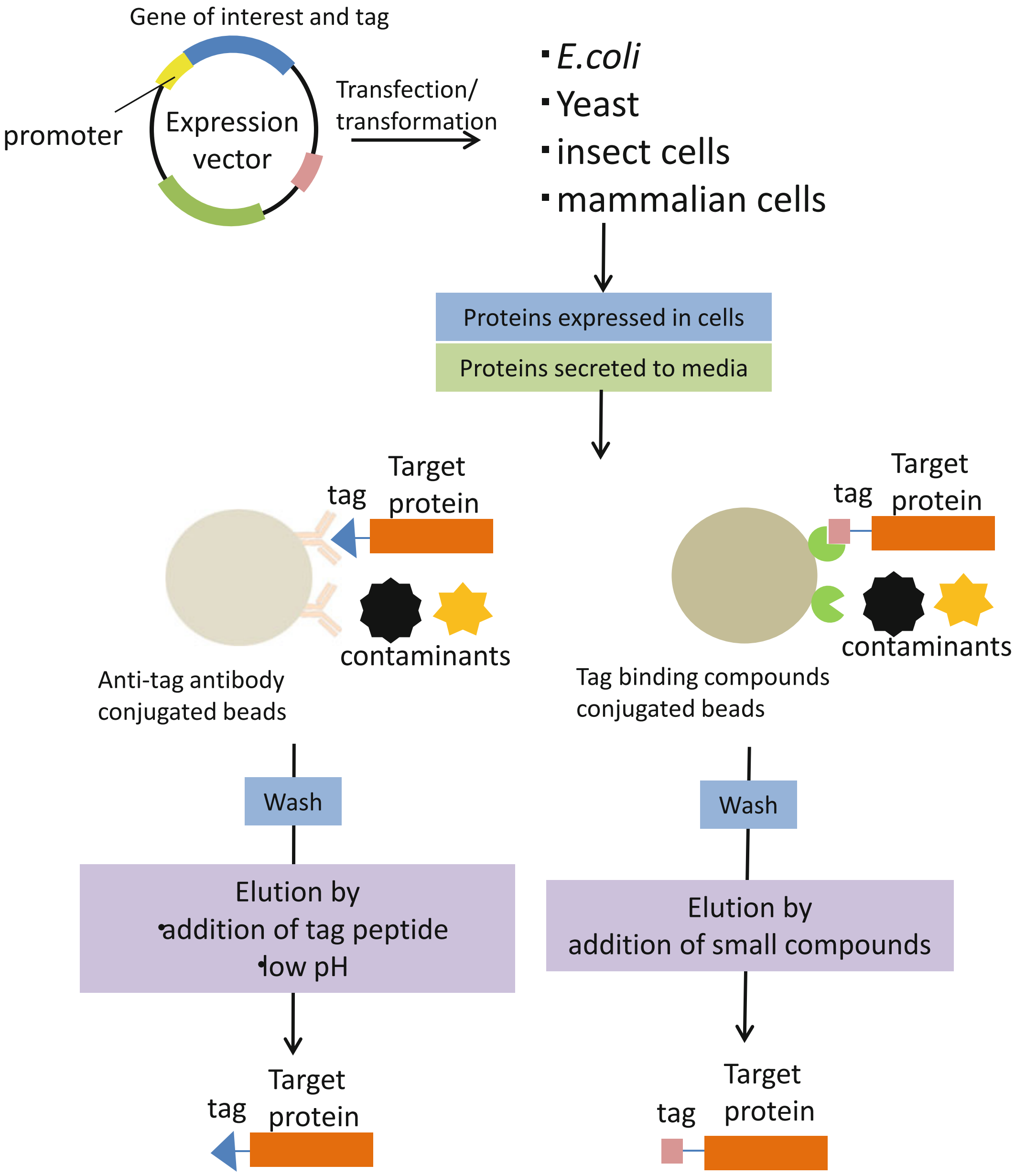


Fig. 1 Schematic image of the process from expression to purification of proteins

containing six consecutive histidines is the most effective tag to purify proteins [43]. While some proteins contain consecutive histidine residues and these proteins can bind with IMAC resins, they can generally be washed out by buffers with a low concentration of imidazole. After the washing procedure, the target protein can be purified in high purity by eluting with a buffer solution containing a high concentration of imidazole (typically 500 mM of imidazole) [43].

2.1 Materials and Methods

2.1.1 Buffers

Lysis buffer	50 mM Tris·HCl, pH 8.0, 150 mM NaCl (0.02 % Triton-×100, 5–10 % glycerol); not below pH 4.0
Wash buffer	50 mM Tris·HCl, pH 8.0, 100 mM NaCl, 10 mM imidazole (~500 mM NaCl)
Elution buffer	50 mM Tris pH 8.0, 100 mM NaCl, 250–500 mM imidazole

2.1.2 Methods

When the protein of interest is in an insoluble form in the lysis solution, a denaturant such as guanidinium chloride or urea can be utilized to solubilize the proteins from the pellet. Adding the IMAC resin to the lysis solution, the mixture is incubated with gentle shaking at 4 °C for 1 h. The supernatant and the resin precipitate are separated by centrifugation or filtration. The IMAC resin is washed a few times using the wash buffer containing the low concentration of imidazole (e.g., 10 mM), which is useful for the removal of nonspecifically bound proteins (but it is also important to note that in some cases, His-tagged proteins are removed even at a low concentration of imidazole). Adding the elution buffer containing 250 mM imidazole or more, IMAC resin is incubated with gentle shaking at 4 °C for 1 h. The target protein is eluted in the supernatant.

3 Glutathione S-Transferase Tag

Glutathione S-transferase (GST) is a major member of detoxification enzymes [4]. GSTs are composed of three superfamilies: cytosolic, mitochondrial, and microsomal GST proteins. These GSTs are also classified in terms of cytosolic and membrane-bound isoenzymes. There are five types of cytosolic enzymes, including alpha, mu, pi, sigma, and theta isoforms. The microsomal GST isoforms, delta, kappa, omega, and zeta isoenzymes, and the mitochondrial superfamily are membrane bound. In the cellular signaling pathways, these GST proteins inhibit some kinases such as those of the MAPK cascade, which regulates cell proliferation and death [2, 24]. Furthermore, GSTs can interact with glutathione (GSH) strongly and regulate the oxidation/reduction environment in the cell. GSH comprises three amino acids, Glu, Cys, and Gly. This peptide plays an important role to protect cells from reactive oxygen species such as peroxides and free radicals. The interaction between GST and GSH is strong with the dissociation constant (K_d) at a nanomolar level. Thus, using this high recognition ability, the GST tag can be applied to GSH-based affinity chromatography. The GST tag can be fused with the protein of interest either at the N- or C-terminal. The GST tag sometimes exhibits the increment of stability and solubility of target proteins. Fusion with GST can facilitate the proper folding of the target proteins because GST is rapidly folded right after translation. After the lysis of the host cells expressing the GST fusion proteins, the lysate is added to the GSH resin, which is generally composed of agarose or sepharose. GST-tagged proteins are immobilized on resins through the interaction of the GST tag with GSH. Notably, the GST tag can bind GSH at above pH 7, and thus, the buffer for the GST purification is normally prepared at pH 8 or above. After washing the resin to remove nonspecific proteins, the GST-tagged protein can be purely eluted from the GSH resin by adding an excess of GSH.

3.1 Materials and Methods

3.1.1 Buffers

Lysis buffer	50 mM Tris pH 7.5, 150 mM NaCl, 0.02 % Triton X-100
Wash buffer	50 mM Tris pH 8.0, 100 mM NaCl (~500 mM NaCl)
Elution buffer	50 mM Tris pH 8.0, 100 mM NaCl, 20 mM *reduced* glutathione (*make fresh every time*)

3.1.2 Methods

1. Lysis of *E. coli* cells by the ultrasonic sonicator or a French press with the lysis buffer. Centrifuge the lysate at 20,000 × *g* at 4 °C for 10 min.
2. Filtrate the supernatant using a 0.22-μM pore size filter unit.
3. Wash the GSH resin with the wash buffer. Add 10 resin bed volumes of the wash buffer and centrifuge the mixture until the resin collects at the bottom. Remove and discard the supernatant. Wash the resin at least two more times with the wash buffer.
4. Transfer the lysate to the tube that contains the GSH resin. Incubate at 4 °C for 30–60 min with agitation.
5. Centrifuge and wash the resin three times with the wash buffer.
6. Add the elution buffer to the GSH resin. Incubate at 4 °C for 30 min and collect the supernatant. If desired, the elution can be repeated multiple times.

4 Maltose-Binding Protein

Maltose-binding protein (MBP) is another popular protein tag, similar to the abovementioned GST tag and polyhistidine tag. MBP is composed of 388 amino acids (42 kDa) and works in the *E. coli* maltose/maltodextrin system, which regulates the uptake and catabolism of maltodextrin. This protein tag binds with the “maltose,” as its name suggests, and, furthermore, can bind a few similar sugar groups, such as trehalose and amylose [7, 33]. Almost all of the marketed resins are conjugated with amylose. Maltose is a reduced disaccharide that consists of two α-glucose monomers joined by the α-1,4 glycosidic bond, and amylose is its polymer. In other words, MBP recognizes the maltose part of amylose. Maltose has a higher affinity toward MBP than amylose; therefore, maltose can be used as an elution compound by the competition of the interaction between the amylose and MBP. Similar to the GST tag, the MBP tag can induce the increment of the solubility and stability in some cases. As a common vector system, the pMAL vector can be purchased from New England Biolabs. In this system, the MBP tag is located at the N-terminal side of the target protein via a specific proteinase cleavage site.

5 Avidin/Biotin System

Avidin is a glycoprotein, approximately 70 kDa, first found in the white of a chicken egg. The physiological functions of the avidin protein in the egg have not been clarified. Biotin is one of the B-group vitamins and is also known as vitamin H. In 1976, the avidin–biotin interaction was first reported as a powerful tool in biological science [14, 18]. The interaction is extremely strong among noncovalent bonds, $K_d = 10^{-15}$ M; this affinity exhibits a much higher value than antigen–antibody reactions [12]. Because avidin is a tetrameric protein and its monomer can bind one biotin, avidin can bind up to four biotins. Other biotin-binding proteins such as streptavidin and neutravidin also have the ability to interact with four biotin molecules. Streptavidin is derived from the bacteria *Streptomyces* and thus does not require the sugar modification. Neutravidin does not have any sugar chains, either. These proteins do not interact with sugar-binding proteins such as lectins. Combined with their neutral isoelectric point (pI), these proteins show less nonspecific interactions compared to the avidin protein and are suitable as an experimental tool. The interaction between avidin and biotin is very rapid and nearly irreversible, so it can be used for enzyme-linked immunosorbent assay (ELISA), immunocytochemistry, pull-down assays, and protein immobilization. Recently, this interaction was modified and used extensively as a purification system. It has become possible to use this interaction reversibly by using desthiobiotin [15] and Strep-tag® composed of octapeptides [40, 41]. Because these molecules or peptides have weaker affinities against the avidin, biotin can be used as an additive for elution from these avidin-related proteins.

6 FLAG Tag

The FLAG tag consists of eight hydrophilic amino acids (DYKDDDDK, from the N-terminus to the C-terminus) [9]. The developmental history of FLAG tag is rather unique. Some other tags (e.g., myc and HA discussed below) are part of native proteins and a monoclonal antibody was first isolated against the proteins, then the epitope was characterized. In contrast, the FLAG epitope was artificially designed first, and then monoclonal antibodies were prepared. So far, monoclonal anti-FLAG antibodies, such as M1, M2, and M5, have been developed and are commercially available (i.e., Sigma-Aldrich). It is known that the aromatic amino acid (tyrosine) of the FLAG tag is the major factor in tag–antibody interactions [19], but each commercially available antibody has a different epitope and affinity to the FLAG tag (Table 2). For instance, if the α-amino group of the first amino acid of FLAG tag is freely accessible, the M1 antibody binds with three to four orders of magnitude higher affinity [37]. Combining the use of the

Table 2
The affinities of monoclonal antibodies (M1, M2, and M5) with different fusion positions of the FLAG tag to the protein are shown

Flag-tag fusion proteins	Affinity with each monoclonal antibody		
	M1	M2	M5
Unprocessed N-terminus tagged proteins Signal peptide — FLAG — protein	ND	+	ND*
Met-N-terminus tagged proteins Methionine — FLAG — protein	–	+	++
N-terminus tagged proteins FLAG — protein	+	+	weak
Taginserted proteins protein — FLAG — protein	–	+	ND*
C-terminus proteins protein — FLAG	–	+	weak
Calcium dependent binding	+	–	–

ND indicates not detected. "++" "+" "weak" "–" indicate the binding affinity from strong to weak, and none

Table 3
Enzymes generally used for tag removal

Protease names and digestion site	Representative available company	Protease capture
Thrombin	GE, Merck Millipore,	Benzamidine–agarose
LVPR▼GS	SIGMA, Roche	
Factor Xa	GE, New England Biolabs,	Benzamidine–agarose
I(D/E)GR▼	Roche	
Enterokinase	New England Biolabs,	Trypsin inhibitor–agarose
DDDDK▼	Merck Millipore, Roche	
TEV protease	Promega, Nacalai, SIGMA	Ni-NTA (6 His recombinant TEV)
ENLYFQ▼G		
PreScission	GE	GSTrap for GST fusion enzyme
LDVLFQ▼GP		
HRV 3C Protease	Takara, Merck Millipore, Pierce	Ni-NTA (6 His recombinant enzyme)
LEVLFQ▼GP		
SUMO Protease	Life Technologies, LifeSensors	Ni-NTA (6 His recombinant enzyme)
Recognize the tertiary structure of SUMO		

FLAG tag with other tags leads to more efficient purification methods [25].

Elution of FLAG-tagged proteins from anti-FLAG antibody can be performed by two different methods. One is a low-pH elution method similar to other antibody-based purification systems. The other is elution by addition of 2–5 mM EDTA, which is a mild elution procedure for many proteins. Furthermore, another advantage is that the tag itself is cleaved by an enterokinase without any insertion of additional amino acids because the sequence of FLAG is recognized by the enzyme (Table 3). A weak point of the system is that the monoclonal antibody matrix for purification is not so stable as others, e.g., Ni^{2+}–NTA or streptavidin beads [43].

7 Myc Tag

The myc tag, EQKLISEEDL sequence, is a short tag derived from the c-myc gene. Myc (c-Myc) is one of the transcription factors, which regulates the cell cycle. The myc gene has been extensively studied as an oncogene because the mutations in myc are found in many cancer cells. A monoclonal antibody, 9E10, which was raised against the myc peptide in mice, is available from the noncommercial Developmental Studies Hybridoma Bank [10, 20]. The agarose gels or beads covalently linked with anti-myc tag antibody are also commercially available from suppliers. The expression of myc fusion proteins in several expression hosts, such as bacteria, yeasts, insect cells, and mammalian cells, has also been successful. Purified c-myc-tagged proteins have been crystallized [32]. The myc tag can be fused to either the C-terminus or the N-terminus of a target protein. It should be noted to avoid fusing the tag directly behind the signal peptide of a secretory protein because it interferes with correct intercellular trafficking.

8 Fc Tag

Fc fusion is also highly used for the expression and purification of proteins. The immunoglobulin Fc domain is a 25 kDa protein with a sugar modification for structural stability. Sugar modification does not generally occur in prokaryotes: therefore, the expression of Fc fusion proteins is normally performed in eukaryotic cells. However, recent biotechnological developments have allowed us to express Fc fusion proteins in *E. coli* by introducing *Campylobacter jejuni* glycosylation machinery into *E. coli* and subsequent enzymatic transglycosylation [27, 42]. Although the Fc regions are originally located at the C-terminus of immunoglobulin, the Fc tag can link to either the N-terminus or the C-terminus of target proteins. Fc-tagged proteins can be utilized for pharmacological purposes (Table 4) [6, 38]. The most important feature of the Fc tag fusion is its ability to increase the protein half-life in the plasma, extending the efficacy of drugs. This phenomenon is mainly thought to be because of the following reasons: (1) Fc-tagged proteins interact with the salvage neonatal Fc receptor (FcRn) [39], and (2) larger molecules have slower renal clearance [23]. The attached Fc domain also enables the fused protein to interact with Fc receptors (FcRs) expressed in immune cells, which is particularly important for their antibody-dependent cellular cytotoxicity (ADCC) in oncological therapies and in the application for vaccines [28, 34]. In addition, in regard to their biophysical features, the Fc domain folds independently and can improve the solubility and stability of the fused proteins both in vitro and in vivo. Furthermore, the Fc region provides easy and cost-effective purification by using protein G/A affinity chromatography [5]. Protein G and protein A are cell-surface proteins from *Streptococcus* and *Staphylococcus* species, respectively. They have different binding

Table 4
Fc fusion proteins used as drugs

Drug name	Description	Indication	Expression system	Approved year	Company
Belatacept	Modified CTLA-4 fused to the Fc of human IgG1	Organ rejection	Mammalian and COS cells	2011	Bristol-Myers Squibb
Aflibercept	Second Ig domain of VEGFR1 and third domain of VEGFR2 fused to the Fc of human IgG1	Age-related macular degeneration	CHO cells	2011	Regeneron Pharmaceuticals
Rilonacept	IL-1R fused to the Fc of human IgG1	Cryopyrin-associated periodic syndromes	CHO cells	2008	Regeneron Pharmaceuticals
Romiplostim	Thrombopoietin-binding peptides fused to the Fc of human IgG1	Thrombocytopenia in chronic immune thrombocytopenic purpura patients	*E. coli*	2008	Amgen/Pfizer
Abatacept	Mutated CTLA-4 fused to the Fc of human IgG1	Rheumatoid arthritis	Mammalian cells	2005	Bristol-Myers Squibb
Alefacept	LFA-3 fused to the Fc of human IgG1	Psoriasis and transplant rejection	CHO cells	2003	Astellas Pharma
Etanercept	Human p75 TNF receptor fused to the Fc of human IgG1	Rheumatoid arthritis	CHO cells	1998	Amgen/Pfizer

affinities depending on the kind of immunoglobulins (Table 5). Cleavage of the Fc domain from the Fc-tagged protein is performed by papain. This enzyme specifically cleaves the hinge region between the target protein and the Fc domain. Recently, for better stability of the cleaved protein, a 3C protease cleavage site was introduced into the hinge region because this enzyme has high specificity and a low optimal reaction temperature [3].

Table 5
The affinities of various kinds of immunoglobulins from several species with protein G or A

Immunoglobulins	Affinity for protein A	Affinity for protein G
Human IgG_1	++++	++++
Human IgG_2	++++	++++
Human IgG_3	–	++++
Human IgG_4	++++	++++
Human IgM	±	–
Human IgA_1	–	–
Human IgA_2	+	–
Human IgD	–	–
Human IgE	±	–
Mouse IgG1	+	++++
Mouse IgG_{2a}	++++	++++
Mouse IgG_{2b}	+++	+++
Mouse Ig_{G3}	++	+++
Mouse IgM	±	–
Mouse IgA	–	–
Mouse IgE	–	–
Rat IgG_1	±	+
Rat IgG_{2a}	–	++++
Rat IgG_{2b}	–	++
Rat IgG_{2c}	+	++
Bovine IgG_1	±	+++
Bovine IgG_2	+++	+++
Bovine IgA	–	–

The number of "+" reflects the affinity strength. "±" indicates slight affinity
"–" indicates no binding affinity

9 SUMO Tag

Small ubiquitin-like modifier (SUMO) tag is a recently developed tag that accelerates the solubility of the target protein. In *Saccharomyces cerevisiae*, the posttranslational modification of SUMO, Smt3, provides proteins with wide biological function, such as nuclear–cytosolic transport, transcriptional regulation, and apoptosis [13]. In contrast to ubiquitin, which is a "tag" for degradation, the SUMO tag often extends the lifetime of the proteins. When the

SUMO tag is used as an N-terminal fusion protein in prokaryotic expression, SUMO promotes folding and structural stability, which leads to enhanced functional production compared to untagged protein [30, 31]. Furthermore, the SUMO tag itself has a unique advantage that a SUMO-specific protease (*S. cerevisiae* UlpI) can digest a Gly–Gly motif of the tag. Thus, the SUMO tag is widely available in both prokaryotic and eukaryotic expression systems. Recently, an engineered SUMO-based tag, SUMOstar, has been established to enhance protein expression in eukaryotic cells because the SUMOstar sequence could not be recognized by the endogenous SUMO protease [22, 26, 35]. Instead of the conventional SUMO protease, this SUMOstar tag could be cleaved by engineered SUMOstar protease. Because of the recent usefulness of the SUMO tag for protein crystals, this tag will become more important and common in the future [1, 21, 29].

10 Other Tags

Hemagglutinin is well known as a surface protein in the human influenza virus and is involved in the adhesion to host cells. The HA tag consists of 9 amino acids, YPYDVPDYA, from the N-terminus to C-terminus, corresponding to the 98–106 amino acid residues in HA. HA monoclonal antibodies (and HA-antibody conjugated agarose) for purification are commercially available.

The GFP tag is widely used to investigate the subcellular localization of a target protein by fluorescence microscopy and the expression of exogenous proteins by FACS or Western blotting. The expression and purification of GFP-tagged recombinant proteins are not common owing to the problem of cost and the amount of protein expression. Recently, it has been reported that the GFP tag has been used for optimization of the expression and purification of a eukaryotic membrane protein [8].

The 1D4 epitope is nine amino acids (TETSQVAPA) derived from the intracellular C-terminus domain of bovine rhodopsin [17, 44]. Combining this epitope and the high-affinity 1D4 monoclonal antibody has established useful tools in antibody-based purification, localization studies, and Western blot analysis of 1D4-tagged proteins [11, 45]. Additionally, the 1D4 enrichment strategy offers a highly specific, non-denaturing method for purifying membrane proteins with yields and purities sufficient enough to use for structural characterization and functional proteomics applications [45].

11 Tag Digestion

Because tags described above have various characteristics, these might affect the physicochemical properties of the target protein fused with these tags. For example, the polyhistidine tag is composed of several consecutive histidine residues and thus confers a high positive charge to the fusion proteins. Conversely, huge tags such as GST and MBP might inhibit the enzyme activities of fusion proteins through their steric hindrance. In these cases, tags can be a useful tool for purification by avoiding undesired enzymatic function, but it is better to remove the tags for functional assays. Generally, there is a linker composed of some amino acids between a tag and a target protein, wherein the digestion sequences of some proteases are inserted into the linker site. The linker can be used for higher purification by removing tagged proteins from the resin by specific proteases. Typical proteases are 3C protease, thrombin, Factor Xa, and enterokinase (Table 3).

11.1 Removal of Tags

An example of the removal of a tag with an enterokinase site is shown below.

1. Dilute enterokinase in the storage buffer to prepare 0.01, 0.04, 0.1, 0.4, and 1 U/μl enterokinase solution.
2. Mix the following materials in a tube. The total volume should be adjusted to 50 μl by the addition of an appropriate amount of water.

10× reaction buffer	5 μl
Target protein	20 μg
Diluted enterokinase	5 μl
H_2O	X μl
Total volume	50 μl

3. Incubate the tube at room temperature (e.g., 25 °C).
4. Take a 10 μl aliquot after 2, 9, and 24 h of incubation. Each aliquot is mixed with 10 μl 2× SDS-PAGE sample/loading buffer for SDS-PAGE analysis.
5. Check the result of the tag cleavage by SDS-PAGE and decide suitable reaction conditions (time and the protease concentration).
6. Apply suitable reaction condition to a large amount of the tagged protein.
7. Purify the cleaved protein by chromatography, such as gel filtration chromatography.

References

1. Abbas YM, Pichlmair A, Górna MW, Superti-Furga G, Nagar B (2013) Structural basis for viral 5′-PPP-RNA recognition by human IFIT proteins. Nature 494(7435):60–64. doi:10.1038/nature11783
2. Adler V, Yin Z, Fuchs SY, Benezra M, Rosario L, Tew KD, Pincus MR, Sardana M, Henderson CJ, Wolf CR, Davis RJ, Ronai Z (1999) Regulation of JNK signaling by GSTp. EMBO J 18(5):1321–1334. doi:10.1093/emboj/18.5.1321
3. Asano R, Ikoma K, Kawaguchi H, Ishiyama Y, Nakanishi T, Umetsu M, Hayashi H, Katayose Y, Unno M, Kudo T, Kumagai I (2010) Application of the Fc fusion format to generate tag-free bi-specific diabodies. FEBS J 277(2):477–487. doi:10.1111/j.1742-4658.2009.07499.x
4. Boyer TD (1989) The glutathione S-transferases: an update. Hepatology 9(3):486–496
5. Carter PJ (2011) Introduction to current and future protein therapeutics: a protein engineering perspective. Exp Cell Res 317(9):1261–1269. doi:10.1016/j.yexcr.2011.02.013
6. Czajkowsky DM, Hu J, Shao Z, Pleass RJ (2012) Fc-fusion proteins: new developments and future perspectives. EMBO Mol Med 4(10):1015–1028. doi:10.1002/emmm.201201379
7. Diez J, Diederichs K, Greller G, Horlacher R, Boos W, Welte W (2001) The crystal structure of a liganded trehalose/maltose-binding protein from the hyperthermophilic Archaeon *Thermococcus litoralis* at 1.85 A. J Mol Biol 305(4):905–915. doi:10.1006/jmbi.2000.4203
8. Drew D, Newstead S, Sonoda Y, Kim H, von Heijne G, Iwata S (2008) GFP-based optimization scheme for the overexpression and purification of eukaryotic membrane proteins in *Saccharomyces cerevisiae*. Nat Protoc 3(5):784–798. doi:10.1038/nprot.2008.44
9. Einhauer A, Jungbauer A (2001) The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins. J Biochem Biophys Methods 49(1–3):455–465
10. Evan GI, Lewis GK, Ramsay G, Bishop JM (1985) Isolation of monoclonal antibodies specific for human c-myc proto-oncogene product. Mol Cell Biol 5(12):3610–3616
11. Farrens DL, Dunham TD, Fay JF, Dews IC, Caldwell J, Nauert B (2002) Design, expression, and characterization of a synthetic human cannabinoid receptor and cannabinoid receptor/ G-protein fusion protein. J Pept Res 60(6):336–347
12. Finn FM, Iwata N, Titus G, Hofmann K (1981) Hormonal properties of avidin-biotinylinsulin and avidin-biotinylcorticotropin complexes. Hoppe Seylers Z Physiol Chem 362(6):679–684
13. Hay RT (2005) SUMO: a history of modification. Mol Cell 18(1):1–12. doi:10.1016/j.molcel.2005.03.012
14. Heggeness MH, Ash JF (1977) Use of the avidin-biotin complex for the localization of actin and myosin with fluorescence microscopy. J Cell Biol 73(3):783–788
15. Hirsch JD, Eslamizar L, Filanoski BJ, Malekzadeh N, Haugland RP, Beechem JM (2002) Easily reversible desthiobiotin binding to streptavidin, avidin, and other biotin-binding proteins: uses for protein labeling, detection, and isolation. Anal Biochem 308(2):343–357
16. Hochuli E, Bannwarth W, Döbeli H, Gentz R, Stüber D (1988) Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate adsorbent. Nat Biotechnol 6(11):1321–1325
17. Hodges RS, Heaton RJ, Parker JM, Molday L, Molday RS (1988) Antigen-antibody interaction. Synthetic peptides define linear antigenic determinants recognized by monoclonal antibodies directed to the cytoplasmic carboxyl terminus of rhodopsin. J Biol Chem 263(24):11768–11775
18. Hofmann K, Kiso Y (1976) An approach to the targeted attachment of peptides and proteins to solid supports. Proc Natl Acad Sci U S A 73(10):3516–3518
19. Hopp TP, Prickett KS, Price VL, Libby RT, March CJ, Pat Cerretti D, Urdal DL, Conlon PJ (1988) A short polypeptide marker sequence useful for recombinant protein identification and purification. Nat Biotech 6(10):1204–1210
20. http://dshb.biology.uiowa.edu/c-myc
21. Hu Z, Yan C, Liu P, Huang Z, Ma R, Zhang C, Wang R, Zhang Y, Martinon F, Miao D, Deng H, Wang J, Chang J, Chai J (2013) Crystal structure of NLRC4 reveals its autoinhibition mechanism. Science 341(6142):172–175. doi:10.1126/science.1236381
22. Hughes SR, Sterner DE, Bischoff KM, Hector RE, Dowd PF, Qureshi N, Bang SS, Grynaviski N, Chakrabarty T, Johnson ET, Dien BS, Mertens JA, Caughey RJ, Liu S, Butt TR, LaBaer J, Cotta MA, Rich JO (2009) Engineered Saccharomyces cerevisiae strain for improved

xylose utilization with a three-plasmid SUMO yeast expression system. Plasmid 61(1):22–38. doi:10.1016/j.plasmid.2008.09.001

23. Kontermann RE (2011) Strategies for extended serum half-life of protein therapeutics. Curr Opin Biotechnol 22(6):868–876. doi:10.1016/j.copbio.2011.06.012
24. Laborde E (2010) Glutathione transferases as mediators of signaling pathways involved in cell proliferation and cell death. Cell Death Differ 17(9):1373–1380. doi:10.1038/cdd.2010.80
25. Li Y (2011) The tandem affinity purification technology: an overview. Biotechnol Lett 33 (8):1487–1499. doi:10.1007/s10529-011-0592-x
26. Liu L, Spurrier J, Butt TR, Strickler JE (2008) Enhanced protein expression in the baculovirus/insect cell system using engineered SUMO fusions. Protein Expr Purif 62 (1):21–28. doi:10.1016/j.pep.2008.07.010
27. Lizak C, Fan YY, Weber TC, Aebi M (2011) N-Linked glycosylation of antibody fragments in Escherichia coli. Bioconjug Chem 22 (3):488–496. doi:10.1021/bc100511k
28. Loureiro S, Ren J, Phapugrangkul P, Colaco CA, Bailey CR, Shelton H, Molesti E, Temperton NJ, Barclay WS, Jones IM (2011) Adjuvant-free immunization with hemagglutinin-Fc fusion proteins as an approach to influenza vaccines. J Virol 85 (6):3010–3014. doi:10.1128/JVI.01241-10
29. Luo D, Ding SC, Vela A, Kohlway A, Lindenbach BD, Pyle AM (2011) Structural insights into RNA recognition by RIG-I. Cell 147 (2):409–422. doi:10.1016/j.cell.2011.09.023
30. Malakhov MP, Mattern MR, Malakhova OA, Drinker M, Weeks SD, Butt TR (2004) SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins. J Struct Funct Genom 5(1–2):75–86. doi:10.1023/B:JSFG.0000029237.70316.52
31. Marblestone JG, Edavettal SC, Lim Y, Lim P, Zuo X, Butt TR (2006) Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. Protein Sci 15(1):182–189. doi:10.1110/ps.051812706
32. McKern NM, Lou M, Frenkel MJ, Verkuylen A, Bentley JD, Lovrecz GO, Ivancic N, Elleman TC, Garrett TP, Cosgrove LJ, Ward CW (1997) Crystallization of the first three domains of the human insulin-like growth factor-1 receptor. Protein Sci 6(12):2663–2666. doi:10.1002/pro.5560061223
33. Nikaido H (1994) Maltose transport system of Escherichia coli: an ABC-type transporter. FEBS Lett 346(1):55–58
34. Nimmerjahn F, Ravetch JV (2008) Fcgamma receptors as regulators of immune responses. Nat Rev Immunol 8(1):34–47. doi:10.1038/nri2206
35. Peroutka RJ, Elshourbagy N, Piech T, Butt TR (2008) Enhanced protein expression in mammalian cells using engineered SUMO fusions: secreted phospholipase A2. Protein Sci 17 (9):1586–1595. doi:10.1110/ps.035576.108
36. Porath J (1992) Immobilized metal ion affinity chromatography. Protein Expr Purif 3 (4):263–281
37. Prickett KS, Amberg DC, Hopp TP (1989) A calcium-dependent antibody for identification and purification of recombinant proteins. Bio Tech 7:580–589
38. Rath T, Baker K, Dumont JA, Peters RT, Jiang H, Qiao SW, Lencer WI, Pierce GF, Blumberg RS (2013) Fc-fusion proteins and FcRn: structural insights for longer-lasting and more effective therapeutics. Crit Rev Biotechnol 35 (2):235–254. doi:10.3109/07388551.2013.834293
39. Roopenian DC, Akilesh S (2007) FcRn: the neonatal Fc receptor comes of age. Nat Rev Immunol 7(9):715–725. doi:10.1038/nri2155
40. Schmidt TG, Skerra A (1993) The random peptide library-assisted engineering of a C-terminal affinity peptide, useful for the detection and purification of a functional Ig Fv fragment. Protein Eng 6(1):109–122
41. Schmidt TG, Skerra A (1994) One-step affinity purification of bacterially produced proteins by means of the “Strep tag” and immobilized recombinant core streptavidin. J Chromatogr A 676(2):337–345
42. Schwarz F, Huang W, Li C, Schulz BL, Lizak C, Palumbo A, Numao S, Neri D, Aebi M, Wang LX (2010) A combined method for producing homogeneous glycoproteins with eukaryotic N-glycosylation. Nat Chem Biol 6 (4):264–266. doi:10.1038/nchembio.314
43. Terpe K (2003) Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biotechnol 60(5):523–533. doi:10.1007/s00253-002-1158-6
44. Wong JP, Reboul E, Molday RS, Kast J (2009) A carboxy-terminal affinity tag for the purification and mass spectrometric characterization of integral membrane proteins. J Proteome Res 8 (5):2388–2396. doi:10.1021/pr801008c
45. Zhong M, Molday RS (2010) Binding of retinoids to ABCA4, the photoreceptor ABC transporter associated with Stargardt macular degeneration. Methods Mol Biol 652:163–176. doi:10.1007/978-1-60327-325-1_9

Chapter 5

Cell-Free Protein Production for Structural Biology

Takaho Terada, Seisuke Kusano, Takayoshi Matsuda, Mikako Shirouzu, and Shigeyuki Yokoyama

Abstract

Cell-free protein synthesis using *E. coli* cell extracts has successfully been applied to protein sample preparation for structure determination by X-ray crystallization and NMR spectroscopy. The standard reaction solution for *E. coli* cell-free protein synthesis by coupled transcription-translation contains the S30 extract of *E. coli* cells, T7 RNA polymerase, and the DNA template (either plasmid or PCR-amplified linear DNA). Milligram quantities of proteins can be synthesized by the dialysis mode of the cell-free reaction in several hours. The *E. coli* cell-free protein synthesis method is suitable for the production of mammalian proteins, heteromultimeric protein complexes, and integral membrane proteins and features numerous advantages over the recombinant protein expression methods with bacterial and eukaryotic host cells. We present examples of structure determinations of mammalian and bacterial heteromultimeric protein complexes prepared by the cell-free production method.

Keywords Cell-free protein synthesis, *Escherichia coli*, Protein complexes, Structure determination

1 Introduction

Proteins are synthesized in cells by translation of their messenger RNAs (mRNAs), which are transcribed from the encoding genes. Large-scale protein synthesis can be performed not only by the recombinant DNA methods using host cells, such as *Escherichia coli*, yeast, insect, and mammalian cells, but also by the cell-free or in vitro protein synthesis methods. Cell-free protein synthesis can be accomplished with cell extracts prepared from a variety of organisms, including *E. coli* [1–10], wheat germ [11–13], insects [14–16], and humans [17, 18]. The cell extracts contain the ribosomes, transfer RNAs (tRNAs), various translation factors, and downstream factors, such as molecular chaperones. As for mRNA, cell-free translation may be performed by either using separately prepared mRNA or coupling translation with transcription from the template DNA ("coupled transcription-translation") by T7 or SP6 RNA polymerase [1, 2]. The reaction solution for

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks, DOI 10.1007/978-4-431-56030-2_5, © Springer Japan 2016

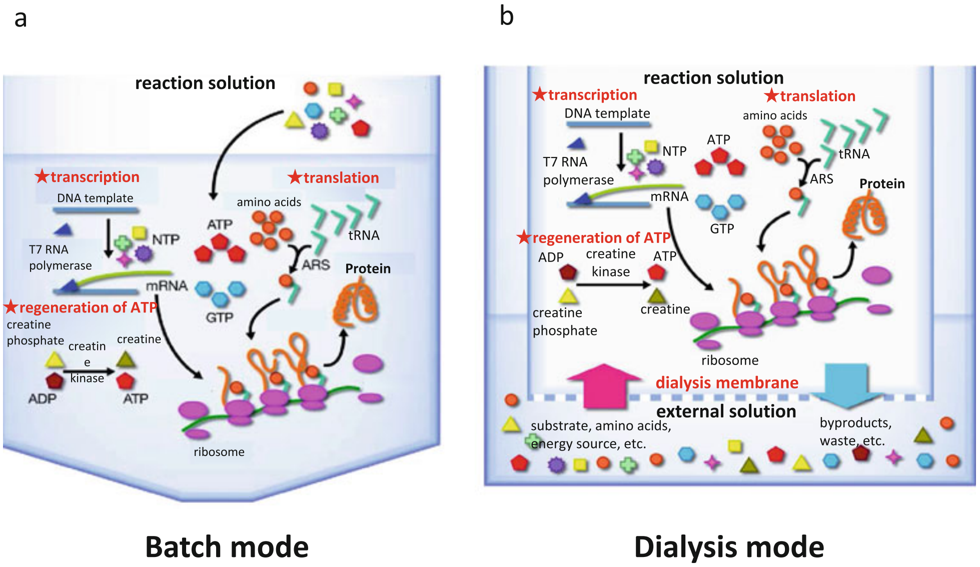


Fig. 1 Schematic illustration of the cell-free protein synthesis reaction modes. (**a**) The batch mode and (**b**) the dialysis mode

cell-free protein synthesis contains the cell extract, the template DNA for coupled transcription-translation or the pre-prepared mRNA, the low-molecular-mass substrates such as amino acids, the ATP regeneration system, and other components (Fig. 1). The cell-free synthesis reaction in a tube (the batch mode) continues for about one hour (Fig. 1a). To produce larger amounts of proteins, the reaction solution is dialyzed against the external solution containing the low-molecular-mass substrates (the dialysis mode) (Fig. 1b) [3, 4, 8–10, 19]. In this dialysis mode, the synthesis reaction continues for several hours, as the reaction solution is replenished with the low-molecular-mass substrates through the dialysis membrane, while the low-molecular-mass by-products are removed by dialysis (Fig. 1b) [3, 4, 8–10, 19].

The cell-free protein synthesis method has a number of advantages over conventional recombinant expression methods with host cells. For example, physiologically toxic proteins can be synthesized well by the cell-free method. The cell-free protein synthesis method actually has a much longer history than that of the host-vector recombinant protein expression, mainly for small-scale synthesis. However, drastic improvements of the cell-free protein synthesis method over the past decade have expanded its use for large-scale protein preparation [8–10, 20–24]. In fact, target proteins are frequently produced at levels of about 1 mg per ml cell-free reaction solution [25, 26]. In the case of the *E. coli* cell-free method, 1 ml of reaction mixture corresponds roughly to 50 ml of *E. coli* cell culture. This high yield of the cell-free protein synthesis method makes it cost-effective, and cell-free protein synthesis systems for large-scale protein production are now commercially available. The DNA template for mg quantity protein production by coupled transcription-translation with an *E. coli* cell extract can be either a pre-prepared plasmid or a PCR-amplified linear DNA template, encoding the protein [27]. This "cloning-free" nature enhances the efficiency of the cell-free method. For example, it only takes a few hours to perform the steps from PCR to cell-free protein synthesis [22–24, 27, 28]. Thus, the cell-free protein synthesis method has become one of the standard methods for protein sample preparation.

For structural biology, the cell-free synthesis method used to be regarded as the "salvage" method, which was only tried when other methods were unsuccessful. In contrast, the cell-free method is now considered as the "first-line" method for structural biology, which should be tried before other methods because of its various advantages over recombinant DNA methods using live host cells. First of all, large amounts of highly purified, homogeneous proteins are characteristically needed for structural biology analyses by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. In this regard, the cell-free protein synthesis method using the *E. coli* cell extract is much more suitable for mammalian protein production, in terms of both quality and quantity, than the host cell-based recombinant expression methods and cell-free synthesis methods using eukaryotic cell extracts. For example, *E. coli* cells may be engineered at the genome level, to tag a nuclease for removal from the cell extract [29]. Therefore, the *E. coli* cell-free method is by far the most frequently chosen for structural biology.

Naturally, the cell-free protein synthesis system is "open" with respect to the addition or subtraction of components. Many parameters, such as the reaction temperature, the incubation time, and the substrate and template concentrations, can be optimized easily. Intra- and intermolecular disulfide bonds may be formed by controlling the redox status of the reaction solution. Molecular chaperones may be added to the reaction solution, in order to facilitate proper folding. The cell-free protein synthesis method is suitable for the production of protein complexes consisting of two or more different components or subunits [30]. First, the components can be co-expressed simply by including their templates in stoichiometric amounts in the reaction solution, which is more strictly controllable than the cell-based recombinant methods. Moreover, a larger number of components may be co-expressed by the cell-free methods than by the recombinant cell-based methods. Otherwise, protein complexes may be reconstituted in a stepwise manner; for example, one or more components may be synthesized in the presence of a subcomplex consisting of the others [30]. Proteins can also be synthesized in complex with ligand(s), such as a low-molecular-mass cofactor, zinc ion [31], substrate, inhibitor, peptide fragment of the binding partner protein, and nucleic acids [30]. The formation of such complexes frequently improves the qualities of the products with respect to proper folding, as compared with the synthesis of the proteins by themselves. Furthermore, the cell-free synthesis of membrane proteins is particularly more advantageous than the recombinant cell-based methods, as described in Chap. 7.

For structural biology, the flexibility of the cell-free method in terms of nonstandard amino acids is very useful. For the multiwavelength anomalous diffraction (MAD) method in protein crystallography, the methionine residues in the protein may be almost completely replaced with selenomethionine, simply by using the same cell extract and selenomethionine in place of methionine in the reaction and external solutions [20, 23, 24, 32], while the recombinant expression method uses a methionine auxotrophic mutant strain of *E. coli*. Stable isotope (SI) labeling of proteins with nitrogen-15 (^{15}N), carbon-13 (^{13}C), and/or deuterium (^{2}H) for NMR measurements can easily be performed by cell-free protein synthesis [5, 7–9, 21–24, 28, 31]. Uniform SI labeling of proteins may be accomplished by using a mixture of uniformly labeled amino acids [8]. In addition, a variety of selective labeling techniques have been developed, using the advantages of the cell-free synthesis method [5, 7, 8, 33, 34]. For instance, unnatural amino acids may be incorporated site specifically into proteins by the cell-free method, using an engineered pair of a tRNA, specific to a special codon such as the UAG "stop" codon, and an aminoacyl-tRNA synthetase, specific to the unnatural amino acid [35–38]. The engineered pair of tRNA and pyrrolysyl-tRNA synthetase from *Methanosarcina mazei* was used, along with an extract of the *E. coli* RFzero strain [39], which lacks the gene-encoding release factor 1 recognizing the UAG and UAA stop codons, to introduce an epigenetic modification, acetyl-lysine, at four sites in the human histone H4 N-terminal tail [38].

Table 1 summarizes the structures of proteins produced by our group, using the cell-free method with the *E. coli* cell extract, deposited in the Protein Data Bank (PDB) as of June 22, 2015. The organisms range from human and mouse to viruses and bacteria. The number of NMR structures is much larger than that of the X-ray crystallographic structures, because most of the NMR structures were determined for human and mouse functional domains in the framework of the Japanese structural genomics project, "The Protein 3000 Project," from 2002 to 2007 [40–42]. We have deposited about 100 crystallographic structures of human and mouse proteins in the PDB, and eight of them are heteromultimeric protein complexes. For the human and mouse proteins with crystal structures determined with cell-free-produced samples, the average molecular masses are about 40 kDa. Therefore, the cell-free synthesis method is applicable for much larger proteins than the

Table 1
The numbers of PDB-deposited structures of proteins produced by *E. coli* cell-free protein synthesis method in our group (Deposited from Apr. 2001 to Dec. 2014)

Source organism	X-ray	NMR
Vertebrate		
Homo sapiens	74	1029
Mus musculus	30	261
Rattus rattus		1
Invertebrate	1	4
Yeast		2
Plant	1	33
Bacteria	17	3
Virus	5	
Total	128	1333

Table 2
The numbers of PDB-deposited structures of proteins produced by the wheat germ cell-free synthesis method (Available from http:/www.uwstructuralgenomics.org/structures.htm, accessed on Jun. 22, 2015)

Source organism	X-ray	NMR
Vertebrate		
Homo sapiens	1	5
Mus musculus		1
Danio rerio		2
Plant	3	9
Bacteria	1	
Total	5	17

functional domains analyzed by NMR spectroscopy (12 kDa). Table 2 summarizes the structures of proteins produced by the cell-free method using wheat germ extract, from The Center for Eukaryotic Structural Genomics (USA), deposited in the PDB as of January 13, 2015. We expect that the *E. coli* cell-free protein synthesis method will be used more extensively in the future, particularly for difficult proteins, such as mammalian proteins, protein complexes, and membrane proteins.

2 Cell-Free Protein Production Methods for Structural Biology

2.1 Workflow of Cell-Free Protein Production

The overall workflow of cell-free protein production is shown in Fig. 2. The preliminary experiment is performed through small-scale reactions to optimize various conditions. Using these optimized conditions, the reaction scale can be increased to the large-scale protein production. Selenomethionine and stable isotope-labeled amino acids may be used to label the product for X-ray crystallography and NMR spectroscopy, respectively.

2.2 Template DNA for Cell-Free Coupled Transcription-Translation

The template DNA for cell-free coupled transcription-translation in the *E. coli* extract contains the coding region of the target protein and the flanking sequences for transcription, translation, and purification. A typical template DNA is shown in Fig. 3. The flanking sequences may be provided by the plasmid vector or PCR primer (s). The two-step PCR method [27] is useful to efficiently construct the designed template DNA and particularly for the preparation of a large number of constructs for comparison. The tag is selected by considering not only the ease of purification but also the folding and/or solubility, from a variety of tags, such as 6×histidine (6His),

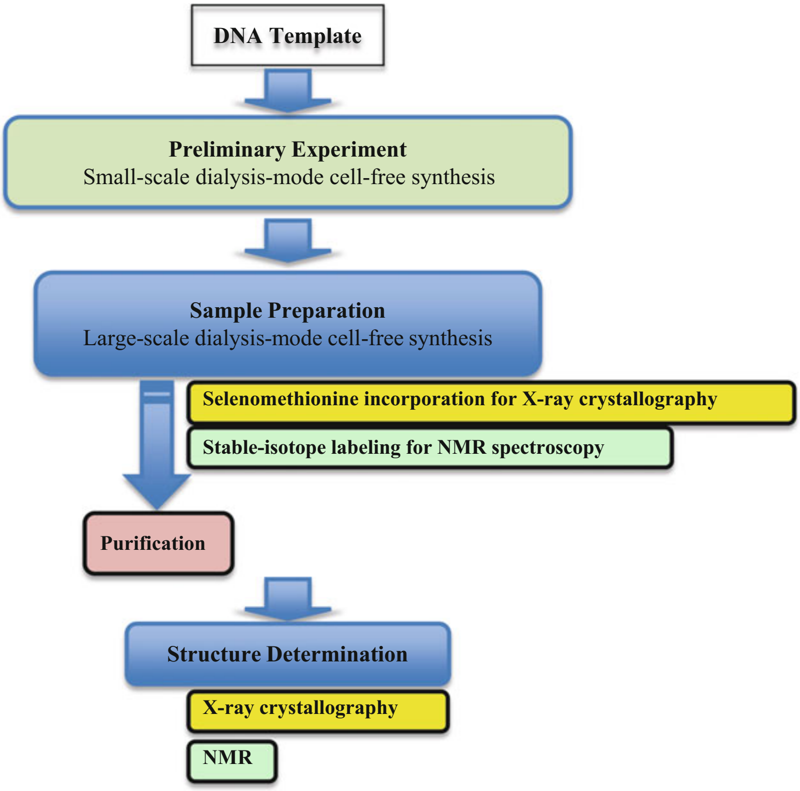

DNA Template
Preliminary Experiment
Small-scale dialysis-mode cell-free synthesis
Sample Preparation
Large-scale dialysis-mode cell-free synthesis
Selenomethionine incorporation for X-ray crystallography
Stable-isotope labeling for NMR spectroscopy
Purification
Structure Determination
X-ray crystallography
NMR

Fig. 2 Workflow of the cell-free protein production

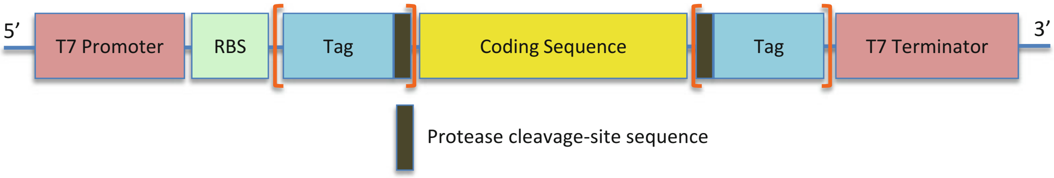


Fig. 3 Typical design of the template DNA for coupled transcription-translation in *E. coli* cell-free protein synthesis. The sequence encoding the target protein (coding sequence) and the preceding RBS (ribosome-binding site) for translation are flanked on the 5′ and 3′ sides by the T7 Promoter and T7 Terminator sequences, respectively. The N- and/or C-terminal tag sequence (usually including a protease cleavage site sequence) may be introduced not only for detection and purification but also for increasing folding and/or solubility

streptavidin-binding peptide (SBP), glutathione-S-transferase (GST), maltose-binding protein (MBP), and small ubiquitin-related modifier (SUMO). The template DNA, in the form of either plasmid DNA or PCR-amplified linear DNA, may be used in the cell-free reaction. The use of a linear template DNA significantly shortens the total duration of the experiment, and a higher protein yield is generally obtained with the use of plasmid DNA. In the latter case, the plasmid DNA must be well purified with a commercially available kit (Qiagen, Promega, etc.).

2.3 E. coli *Cell-Free Protein Synthesis System*

The S30 fraction (the supernatant fraction obtained after cell disruption and centrifugation at 30,000 × *g*) of *E. coli* cells is used as the cell extract for cell-free protein synthesis. We usually use the S30 extract of *E. coli* strain BL21 CodonPlus-RIL (Agilent Technologies), containing extra copies of the genes encoding minor tRNAs [10, 43]. Various kits for cell-free protein synthesis with *E. coli* cell extracts are commercially available, and those suitable for structural biology sample preparation should be chosen. For structural biology purposes, the cell-free protein expression kit "Musaibo-Kun" (Taiyo Nippon Sanso, Japan), "*i*PE Kit" (Sigma-Aldrich, USA), and the Remarkable Yield Translation System (RYTS) Kit (Protein Express, Japan) are useful and based on the method by Kigawa et al. [9]. The RTS 100 *E. coli* HY Kit (Biotechrabbit GmbH, Germany), the EasyXpress Protein Synthesis Kit (QIAGEN, The Netherlands), and the S30 T7 High-Yield System (Promega, USA) are also suitable. Some products are optimized for special purposes, such as the use of a linear template DNA and disulfide bond formation. To facilitate proper folding, we prepare the S30 extract from *E. coli* BL21 cells expressing a set of *E. coli* chaperones (DnaK/DnaJ/GrpE and/or GroEL/GroES) in addition to the minor tRNAs for rare codons, such as AGA/AGG, AUA, and CUA. Notably, non-natural amino acids can be incorporated into proteins in response to UAG codons much more efficiently by using the S30 extract of the *E. coli* RFzero strain, which lacks the release factor 1 gene [35, 36]. The detailed protocols for *E. coli* cell extract preparation have been published [9, 10, 29].

2.4 Cell-Free Protein Synthesis Reaction Solution

The reaction solution for *E. coli* cell-free coupled transcription-translation contains the *E. coli* S30 extract, the DNA template, the T7 RNA polymerase, and the substrates for transcription and translation. The components of the standard reaction solution are listed in Table 3. The order of the components in Table 3 roughly corresponds to that used to set up the reaction solution. For transcription, T7 RNA polymerase, prepared as reported in [44], is used. For translation, the S30 extract is prepared in 10 mM Tris-acetate buffer (pH 8.2), containing 60 mM potassium acetate, 16 mM magnesium acetate, and 1 mM DTT, and used at a final concentration of 30 % (v/v) in the reaction solution. The S30 extract contains the endogenous tRNAs from *E. coli* cells, but is supplemented with *E. coli* MRE600-derived tRNA (Roche Applied Science, 109550). The low-molecular-mass component mixture solution, *l*ow-*m*olecular-weight *c*reatine *p*hosphate t*y*rosine (LMCPY), contains 160 mM HEPES-KOH buffer (pH 7.5), 4.13 mM L-tyrosine, 534 mM potassium L-glutamate, 5 mM DTT, 3.47 mM ATP, 2.40 mM GTP, 2.40 mM CTP, 2.40 mM UTP, 0.217 mM folic acid, 1.78 mM cAMP, 74 mM ammonium acetate, and 214 mM creatine phosphate. The other amino acids besides L-tyrosine, which is included in LMCPY, are provided as "A.A.(-Y)", containing 10 mM DTT and 20 mM each of the 19 amino acids, as shown in Table 3.

As DTT is used in the standard reaction solution, protein synthesis is fundamentally performed under reducing conditions, whereas the use of DTT-free LMCPY is recommended for the production of disulfide bond-forming proteins, such as secreted

Table 3
Standard composition of the *E. coli* cell-free protein synthesis reaction solution

Reagent	Stock conc	Final conc
LMCP(Y)		37.33 % (v/v)
NaN_3	5 % (w/v)	0.05 % (w/v)
$Mg(OAc)_2$	1.6 M	9.28 mM
A.A.(-Y)	20 mM	1.5 mM each
tRNA	17.5 mg/ml	0.175 mg/ml
Creatine kinase	3.75 mg/ml	0.25 mg/ml
S30 extract		30 % (v/v)
T7 RNA polymerase	10 mg/ml	66.7 μ g/ml
Other factors		
Milli-Q water		To the desired volume

proteins and membrane proteins, as described in the next section (Sect. 2.5). The optimal magnesium concentration depends to some extent on the target proteins, and it should therefore be optimized for each target protein, in the range of 5–20 mM. For ATP regeneration, creatine kinase and its substrate, creatine phosphate, are used. The optimal DNA template concentration for coupled transcription-translation should be determined by a preliminary small-scale cell-free experiment.

2.5 Protein Folding

Metal Ligation, Ligand Binding, and Complex Formation: For Zn-binding proteins, an appropriate concentration (usually around 50 μM) of $ZnCl_2$ or $ZnSO_4$ should be added [30, 31]. Ligand-binding proteins are synthesized in the presence of the ligand (cofactor, substrate, inhibitor, etc.) in the reaction solution, since the ligand is expected to help the protein fold properly. For protein complex formation, two or more DNA templates are simultaneously used. The ratio of these templates should be adjusted prior to the large-scale cell-free production [30].

Molecular Chaperones: To facilitate correct folding, appropriate molecular chaperones [45] are prepared separately, and their mixture is added to the cell-free reaction. Otherwise, the S30 extract for the correct folding of the target protein(s) and protein complex(es) should be added. Among the *E. coli* chaperones [45], DnaK/DnaJ/GrpE and GroEL/GroES may function in the early and late stages, respectively, of chaperone-assisted protein folding. Therefore, single and/or dual uses of the two sets of chaperones in the cell-free protein synthesis are usually tested for precipitating or aggregating proteins.

Disulfide Bonds: For disulfide bond-containing proteins, cell-free synthesis is performed under more oxidative redox conditions than the standard conditions. The ratio between reduced glutathione (GSH) and oxidized glutathione (GSSG) may be optimized, by testing ratios between 1:9 and 9:1. A disulfide isomerase [46], such as *E. coli* DsbC, is usually added to facilitate proper protein folding. *E. coli* Skp [47, 48] may be used as a chaperone in addition to DsbC.

Reaction Temperature: For proper protein folding, the incubation temperature may be selected according to the efficiency of folding in the range of 15–37 °C, while the standard temperature is about 25 °C.

2.6 Amino Acid Labeling for Structure Determination

Selenomethionine Incorporation for X-ray Crystallography: Selenomethionine-substituted proteins for MAD phasing can be obtained by cell-free protein synthesis, in which the L-methionine in the reaction and external solutions is simply replaced by L-selenomethionine. The amino acid mixture lacking L-methionine and the 20 mM selenomethionine solution with 10 mM DTT are prepared separately and used in place of the standard amino acid mixture. Selenocysteines may be used instead of selenomethionine in the reaction solution and incorporated in place of cysteine in the protein for the MAD method. Iodine-/bromine-substituted amino acids such as tyrosine can be incorporated into specified site(s) of the protein, by using the "expanded genetic code system," and may also be used for MAD phasing [49].

Stable Isotope Labeling for NMR Spectroscopy: The production of stable isotope (SI)-labeled protein samples for multinuclear NMR spectroscopy is performed by replacing the amino acid(s) to be labeled in the cell-free reaction solution with SI-labeled ones. The mixture solution containing 10 mM DTT and 20 mM each of the SI-labeled amino acids should be used. Uniform SI labeling of proteins is accomplished with mixtures of the 20 amino acids uniformly labeled with ^{15}N, ^{13}C, and/or ^{2}H. Amino acid-selective SI labeling with respect to one or several kinds of amino acids can be performed more easily by the cell-free method than by the conventional recombinant method, because the SI scrambling between amino acids is minimized in the cell-free reaction. The cell-free protein synthesis method is quite useful for the stereo-array isotope labeling (SAIL) method [50]. We developed a cell-free system that utilizes potassium D-glutamate in place of L-glutamate, for efficient SI labeling [21, 34].

2.7 Reaction Modes of Cell-Free Protein Synthesis

The Batch and Dialysis Modes: The cell-free coupled transcription-translation may be performed in either the batch or dialysis mode (Fig. 1). The batch mode of cell-free protein synthesis is the simplest: the reaction is performed by incubating the reaction solution in a container, such as a test tube. In the reaction solution, the low-molecular-mass substrates for coupled transcription-translation and ATP regeneration become exhausted and by-products accumulate. Thus, the batch reaction reaches a plateau in a few hours. In order to achieve higher yields, the dialysis-mode cell-free synthesis reaction is performed by placing the reaction solution in a compartment with a dialysis membrane, such as a dialysis bag, and incubating it with the external solution, containing the same low-molecular-mass components as those in the reaction solution. In this mode, the substrates and the by-products are continuously provided and removed, respectively, by the external solution through the dialysis membrane. In the standard conditions, the molecular weight cutoff of the dialysis membrane is 10–15 kDa, and the ratio of the volume of the external solution to that of the reaction solution is equal to or greater than 10. Therefore, the protein synthesis reaction continues much longer in the dialysis mode than in the batch mode. The synthesis yield at 25–30 °C may reach 1–5 mg/ml reaction in 3–4 h. In practice, we usually stop the reaction at 3–4 h to avoid denaturation of the products, although it may continue longer. For synthesis at 15 °C, the reaction may be continued up to overnight.

For large-scale structural biology sample preparation, the cell-free synthesis reaction is performed in the dialysis mode, usually with a 1–10 ml reaction solution, while difficult targets such as large complexes and membrane proteins may be synthesized with a 30 ml or larger reaction solution. We recommend optimizing the construct and the conditions of the cell-free synthesis reaction, by performing small-scale reactions (5–30 μl) prior to the large-scale synthesis. For example, multiple PCR-amplified linear template DNAs encoding protein constructs with different terminal deletions may be generated and tested, with no cloning steps, by small-scale cell-free synthesis in multi-well plates, in either the batch or dialysis mode. Typically, multiple dialysis-mode cell-free reactions are performed in 96-well plates equipped with a dialysis membrane, and the volumes of the reaction solutions are 5 μl per well. The optimal construct/conditions are selected with respect to the yield, the solubility, etc., toward the larger-scale cell-free production, as described above, for structure determination.

2.8 Purification of Synthesized Proteins

After the large-scale protein synthesis reaction, the product is purified by affinity chromatography, incubated with the specific protease to cleave the affinity tag, and then purified again with the affinity column to remove the affinity-tag peptides. When the protease is fused with the same affinity tag without the cleavage site, the affinity-tagged protease can be removed together with the affinity-tag peptides in one step. The resultant fraction is subjected to further ion-exchange chromatography and gel-filtration chromatography for X-ray analysis.

3 Examples of Heteromultimeric Complexes Produced by the Cell-Free Method for Structure Determination

The cell-free protein synthesis method is highly advantageous for the production of heteromultimeric complexes consisting of two or more different component proteins [30]. Here, we describe several examples of cell-free heteromultimeric proteins produced for structural biology.

3.1 DOCK2•ELMO1

DOCK2 (dedicator of cytokinesis 2), which is specifically expressed in hematopoietic cells, activates the small GTP-binding protein Rac and thereby plays a critical role in cellular signaling events. The formation of a complex between DOCK2 and ELMO1 (engulfment and cell motility 1) is required for DOCK2-mediated Rac signaling. In 2012, we identified the regions of DOCK2 and ELMO1 required for their association and determined the complex structure by the following experimental strategies [51].

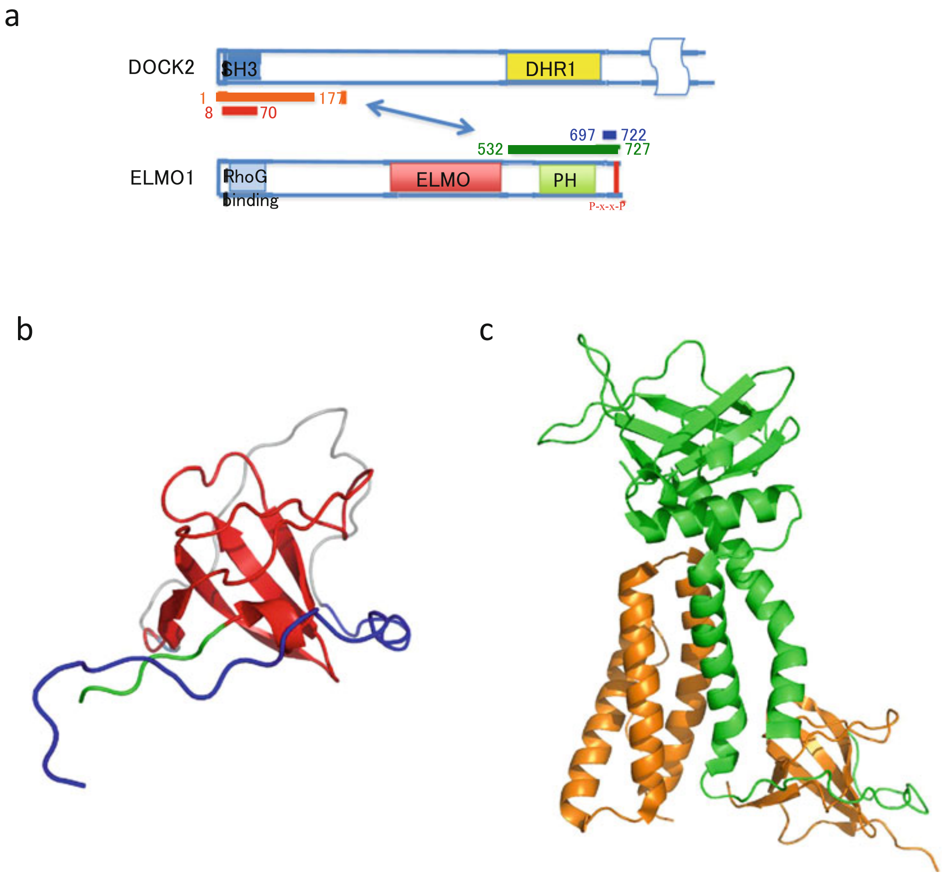


Fig. 4 Structures of the interactive regions of DOCK2 and ELMO1. (**a**) The domain organizations of DOCK2 and ELMO1. The *red* and *blue bars* indicate the DOCK2 and ELMO1 regions included in the fusion construct for NMR. The *orange* and *green bars* indicate the regions co-expressed for crystallization. (**b**) The NMR structure of the DOCK2 SH3-ELMOI peptide fusion protein (PDB ID: 2RQR) (*ribbon* representation). The DOCK2 SH3 domain and the ELMO1 peptide are colored *red* and *blue*, respectively. (**c**) The crystal structure of the DOCK2 (1–177)•ELMOI(532–727) complex (PDB ID: 3A98) (*ribbon* representation). The DOCK2(1–177) and ELMOI (532–727) proteins are colored *orange* and *green*, respectively

First, the N-terminal SH3 domain of human DOCK2 was found to bind to the C-terminal Pro-rich sequence of human ELMO1. Therefore, 87 differently designed DNA fragments encoding the human DOCK2 SH3 domain, with a human ELMO1 Pro-rich sequence peptide fused to its N- or C-terminus (Fig. 4a), were generated by PCR. The fragments were cloned into the pCR2.1 vector (Invitrogen) as fusions with an N-terminal histidine tag (a modified HAT tag) and a tobacco etch virus (TEV) protease cleavage site. Among these constructs, one fusion construct including an ELMO1 peptide (residues 697–722) fused to the N-terminus of the DOCK2 SH3 domain (residues 8–70) (designated as the DOCK2 SH3-ELMO1 peptide fusion protein) was selected as a suitable construct for NMR analysis, after checking the productivity and solubility of the constructs by the small-scale dialysis-mode of cell-free protein synthesis.

For NMR structure determination, the $^{13}C/^{15}N$-labeled DOCK2-ELMO1 peptide fusion protein was prepared by the large-scale dialysis-mode cell-free method. The solution structure determined by NMR (Fig. 4b, PDB ID: 2RQR) confirmed that the C-terminal Pro-rich region, especially P_{714}-x-x-P_{717}, of ELMO1 interacts with the SH3 domain of DOCK2, and prompted us to investigate the more detailed interactions between DOCK2 and ELMO1 by X-ray crystallography.

To identify the precise interacting regions of DOCK2 and ELMO1, a variety of N-terminal fragments of DOCK2 (residues 1–160, 1–177, 1–190, 9–160, 9–177, 9–190, 21–160, 21–177, and 21–190) and C-terminal fragments of ELMO1 (residues 532–717, 541–717, 550–717, 532–727, 541–727, and 550–727), with the N-terminal histidine tag (a modified HAT tag) sequence and the TEV cleavage site sequence, were generated by the two-step PCR method [27]. Using these PCR products as the templates for the small-scale dialysis-mode cell-free synthesis reactions, co- and separate protein expression studies were conducted. Among the above fragments, the DOCK2(1–177) fragment, consisting of the SH3 domain and the flanking region, and the ELMO1(532–727) fragment, consisting of the PH domain and the Pro-rich sequence, were selected as suitable fragments for crystallographic analyses and separately cloned into the pCR2.1 vector. By co-expression of the DOCK2(1–177) and ELMO1(532–727) fragments by the large-scale dialysis-mode cell-free synthesis method, the DOCK2(1–177)•ELMO1(532–727) complex protein was obtained in a soluble, selenomethionine-labeled form, whereas the DOCK2(1–177) fragment alone precipitated during the cell-free synthesis. The DOCK2(1–177)•ELMO1(532–727) complex protein, purified by histidine-tag affinity chromatography, histidine-tag cleavage with TEV protease, ion-exchange chromatography, and gel-filtration chromatography, was crystallized and the structure was determined at 2.1-Å resolution, as shown in Fig. 4c (PDB ID: 3A98, Structure Weight: 89,622.52). The complex structure revealed the structural basis for the mutual relief of DOCK2 and ELMO1 from their autoinhibited forms.

3.2 Rab27B•Slac2-a

Rab27A is required for actin-based melanosome transport in mammalian skin melanocytes. Rab27A (221 residues) and its isoform Rab27B (218 residues) bind to several effectors in common, including their specific effector, Slac2-a/melanophilin (590 residues).

We chose a C-terminally truncated form of the GTPase-deficient mutant Rab27B(Q78L) (residues 1–201; designated simply as Rab27B(1–201) hereafter) and the minimum effector region of Slac2-a that specifically binds to the GTP-bound form of Rab27

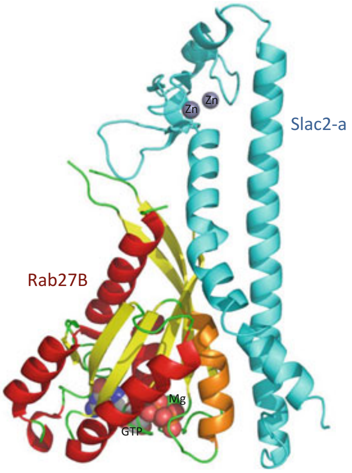


Fig. 5 Crystal structure of the Rab27B•Slac2-a complex. *Ribbon* representation of the Rab27B•Slac2-a complex structure (PDB ID: 2ZET). Rab27B is colored *red*, *orange*, and *yellow*. Slac2-a is colored cyan. Zn^{2+}, GTP, and Mg^{2+} are represented by spheres

(residues 1–146; designated as Slac2-a(1–146)) and produced their complex, Rab27B•Slac2-a, by the *E. coli* cell-free production method. First, two PCR-amplified DNA fragments encoding the proteins were independently cloned into the pCR2.1 vector (Invitrogen), as fusions with an N-terminal histidine tag (a modified HAT tag) and TEV protease cleavage site. The selenomethionine-labeled Rab27B•Slac2-a complex was obtained in a soluble form by the cell-free co-expression synthesis method, with 50 μM $ZnCl_2$ present in the reaction solution. The Rab27B•Slac2-a complex was stable and monomeric with 1:1 stoichiometry, as determined by gel filtration. The purified Rab27B•Slac2-a complex was crystallized and the structure was determined at 3.0-Å resolution, as shown in Fig. 5 (PDB ID: 2ZET, Structure Weight: 83,464.15). The crystal structure revealed the residues involved in the specific Rab27B•Slac2-a interaction [52].

3.3 V-ATPase

By using cell-free synthesized protein complex samples, high-quality structures of the *Enterococcus hirae* V1-ATPase A_3B_3 [53], DF [54, 55], and A_3B_3DF [53–55] complexes were determined by X-ray crystallography.

The PCR-amplified DNA fragments encoding the *E. hirae* V1-ATPase subunits A, B, D, and F (Eh-A, -B, -D, -F) were independently subcloned into the pCR2.1 vector (Invitrogen). These subunit proteins could only be expressed in the soluble forms by co-expression, and they formed the stoichiometric complexes by the cell-free protein synthesis method. To form the stable subcomplex and whole complex, the optimum concentrations of the plasmid

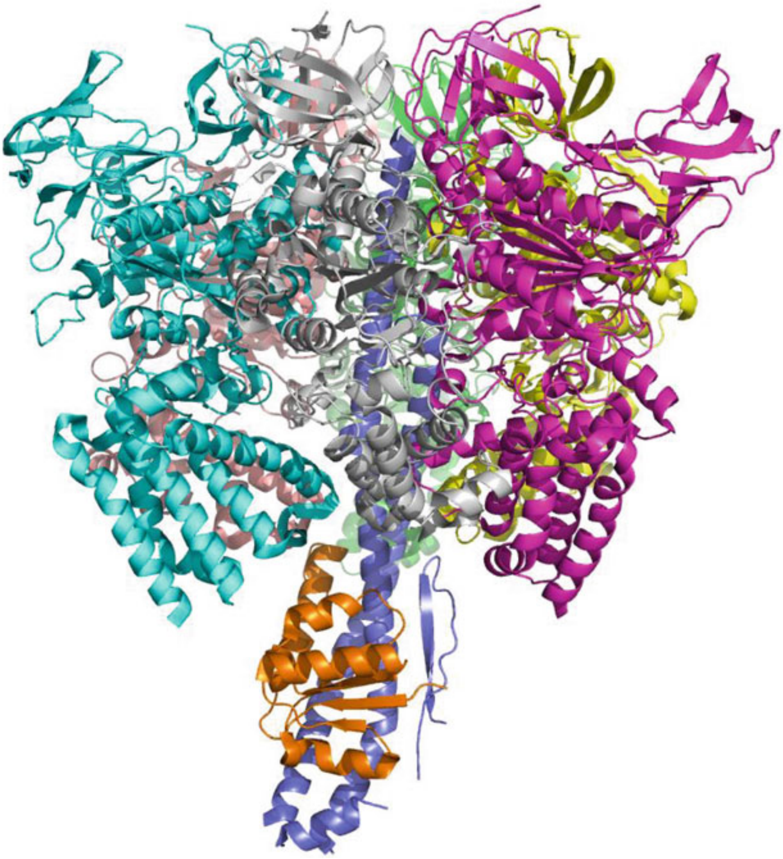


Fig. 6 Crystal structure of the *E. hirae* V-ATPase A_3B_3DF complex. *Ribbon* representation of the *E. hirae* V-ATPase A_3B_3DF complex (PDB ID: 3VR4)

DNA templates were determined by small-scale cell-free expression.

To promote X-ray crystallographic analyses, the selenomethionine-substituted Eh-A_3B_3 [53] and Eh-DF [54, 55] proteins were synthesized by the large-scale dialysis-mode *E. coli* cell-free method. The Eh-A_3B_3DF complex (Fig. 5.6, PDB ID: 3VR4, Structure Weight: 399,405.1) [53] was reconstituted from the Eh-A_3B_3 [53] and Eh-DF [54, 55] subcomplexes.

Using 27 mL of the cell-free reaction solution, more than 15 mg of the purified complex proteins were produced [54, 55].

3.4 Complexes of Disulfide-Bonded Proteins

Disulfide bond formation is required for the correct folding and structural stabilization of secreted and membrane proteins [56]. Cells have protein-folding catalysts to ensure that the correct pairs of cysteine residues interact during the folding process [57]. These enzymatic systems are located in the endoplasmic reticulum (ER) of eukaryotes and the periplasm of Gram-negative bacteria [58]. In bacteria, electron transfer occurs through cascades of disulfide bond formation/reduction between a series of proteins (DsbA, DsbB, DsbC, and DsbD) [46]. However, the overproduction of disulfide-bonded proteins by *E. coli* cells tends to result in precipitation, aggregation, or inclusion body formation, thus requiring protein solubilization and refolding.

We applied the cell-free synthesis method to the large-scale preparation of a variety of heterodimeric complexes of disulfide-bonded proteins and determined their crystal structures, including the complexes such as (1) the extracellular domains (ECDs) of the calcitonin receptor-like receptor (CLR) and the receptor activity-modifying protein 2 (RAMP2) [the adrenomedullin 1 (AM_1) receptor] [59] and (2) the secreted homodimeric interleukin-5 (IL-5) and the IL-5 receptor α-subunit (IL-5RA) ECDs [60]. Human CLR (residues 23–136, including three pairs of Cys residues forming disulfide bonds) and human RAMP2 (residues 56–139, including two pairs of Cys residues forming disulfide bonds) were cloned into the TA vector pCR2.1TOPO (Life Technologies). The CLR and RAMP2 ECDs were produced as fusions with an N-terminal histidine tag and a TEV cleavage site. The selenomethionine-labeled proteins were synthesized by the *E. coli* cell-free method, using the large-scale dialysis mode [9, 22]. The CLR and RAMP2 ECDs both precipitated during synthesis. The precipitated proteins were denatured with 50 mM Tris-HCl buffer (pH 8.3), containing 8 M guanidine hydrochloride and 20 mM DTT, and were refolded together (co-refolded) by rapid dilution into 50 mM Tris-HCl buffer (pH 8.3), containing 1 M arginine hydrochloride, 5 mM reduced glutathione, and 0.5 mM oxidized glutathione. The co-refolded CLR•RAMP2 ECD complex was successfully purified to homogeneity by chromatography, after the affinity tags were enzymatically removed by TEV protease. The disulfide bonds were properly formed during the co-refolding process. By a similar method, human IL-5 (residues 23–134, including two pairs of Cys residues for disulfide bonding per subunit) and human IL-5RA (residues 21–335, including three pairs of Cys residues for disulfide bonding) were synthesized and co-refolded. The co-refolded IL-5•IL-5RA ECD complex was also successfully purified [60].

For larger proteins and more difficult complexes with numerous disulfide bonds, co-translational disulfide bonding is necessary. The openness of the cell-free system offers direct and flexible control of the reaction environment to promote proper disulfide-bond formation. Several groups have developed cell-free synthesis methods for disulfide-bonded proteins, based on crude extracts from *E. coli*, wheat germ, or insect cells [9, 22, 61–65]. To facilitate disulfide-bond formation, glutathione buffer is used to control the relatively oxidative environment. Usually, glutathione buffer is composed of 0–5 mM oxidized glutathione (glutathione-S-S-glutathione, GSSG) or a mixture of various ratios of oxidized glutathione (GSSG) and reduced glutathione (GSH). In addition, incorrectly formed disulfide bonds are reshuffled by the addition of 0.2–0.8 mg/ml disulfide isomerase, DsbC, or another protein disulfide isomerase (PDI). To favor disulfide bond formation, the reducing agent should be removed to maintain the oxidizing conditions. In fact, by the cell-free co-expression method, the above-mentioned CLR•RAMP2 ECD complex can be produced in the disulfide-bonded and soluble form without refolding (the final yield is 0.3 mg purified complex protein/ml reaction solution).

The cell-free synthesis method enables the efficient synthesis of antibody fragments by co-expression of the heavy chain (Hc) and light chain (Lc) genes encoding the Fv or Fab fragment. Several hundred micrograms of functional anti-human IL-23 single-chain Fv and anti-human IL-13α1R Fab fragment were produced from a 1 ml batch reaction [66, 67]. Intact mouse IgG1 against human creatine kinase was successfully produced, although the productivity was relatively low (0.5 μg/ml reaction) even with the dialysis-mode cell-free system [68]. Structural analyses require milligram quantities of protein samples. The batch-mode cell-free synthesis method is unable to produce sufficient amounts for this purpose. For large-scale disulfide-bonded protein production, the dialysis-mode cell-free synthesis method has been improved. For example, 3.3 mg/ml of human lysozyme-C was obtained from 1 ml reaction solution in 6 h [69]. This method can be applied to produce antibody fragments, including Fv, scFv, and Fab, for structural analysis.

Acknowledgments

We thank Mr. K. Ake, Ms. T. Imada, and Ms. T. Nakayama for their assistance in the manuscript preparation. This work was supported by the RIKEN Structural Genomics/Proteomics Initiative (RSGI), the National Project on Protein Structural and Functional Analyses, the Targeted Proteins Research Program, and the Platform for Drug Discovery, Informatics, and Structural Life Science, of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan (to S.Y.).

References

1. Zubay G (1973) In vitro synthesis of protein in microbial systems. Annu Rev Genet 7:267–287
2. Pratt JM (1984) Coupled transcription-translation in prokaryotic cell-free system. In: Hames BD, Higgins SJ (eds) Transcription and translation. IRL Press, Washington, DC, pp 179–209
3. Spirin AS, Baranov VI, Ryabova LA et al (1988) A continuous cell-free translation system capable of producing polypeptides in high yield. Science 242:1162–1164
4. Kigawa T, Yokoyama S (1991) A continuous cell-free protein synthesis system for coupled transcription-translation. J Biochem 110:166–168
5. Kigawa T, Muto Y, Yokoyama S (1995) Cell-free synthesis and amino acid-selective stable

isotope labeling of proteins for NMR analysis. J Biomol NMR 6:129–134

6. Kim DM, Kigawa T, Choi C-Y et al (1996) A highly efficient cell-free protein synthesis system from *Escherichia coli*. Eur J Biochem 239:881–886
7. Yabuki T, Kigawa T, Dohmae N et al (1998) Dual amino acid-selective and site-directed stable-isotope labeling of the human c-Ha-Ras protein by cell-free synthesis. J Biomol NMR 11:295–306
8. Kigawa T, Yabuki T, Yoshida Y et al (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett 442:15–19
9. Kigawa T, Yabuki T, Matsuda N et al (2004) Preparation of *Escherichia coli* cell extract for highly productive cell-free protein expression. J Struct Funct Genomics 5:63–68
10. Kigawa T (2010) Cell-free protein preparation through prokaryotic transcription-translation methods. Methods Mol Biol 607:1–10
11. Madin K, Sawasaki T, Ogasawara T et al (2000) A highly efficient and robust cell-free protein synthesis system prepared from wheat embryos: plants apparently contain a suicide system directed at ribosomes. Proc Natl Acad Sci U S A 97:559–564
12. Takai K, Endo Y (2010) The cell-free protein synthesis system from wheat germ. Methods Mol Biol 607:23–30
13. Takai K, Sawasaki T, Endo Y (2010) Practical cell-free protein synthesis system using purified wheat embryos. Nat Protoc 5(2):227–238
14. Tarui H, Imanishi S, Hara T (2000) A novel cell-free translation/glycosylation system prepared from insect cells. J Biosci Bioeng 90:508–514
15. Wakiyama M, Kaitsu Y, Yokoyama S (2006) Cell-free translation system from Drosophila S2 cells that recapitulates RNAi. Biochem Biophys Res Commun 343:1067–1071
16. Suzuki T, Ezure T, Ito M et al (2009) An insect cell-free system for recombinant protein expression using cDNA resources. Methods Mol Biol 577:97–108
17. Mikami S, Masutani M, Sonenberg N et al (2006) An efficient mammalian cell-free translation system supplemented with translation factors. Protein Expr Purif 46:348–357
18. Mikami S, Kobayashi T, Masutani M et al (2008) A human cell-derived *in vitro* coupled transcription/translation system optimized for production of recombinant proteins. Protein Expr Purif 62:190–198
19. Kim DM, Choi CH (1996) A semicontinuous prokaryotic coupled transcription/translation system using a dialysis membrane. Biotechnol Prog 12:645–649
20. Kigawa T, Yamaguchi-Nunokawa E, Kodama K et al (2002) Selenomethionine incorporation into a protein by cell-free synthesis. J Struct Funct Genomics 2:29–35
21. Matsuda T, Koshiba S, Tochio N et al (2007) Improving cell-free protein synthesis for stable-isotope labeling. J Biomol NMR 37:225–229
22. Kigawa T, Matsuda T, Yabuki T, et al (2008) Bacterial cell-free system for highly efficient protein synthesis. In: Spirin AS, Swartz JR (eds) Cell-free protein synthesis. Wiley-VCH, pp 83–97
23. Kigawa T, Inoue M, Aoki M, et al (2008) The use of the *Escherichia coli* cell-free protein synthesis for structural biology and structural proteomics. In: Spirin AS, Swartz JR (eds) Cell-free protein synthesis. Wiley-VCH, pp 99–109
24. Kigawa T (2010) Cell-free protein production system with the *E. coli* crude extract for determination of protein folds. Methods Mol Biol 607:101–111
25. Jackson AM, Boutell J, Cooley N et al (2003) Cell-free protein synthesis for proteomics. Brief Funct Genomic Proteomic 2:308–319
26. Carlson ED, Gan R, Hodgman CE et al (2012) Cell-free protein synthesis: applications come of age. Biothechnol Adv 30:1185–1194
27. Yabuki T, Motoda Y, Hanada K et al (2007) A robust two-step PCR method of template DNA production for high-throughput cell-free protein synthesis. J Struct Funct Genomics 8:173–191
28. Aoki M, Matsuda T, Tomo Y et al (2009) Automated system for high-throughput protein production using the dialysis cell-free method. Protein Expr Purif 68:128–136
29. Seki E, Matsuda N, Yokoyama S et al (2008) Cell-free protein synthesis system from *Escherichia coli* cells cultured at decreased temperatures improves productivity by decreasing DNA template degradation. Anal Biochem 377:156–161
30. Terada T, Murata T, Shirouzu M et al (2014) Cell-free expression of protein complexes for structural biology. Methods Mol Biol 1091:151–159
31. Matsuda T, Kigawa T, Koshiba S et al (2006) Cell-free synthesis of zinc-binding proteins. J Struct Funct Genomics 7:93–100
32. Wada T, Shirouzu M, Terada T et al (2003) Structure of a conserved CoA-binding protein synthesized by a cell-free system. Acta Crystallogr D Biol Crystallogr 59:1213–1218
33. Yokoyama J, Matsuda T, Koshiba S et al (2010) An economical method for producing stable-

isotope labeled proteins by the *E. coli* cell-free system. J Biomol NMR 48(4):193–201

34. Yokoyama J, Matsuda T, Koshiba S et al (2011) A practical method for cell-free protein synthesis to avoid stable isotope scrambling and dilution. Anal Biochem 411(2):223–229
35. Hirao I, Ohtsuki T, Fujiwara T et al (2002) An unnatural base pair for incorporating amino acid analogs into proteins. Nat Biotechnol 20:177–182
36. Kiga D, Sakamoto K, Kodama K et al (2002) An engineered Escherichia coli tyrosyl-tRNA synthetase for site-specific incorporation of an unnatural amino acid into proteins in eukaryotic translation and its application in a wheat germ cell-free system. Proc Natl Acad Sci U S A 99:9715–9720
37. Kodama K, Fukuzawa S, Nakayama H et al (2006) Regioselective carbon-carbon bond formation in proteins with palladium catalysis; new protein chemistry by organometallic chemistry. Chembiochem 7:134–139
38. Mukai T, Yanagisawa T, Ohtake K et al (2011) Genetic-code evolution for protein synthesis with non-natural amino acids. Biochem Biophys Res Commun 411:757–761
39. Mukai T, Hayashi A, Iraha F et al (2010) Codon reassignment in the *Escherichia coli* genetic code. Nucleic Acids Res 38:8188–8195
40. Yokoyama S, Hirota H, Kigawa T et al (2000) Structural genomics project in Japan. Nat Struct Biol 7(Suppl):943–945
41. Yokoyama S (2003) Protein expression systems for structural genomics and proteomics. Curr Opin Chem Biol 7:39–43
42. Yokoyama S, Terwilliger TC, Kuramitsu S et al (2007) RIKEN aids international structural genomics efforts. Nature 445:21
43. URL: http://www.genomics.agilent.com/article.jsp?pageId=484
44. Davanloo P, Rosenberg AH, Dunn JJ et al (1984) Cloning and expression of the gene for bacteriophage T7 RNA polymerase. Proc Natl Acad Sci U S A 81:2035–2039
45. Thomas JG, Ayling A, Baneyx F (1997) Molecular chaperones, folding catalysts, and the recovery of active recombinant proteins from *E. coli*. To fold or to refold. Appl Biochem Biotechnol 66:197–238
46. Kadokura H, Katzen F, Beckwith J (2003) Protein disulfide bond formation in prokaryotes. Annu Rev Biochem 72:111–135
47. Muller M, Koch HG, Beck K et al (2001) Protein traffic in bacteria: multiple routes from the ribosome to and across the membrane. Prog Nucleic Acid Res Mol Biol 66:107–157
48. Weski J, Ehrmann M (2012) Genetic analysis of 15 protein folding factors and proteases of the *Escherichia coli* cell envelope. J Bacteriol 194:3225–3233
49. Sakamoto K, Murayama K, Oki K et al (2009) Genetic encoding of 3-iodo-L-tyrosine in *Escherichia coli* for single-wavelength anomalous dispersion phasing in protein crystallography. Structure 17:335–344
50. Kainosho M, Torizawa T, Iwashita Y et al (2006) Optimal isotope labelling for NMR protein structure determinations. Nature 440:52–57
51. Hanawa-Suetsugu K, Kukimoto-Niino M, Mishima-Tsumagari C et al (2012) Structural basis for mutual relief of the Rac guanine nucleotide exchange factor DOCK2 and its partner ELMO1 from their autoinhibited forms. Proc Natl Acad Sci U S A 109:3305–3310
52. Kukimoto-Niino M, Sakamoto A, Kanno E et al (2008) Structural basis for the exclusive specificity of Slac2-a/melanophilin for the Rab27 GTPases. Structure 16:1478–1490
53. Arai S, Saijo S, Suzuki K et al (2013) Rotation mechanism of *Enterococcus hirae* V1-ATPase based on asymmetric crystal structures. Nature 493:703–707
54. Arai S, Yamato I, Shiokawa A et al (2009) Reconstitution *in vitro* of the catalytic portion (NtpA$_3$-B$_3$-D-G complex) of *Enterococcus hirae* V-type Na$^+$-ATPase. Biochem Biophys Res Commun 390:698–702
55. Saijo S, Arai S, Hossain KM et al (2011) Crystal structure of the central axis DF complex of the prokaryotic V-ATPase. Proc Natl Acad Sci U S A 108:19955–19960
56. Creighton TE (1988) Toward a better understanding of protein folding pathways. Proc Natl Acad Sci U S A 85:5082–5086
57. Paget MS, Buttner MJ (2003) Thiol-based regulatory switches. Annu Rev Genet 37:91–121
58. Sevier CS, Kaiser CA (2002) Formation and transfer of disulphide bonds in living cells. Nat Rev Mol Cell Biol 3:836–847
59. Kusano S, Kukimoto-Niino M, Hino N et al (2012) Structural basis for extracellular interactions between calcitonin receptor-like receptor and receptor activity-modifying protein 2 for adrenomedullin-specific binding. Protein Sci 21:199–210
60. Kusano S, Kukimoto-Niino M, Hino N et al (2012) Structural basis of interleukin-5 dimer recognition by its α receptor. Protein Sci 21:850–864
61. Goerke AR, Swartz JR (2008) Development of cell-free protein synthesis platforms for disulfide bonded proteins. Biotechnol Bioeng 99:351–367

62. Michel E, Wüthrich K (2012) Cell-free expression of disulfide-containing eukaryotic proteins for structural biology. FEBS J 279:3176–3184
63. Kawasaki T, Gouda MD, Sawasaki T et al (2003) Efficient synthesis of a disulfide-containing protein through a batch cell-free system from wheat germ. Eur J Biochem 270:4780–4786
64. Ezure T, Suzuki T, Shikata M et al (2007) Expression of proteins containing disulfide bonds in an insect cell-free system and confirmation of their arrangements by MALDI-TOF MS. Proteomics 7:4424–4434
65. Stech M, Merk H, Schenk JA et al (2012) Production of functional antibody fragments in a vesicle-based eukaryotic cell-free translation system. J Biotechnol 164:220–231
66. Yin G, Garces ED, Yang J, et al (2012) Aglycosylated antibodies and antibody fragments produced in a scalable *in vitro* transcription-translation system. MAbs 4(2)
67. Matsuda T, Furumoto S, Higuchi K et al (2012) Rapid biochemical synthesis of ^{11}C-labeled single chain variable fragment antibody for immuno-PET by cell-free protein synthesis. Bioorg Med Chem 20(22):6579–6582
68. Frey S, Haslbeck M, Hainzl O et al (2008) Synthesis and characterization of a functional intact IgG in a prokaryotic cell-free expression system. Biol Chem 389:37–45
69. Matsuda T, Watanabe S, Kigawa T (2013) Cell-free synthesis system suitable for disulfide-containing proteins. Biochem Biophys Res Commun 431:296–301

Part II

Purification and Crystallization of Membrane Proteins

Chapter 6

Overview of Membrane Protein Purification and Crystallization

Tatsuro Shimamura

Abstract

The three-dimensional structures of proteins provide important information for elucidation of the mechanisms and functions of the proteins. However, membrane proteins are difficult to crystallize and available structural information on membrane proteins is very limited. The difficulty is mainly due to the hydrophobic nature and the instability of membrane proteins, which increase some parameters in their purification and crystallization procedures. Recently, some new techniques such as the antibody technique and the lipidic cubic phase crystallization technique were applied to the production of high-quality crystals of membrane proteins. In this chapter, the protocols for the purification of the membrane protein and the lipidic cubic phase crystallization technique are described.

Keywords Membrane protein, Lipidic cubic phase, Crystallization, Detergent, In meso, Antibody

1 Introduction

Approximately 30 % of proteins encoded in the human genome are membrane proteins [1, 2]. They are involved in a variety of essential biological functions such as signal transduction, solute transport, and energy conversion. Despite their essential roles, membrane proteins are known to be difficult to crystallize compared with soluble proteins. Actually, of nearly 120,000 entries in the Protein Data Bank (PDB) [3], only around 610 structures are of unique integral membrane proteins [4]. The difficulty is mainly due to the hydrophobic nature and the instability of membrane proteins, which increase some parameters in the purification and crystallization procedures of membrane proteins [5, 6].

Membrane proteins are embedded within the lipid bilayer and very insoluble. The first step of the purification process is thus to solubilize the membrane protein from the membrane using detergent. Detergent molecules are amphiphilic with a polar head and a hydrophobic tail. At lower concentrations, the detergent molecules exist as monomers in aqueous solution. At the critical micelle

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks, DOI 10.1007/978-4-431-56030-2_6, © Springer Japan 2016

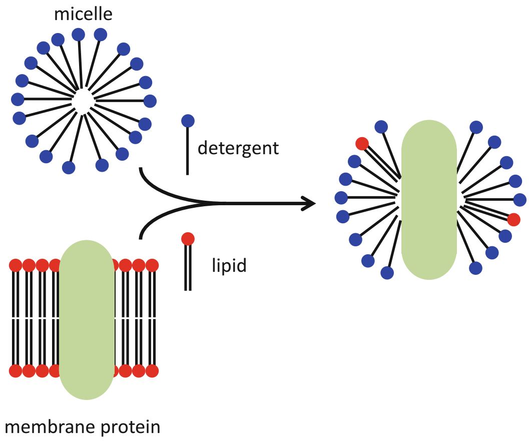


Fig. 1 Solubilization of membrane proteins

concentration (CMC), the detergent molecules begin to self-associate and form micelles [7]. When added to the membrane, the detergent molecules disrupt the membrane structure and cover the hydrophobic surface of the membrane protein, generating water-soluble protein-detergent micelles (Fig. 1). Nonionic sugar detergents such as maltosides and glucosides are most often used for membrane protein purification and crystallization (Table 1). The recently developed maltose neopentyl glycol (MNG) amphiphiles have shown effectively to stabilize several membrane proteins compared with conventional detergents, leading to successful crystallization [8–11]. Using these mild detergents, the membrane proteins are extracted from the membrane in their native conformation. Generally the concentration of the detergent required for the solubilization of membrane proteins is much higher than the CMC. For example, n-dodecyl-β-D-maltoside (DDM), one of the most popular sugar detergents whose CMC is ~0.0087 % (Table 1) in water, is used for solubilization at a concentration of 0.5–1 %. The concentration of the detergent can be decreased to two to three times higher than the CMC at later steps in the purification procedure. The detergent used for solubilization does not need to be the same as the detergent in the later steps of the purification and crystallization; it can be exchanged for another detergent during purification.

Except for solubilization, the purification procedures for membrane proteins are essentially the same as those for soluble proteins. Immobilized metal affinity chromatography (IMAC) is an efficient and high-speed method for the purification of membrane proteins [12, 13]. Cleavage of the affinity tag from the protein increases the likelihood of crystallization. The presence of detergent molecules may decrease the efficiency of proteolysis by blocking the access of the protease to the cleavage site or by inhibiting the protease

Table 1
Common detergents

	M_r	CMC[a]
Nonionic		
Glucosides or maltosides		
n-Octyl-β-D-glucoside	292.4	18–20 mM (0.53 %)
n-Nonyl-β-D-glucoside	306.4	6.5 mM (0.20 %)
n-Decyl-β-D-glucoside	320.4	2.2 mM (0.0070 %)
n-Nonyl-β-D-maltoside	468.5	6 mM (0.28 %)
n-Decyl-β-D-maltoside	482.6	1.8 mM (0.087 %)
n-Undecyl-β-D-maltoside	496.6	0.59 mM (0.029 %)
n-Dodecyl-β-D-maltoside	510.6	0.17 mM (0.0087 %)
n-Tridecyl-β-D-maltoside	524.6	0.033 mM (0.0017 %)
n-Octyl-β-D-thioglucoside	308.4	9.0 mM (0.28 %)
n-Nonyl-β-D-thioglucoside	322.4	2.9 mM (0.093 %)
n-Octyl-β-D-thiomaltoside	470.6	8.5 mM (0.4 %)
n-Nonyl-β-D-thiomaltoside	484.6	3.2 mM (0.15 %)
n-Decyl-β-D-thiomaltoside	498.6	0.9 mM (0.045 %)
n-Undecyl-β-D-thiomaltoside	512.7	0.21 mM (0.011 %)
n-Dodecyl-β-D-thiomaltoside	526.6	0.05 mM (0.0026 %)
CYGLU-4	318.4	1.8 mM (0.058 %)
CYMAL-5	494.5	2.4 mM (0.12 %)
CYMAL-6	508.5	0.56 mM (0.028 %)
CYMAL-7	522.5	0.19 % (0.0099 %)
Octyl glucose neopentyl glycol	569.7	1.02 mM (0.058 %)
Decyl maltose neopentyl glycol	949.1	0.036 mM (0.0034 %)
Lauryl maltose neopentyl glycol	1005.2	0.01 mM (0.001 %)
Polyoxyethylene glycols		
C8E4	306.5	8 mM (0.25 %)
C10E5	378.6	0.81 mM (0.031 %)
C10E6	423.0	0.9 mM (0.038 %)
C12E8	538.8	0.09 mM (0.0048 %)
C12E9	583.0	0.05 mM (0.003 %)
Triton X-100	avg. 647	0.23 mM (0.015 %)

(continued)

Table 1 (continued)

	M_r	CMC[a]
Zwitterionic detergents		
CHAPS	614.9	8 mM (0.49 %)
LDAO	229.4	1–2 mM (0.023 %)

[a]CMC values were from the Affymetrix Anatrace Products catalog [46]

activity [14]. These can be sometimes avoided by increasing the amount of protease, or by changing the location of the tag, or by inserting few hydrophilic amino acid residues between the protein and the tag to expose the cleavage site to the protease. Other chromatographic techniques such as size-exclusion chromatography and ion-exchange chromatography are also available. It should be remembered that the membrane proteins are covered by detergent molecules, which expands the hydrodynamic radius of the protein. Moreover, the membrane protein-detergent micelles tend to interact strongly with the chromatography matrix, lowering the column efficiency. The purity should be as high as possible but overpurification sometimes loses structural components such as subunits of membrane protein complexes or lipids [15]. Homogeneity of the purified protein can be assessed using size-exclusion chromatography. Monodispersity is a critical prerequisite for successful crystallization.

Once sufficient amounts (at least ~0.5 mg) of the membrane protein have been obtained with high purity and monodispersity, one can try to crystallize the protein. Membrane protein three-dimensional (3D) crystals are classified into two types, type I and type II (Fig. 2) [5, 6]. Type I crystals are built by stacks of two-dimensional (2D) crystals. The crystals formed in a lipidic cubic phase (LCP) belong to type I. Type II crystals are obtained using the standard crystallization methods routinely applied to soluble proteins, and most membrane crystals belong to this type. In type II crystals, only the hydrophilic surfaces of the membrane protein can be involved in the rigid crystal contacts. Therefore, the membrane proteins with small hydrophilic surfaces are especially difficult to crystallize. Moreover, the detergent molecules covering the hydrophobic surfaces need space in the crystal lattice, meaning that the crystals have a very high solvent content (65–80 %) and diffract poorly. These issues can be overcome in several ways [5, 6]. Firstly, the shorter alkyl chain detergents generally form smaller micelles, producing more surface area for the crystal contacts. It is recommended to solubilize the membrane protein using a longer chain detergent and exchange it for a shorter chain detergent. This is because the shorter chain detergent generally has a larger CMC

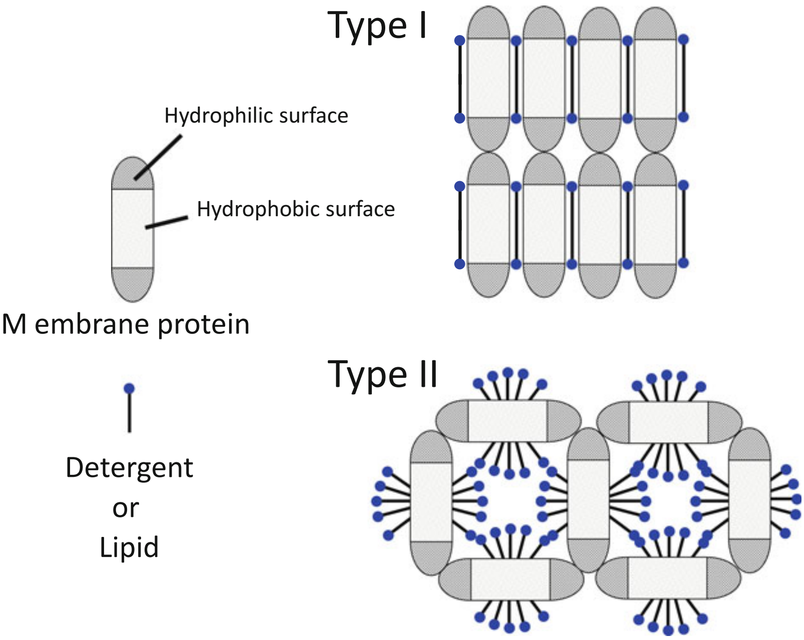


Fig. 2 Basic types of membrane protein 3D crystals

and is required at a higher concentration for solubilization, although the longer chain detergent with very low CMC is difficult to replace completely with the shorter chain detergent. In the case of the crystallographic study of Mhp1, a hydantoin transporter, DDM was used for solubilization and exchanged for n-nonyl-β-D-maltoside at the Ni-affinity chromatography step by washing extensively with buffer containing the detergent, which was essential for the successful crystallization of Mhp1 [16–18]. However, it should be noted that membrane proteins are less stable when covered by a shorter chain detergent. The use of the thermostabilized mutant is sometimes effective to overcome the instability as was shown in the structural study of the turkey β_1 adrenergic receptor [19]. The second way to reduce the micelle size is the addition of small amphiphilic molecules. For example, the addition of 5 % 1,2,3-heptanetriol has shown to reduce the number of *N*,*N*-Dimethyldodecylamine *N*-oxide (LDAO) associated with the reaction center from *Rhodopseudomonas viridis* [20]. The third way is to expand the hydrophilic surface by the specific binding of a soluble protein. As the soluble protein, antibody (Fv fragment, Fab fragment, nanobody), DARPin (designed ankyrin repeat protein) [21] and monobody (fibronectin type III domain) [22] have been used so far for crystallization. The antibody technique was first applied to crystallographic studies of cytochrome *c* oxidase (Fig. 3a, b) [23–25] and has been successfully used for the structure determination of several membrane proteins such as the cytochrome *bc*$_1$ complex [26], the KcsA potassium channel [27],

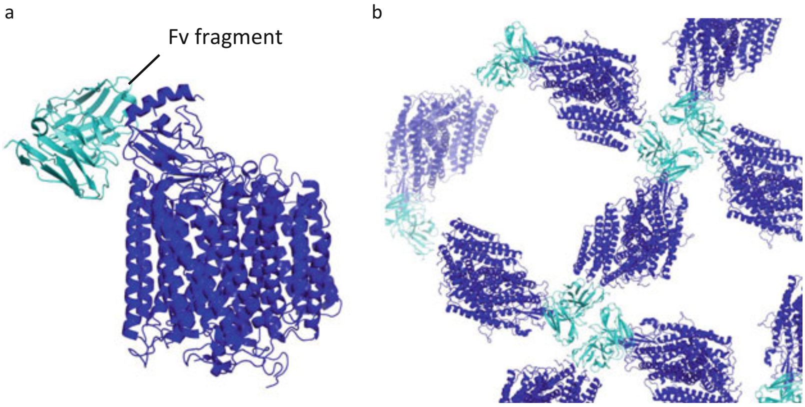


Fig. 3 Fv fragment essential for the crystallization of cytochrome *c* oxidase. (**a**) Structure of cytochrome *c* oxidase-Fv complex. Cytochrome *c* oxidase is shown in *blue* and Fv fragment *cyan*. (**b**) Crystal packing of cytochrome *c* oxidase-Fv complex

and adenosine A_{2a} receptor [28]. In the case of adenosine A_{2a} receptor, the Fab fragment recognizes the 3D structure of the receptor and contributes not only to the expansion of the hydrophilic area but also to the stabilization of the inactive conformation. Nanobody is a small single chain antibody of a llama and was used for structure determinations of the active conformation of G protein-coupled receptors [9, 29, 30].

Crystallization in lipidic mesophase (also known as LCP crystallization or *in meso* crystallization) has been successfully used to determine the high-resolution structures of membrane proteins since it was first applied to the crystallization of bacteriorhodopsin [31]. The first step of the technique is to reconstitute the purified membrane protein in the lipid bilayer prepared by mixing aqueous buffer with lipids such as monoolein under appropriate conditions. Addition of salts and precipitants may produce tiny crystals in LCP. The LCP technique has several advantages compared with the crystallization technique in detergent micelles. First, the membrane protein is more stable in a more native-like environment [32]. Second, the crystals in LCP belong to type I and crystal contacts are established by the hydrophobic surface as well as the hydrophilic surface of the protein, resulting in lower solvent content and higher crystal quality [33]. However, the LCP technique has some disadvantages. The crystals in LCP are generally very tiny and difficult to detect. Moreover, the curved nature of the lipid membrane and the specific microstructure sets a limit to the size of membrane proteins to be crystallized [34]. This obstacle can be overcome by the use of specific precipitants, such as nonvolatile alcohols, small PEGs, etc., that swell and transform LCP to a sponge phase [34–36] or special lipids that enlarge the size of the water channel in LCP [9].

Here, protocols for the membrane protein purification and the LCP crystallization techniques are presented. These are essentially the same methods used for the crystallographic study of human histamine H_1 receptor [37, 38]. Other textbooks [5, 6], papers [39, 40], a web site [41], movies [42, 43], and manufacturer's manuals [44–46] also provide very useful information about these techniques.

2 Materials

1. Glass beads (0.5 mm diameter).
2. n-Dodecyl-β-D-maltoside (DDM).
3. Breaking buffer: 50 mM HEPES pH7.5, 120 mM NaCl, 5 % glycerol, 2 mM EDTA, and one tablet of protein inhibitor cocktail/50 ml.
4. Lysis buffer: 10 mM HEPES pH7.5, 10 mM $MgCl_2$, 20 mM KCl, and one tablet of protein inhibitor cocktail/50 ml.
5. High salt buffer: 10 mM HEPES pH7.5, 10 mM $MgCl_2$, 20 mM KCl, 1 M NaCl, and one tablet of protein inhibitor cocktail/50 ml.
6. Membrane buffer: 50 mM HEPES pH7.5, 120 mM NaCl, 20 % glycerol, and one tablet of protein inhibitor cocktail/50 ml.
7. Iodoacetamide.
8. Solubilization buffer: 50 mM HEPES pH7.5, 500 mM NaCl, 20 % glycerol, 1 % DDM, and one tablet of protein inhibitor cocktail/50 ml.
9. Imidazole.
10. Talon resin.
11. Talon wash buffer 1: 50 mM HEPES pH7.5, 500 mM NaCl, 10 % glycerol, 0.025 % DDM, 20 mM imidazole, 10 mM $MgCl_2$, 8 mM ATP, and one tablet of protein inhibitor cocktail/50 ml.
12. Talon wash buffer 2: 50 mM HEPES pH7.5, 500 mM NaCl, 10 % glycerol, 0.025 % DDM, 20 mM imidazole, and one tablet of protein inhibitor cocktail/50ml.
13. Talon elution buffer: 50 mM HEPES pH7.5, 500 mM NaCl, 10 % glycerol, 0.025 % DDM, 200 mM imidazole, and one tablet of protein inhibitor cocktail/50ml.
14. PD10 desalting column.
15. Ni-sepharose resin.

16. Ni-elute buffer: 20 mM HEPES pH7.5, 500 mM NaCl, 10 % glycerol, 0.025 % DDM, 400 mM imidazole, and one tablet of protein inhibitor cocktail/50 ml.
17. Reverse IMAC buffer: 50 mM HEPES pH7.5, 500 mM NaCl, 10 % glycerol, 0.025 % DDM, and one tablet of protein inhibitor cocktail/100 ml.
18. Ni-sepharose high-performance resin.
19. BCA protein assay kit.
20. Monoolein.
21. Crystallization screens.
22. 100 μl Hamilton gas-tight syringe.
23. Coupler.

3 Methods

3.1 Membrane Preparation from Pichia pastoris *(see Note 1)*

1. Harvest cells by centrifuging at 5000 *g* for 5 min at 4 °C.
2. Discard the supernatant.
3. Resuspend the cells in cold water. A paintbrush is helpful when resuspending the cells.
4. Harvest cells by centrifuging at 5000 *g* for 5 min at 4 °C.
5. Resuspend 20–25 g of cell pellets in 100 ml of the breaking buffer.
6. Take a 10 μl sample of the resuspension to check the cell disruption and store at 4 °C.
7. Transfer the resuspension to a 2 L flask.
8. Add 150 g of glass beads to the flask.
9. Place the flask on an incubator shaker and disrupt the cells by shaking at 350 rpm at 4 °C for ~2 h.
10. Take a 10 μl sample of the homogenate and check the cell disruption by comparing it with the sample from step 6 using a microscope. More than 90 % of cells should be disrupted.
11. Transfer the homogenate to clean centrifuge tubes chilled on ice.
12. Remove intact cells and particles by centrifugation at 2000 *g* for 20 min at 4 °C.
13. Transfer the supernatant to clean ultracentrifugation tubes chilled on ice.
14. Balance the tubes and ultracentrifuge the supernatant at 100,000 *g* for 30 min at 4 °C.
15. Discard the supernatant.

16. Resuspend the pellets in 100 ml of lysis buffer.
17. Transfer the suspension to ultracentrifugation tubes.
18. Ultracentrifuge the suspension at 100,000 *g* for 30 min at 4 °C.
19. Discard the supernatant.
20. Resuspend the pellets in 100 ml of the high salt buffer.
21. Transfer the suspension to ultracentrifugation tubes.
22. Ultracentrifuge the suspension at 100,000 *g* for 30 min at 4 °C.
23. Discard the supernatant.
24. Repeat steps 20–23. The resultant membrane pellets are used for the purification.

3.2 Solubilization of the Membrane Protein

1. Resuspend ~10 g of the membrane pellets in 25 ml of the membrane buffer.
2. Transfer the resuspension to a clean chilled dounce homogenizer.
3. Add iodoacetamide (10 mg/ml).
4. Dounce ~40 times on ice.
5. Transfer the resuspension to a clean chilled beaker.
6. Keep the beaker on ice for 30 min.
7. Pour 50 ml of the solubilization buffer into the membrane suspension and stir gently at 4 °C for ~2 h (See Note 2).
8. Transfer the solubilization mixture to ultracentrifugation tubes.
9. Ultracentrifuge the solubilization mixture at 100,000 *g* for 30 min at 4 °C to remove the unsolubilized material.
10. Pool the supernatant in a clean chilled beaker.

3.3 First Affinity Purification Using Talon Resin [45]

1. Add imidazole (final 5 mM) and NaCl (final 800 mM) in the supernatant from step 10 in 3.2.
2. Add 10 ml of Talon resin equilibrated with the Talon wash buffer 1.
3. Agitate the mixture with a magnetic stir bar at 4 °C for 3–12 h.
4. Collect the Talon resin in a 50 ml Falcon tube by repeating centrifugation at 800 g and discarding the supernatant at 4 °C.
5. Wash the Talon resin with 10 bed volumes of the Talon wash buffer 1. Add the Talon wash buffer 1 in the Falcon tubes and agitate gently on a rotary shaker at 4 °C.
6. Centrifuge at 800 *g* for 5 min at 4 °C.
7. Discard the supernatant.

8. Repeat steps 6–8 with 10 bed volumes of the Talon wash buffer 1.
9. Wash the resin with the Talon wash buffer 2. Repeat steps 6–8 with 5 bed volumes of the Talon wash buffer 2.
10. Add 30 ml of the Talon wash buffer 2 in the tube and resuspend by vortexing.
11. Load the resin into a 20 ml gravity-flow column with the bottom outlet capped.
12. Remove the bottom cap and allow the buffer to drain. Save flow-through for SDS-PAGE analysis.
13. Elute the His-tagged protein by loading 10 ml of the Talon elution buffer on the resin in the column.
14. Collect the eluate in a 15 ml disposable tube.
15. Repeat steps 14–15 ten times.
16. Analyze the fractions by SDS-PAGE.
17. Pool the fractions containing the His-tagged protein.
18. Concentrate to 2.5 ml with a 100 kDa molecular weight cutoff concentrator.

3.4 Remove Imidazole Using PD10 Column

1. Take off the top cap of the PD10 column and remove the storage solution.
2. Cut the sealed end of the column.
3. Equilibrate the column with ~30 ml of the Talon wash buffer 2.
4. Apply the 2.5 ml of the concentrated sample to the column.
5. Let the sample enter the packed bed completely and discard the flow-through.
6. Place a 15 ml tube under the column for sample collection.
7. Load the 3.5 ml of the Talon wash buffer 2 to the column.
8. Collect the eluate.

3.5 Second Affinity Purification Using Ni-Sepharose Resin [44]

1. Equilibrate 4 ml of Ni-NTA resin with ~20 ml of the Talon wash buffer 2.
2. Add 4 ml of the Ni-NTA resin to the eluate from step 8 in 3.4.
3. Gently agitate on a rotary shaker at 4 °C for 2–12 h.
4. Transfer the mixture to a 20 ml disposable column.
5. Wash the resin by applying 45 ml of the Talon wash buffer 2.
6. Elute the His-tagged protein by with 24 ml of the Ni-elute buffer.
7. Collect the eluate and analyze the fractions by SDS-PAGE.
8. Pool the fractions containing the His-tagged protein.

9. Concentrate to 2.5 ml with a 100 kDa molecular weight cutoff concentrator.

3.6 Remove Imidazole Using PD10 Column

1. Take off the top cap of the PD10 column and remove the storage solution.
2. Cut the sealed end of the column.
3. Equilibrate the column with ~30 ml of the reverse IMAC buffer.
4. Apply the 2.5 ml of the concentrated sample to the column.
5. Let the sample enter the packed bed completely and discard the flow-through.
6. Place a 15 ml tube under the column for sample collection.
7. Load the 3.5 ml of the reverse IMAC buffer to the column.
8. Collect the eluate.
9. To cleave off the GFP-His-tag of the protein, add appropriate amount of His-tagged TEV protease to the eluate and incubate overnight at 4 °C.

3.7 Reverse IMAC (Immobilized-Metal Affinity Chromatography)

1. Equilibrate 0.6 ml of the Ni-sepharose high-performance resin with 6 ml of the reverse IMAC buffer.
2. Pour 0.6 ml of the Ni-sepharose high-performance resin into a 10 ml column.
3. Apply the protein mixture from step 9 in 3.6 on the column.
4. Collect the flow-through fraction (see Note 3).
5. Apply 8 ml of the reverse IMAC buffer to the column.
6. Collect the flow-through and add to the fraction from step 4.
7. Determine the protein concentration using a BCA protein assay kit following the manufacturer's instructions.
8. Concentrate the purified protein to ~30 mg/ml with a 100 kDa molecular weight cutoff concentrator.
9. Check the purity and monodispersity of the purified sample by SDS-PAGE and size-exclusion chromatography (See Note 4).

3.8 Lipidic Cubic Phase Formation [40–43]

1. Take out monoolein from the freezer and melt it using a heating block at ~40 °C. It takes ~10 min.
2. Remove needles, Teflon ferrules, and plungers from the two 100 μl Hamilton gas-tight syringes (Fig. 4a).
3. Centrifuge the purified membrane protein solution in a 1.5 ml tube at 20,400 g for ~10 min at 4 °C to remove aggregates and collect the supernatant.
4. Check the volume of the supernatant at step 3.

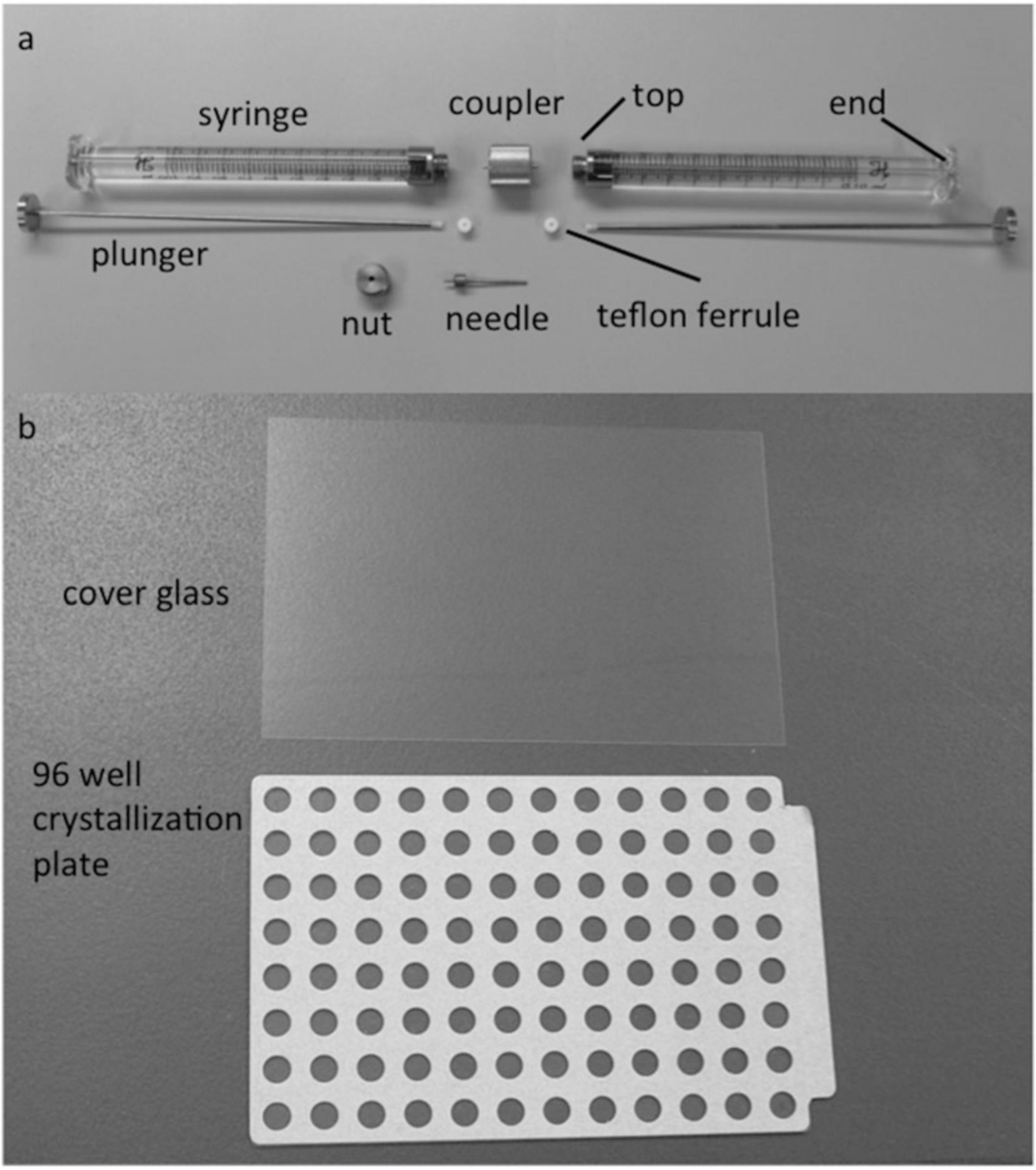


Fig. 4 Apparatus for LCP crystallization. (**a**) Two syringes, two plungers, two teflon ferrules, a coupler, a needle and a nut. (**b**) A cover glass and a 96 well crystallization plate

5. Calculate the required amount of monoolein. The volume of monoolein needs to be ~150 % of the protein solution volume to form the cubic phase (see Note 5).
6. Using a 20 μl or 100 μl pipette and the appropriate disposable tip, put the required amount of melted monoolein into one Hamilton gas-tight syringe from the bottom end (Fig. 4a).
7. Insert a plunger from the bottom end of the syringe. The Teflon part of the plunger will contact with the monoolein.

8. Hold the syringe vertically and push the plunger till monoolein reaches the top end of the syringe. This manipulation will remove air bubbles from the lipid.
9. Using a 20 μl or 100 μl pipette and the appropriate disposable tip, put the membrane protein solution into the other Hamilton gas-tight syringe from the bottom end. Be careful not to trap air bubbles, although air bubbles appear easily because the protein solution contains detergents.
10. Insert a plunger from the bottom of the syringe. The Teflon part of the plunger will contact with protein solution.
11. Hold the syringe vertically and push the plunger until the protein solution reaches the top end of the syringe. If air bubbles are trapped in the solution, remove them by moving the plunger up and down. If this fails, collect the protein solution from the syringe and centrifuge it to remove air bubbles and start again from step 9.
12. Place the Teflon ferrules in the syringes.
13. Connect two syringes by a coupler (Fig. 5a, b).
14. Push slowly the plunger of the syringe that stores the protein solution and transfer all the protein solution into the other

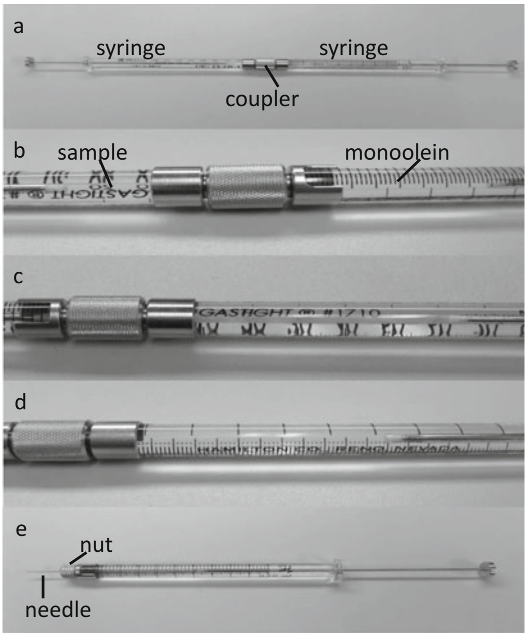


Fig. 5 Outline of the LCP formation. (**a**) Connect two syringes using a coupler. (**b**) One syringe has sample solution and the other has monoolein. (**c**) When mixed, the mixture is clouded. (**d**) Homogeneous cubic phase. (**e**) The syringe with a needle ready for crystallization

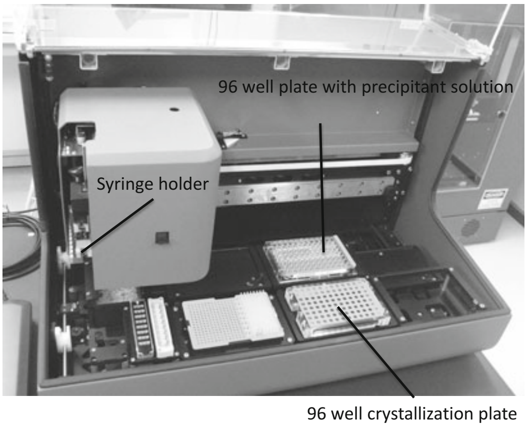

96 well plate with precipitant solution
Syringe holder
96 well crystallization plate

Fig. 6 LCP crystallization robot

syringe through the coupler. This will form a partly clouded mixture (Fig. 5c). Steps 14–16 should be performed at 20 °C.

15. Slowly push the plunger of the syringe that stores the protein solution and monoolein, and transfer them into the other syringe through the coupler.
16. Repeat moving the plungers back and force more than 100 times till the mixture forms a transparent homogeneous cubic phase (Fig. 5d) (See Note 6).

3.9 Crystallization [40–43]

1. Transfer the mixture to one of the two syringes.
2. Disconnect the coupler with the empty syringe from the syringe but keep the Teflon ferrule.
3. Set a needle at the top end of the syringe (Fig. 5e).
4. Place the syringe in the syringe holder of a LCP robot (Fig. 6).
5. Put the 96-well crystallization plate (Fig. 4b) and the 96-well plate containing precipitant solutions in their proper positions on the robot (Fig. 6) (See Note 7).
6. Start the robot. Control the humidity using a humidifier. The volumes of the mesophase and the precipitant solution should be set at 30–50 nl and ~800 nl, respectively.
7. When the robot finishes dispensing, place a cover glass on the crystallization plate (Fig. 4b).
8. Keep the plate in a 20 °C incubator.
9. Check all wells in the plate regularly under a microscope. Use crossed-polarizers as well as normal light. Typically crystals appear in a week, although it may range from a few hours to 2 months (Fig. 7).

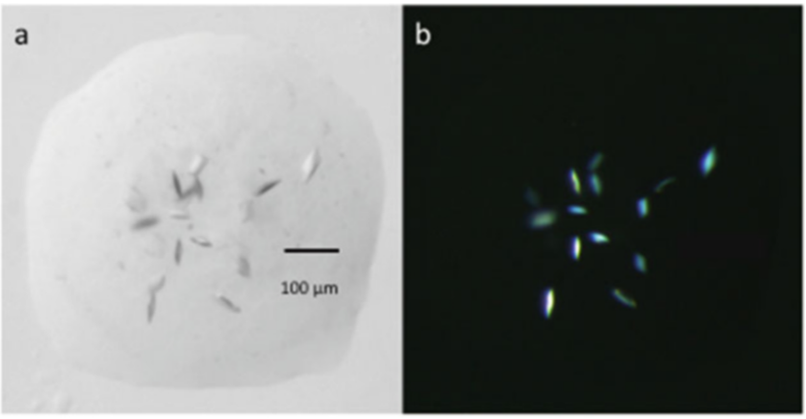


Fig. 7 Crystals of a membrane protein in LCP. Crystals were observed using normal light (**a**) or cross polarizers (**b**)

4 Notes

1. Sonication does not work for the disruption of yeast cells because the cell wall of yeast is harder than bacteria.
2. If an inhibitor, a substrate, or a ligand of the target protein is available, it is advisable to add that compound in the buffers used in the purification steps because the compound will stabilize the protein in a certain conformation and increase the possibility of crystallization.
3. After the TEV protease digestion, the target protein has no His-tag and comes to the flow-through fraction.
4. Purity is less important for the LCP method because LCP can act as a size filter and remove large-size contaminants and protein aggregates [34, 47].
5. For example, if the volume of the membrane protein solution is 20 μl, the volume of monoolein should be 30 μl. Note that monoolein has a density of 0.942 g/ml at 20 °C [48].
6. The recommended rate of mixing is ~1 stroke per second or slower. Faster mixing can raise the temperature of the mixture due to frictional heating which destabilizes the protein.
7. Homemade screens are used. Typically, precipitant solutions contain 30 ~40 % low-molecular-weight PEG (PEG200, PEG300, PEG400, PEG600, PEG500MME, PEG500DME etc.), salts, and buffer at pH 6–8. Note that monoolein is unstable at lower or higher pH.

Acknowledgments

This work was supported partly by JSPS KAKENHI Grant Numbers 26102725 and 15H04338.

References

1. Almen MS, Nordstrom KJ, Fredriksson R, Schioth HB (2009) Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biol 7:50. doi:10.1186/1741-7007-7-50
2. Fagerberg L, Jonasson K, von Heijne G, Uhlen M, Berglund L (2010) Prediction of the human membrane proteome. Proteomics 10 (6):1141–1149. doi:10.1002/pmic.200900258
3. http://www.rcsb.org/
4. http://blanco.biomol.uci.edu/mpstruc/
5. Iwata S (2003) Methods and results in crystallization of membrane proteins. Internat'l University Line, La Jolla
6. Carola Hunte GJ, Schägger H (2003) Membrane protein purification and crystallization, a practical guide. Academic Press, San Diego
7. Garavito RM, Ferguson-Miller S (2001) Detergents as tools in membrane biochemistry. J Biol Chem 276(35):32403–32406. doi:10.1074/Jbc.R100031200
8. Chae PS, Rasmussen SGF, Rana RR, Gotfryd K, Chandra R, Goren MA, Kruse AC, Nurva S, Loland CJ, Pierre Y, Drew D, Popot JL, Picot D, Fox BG, Guan L, Gether U, Byrne B, Kobilka B, Gellman SH (2010) Maltose-neopentyl glycol (MNG) amphiphiles for solubilization, stabilization and crystallization of membrane proteins. Nat Methods 7 (12):1003–U1090. doi:10.1038/Nmeth.1526
9. Rasmussen SGF, DeVree BT, Zou YZ, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, Mathiesen JM, Shah STA, Lyons JA, Caffrey M, Gellman SH, Steyaert J, Skiniotis G, Weis WI, Sunahara RK, Kobilka BK (2011) Crystal structure of the beta(2) adrenergic receptor-Gs protein complex. Nature 477(7366):549–U311. doi:10.1038/Nature10361
10. Kruse AC, Hu JX, Pan AC, Arlow DH, Rosenbaum DM, Rosemond E, Green HF, Liu T, Chae PS, Dror RO, Shaw DE, Weis WI, Wess J, Kobilka BK (2012) Structure and dynamics of the M3 muscarinic acetylcholine receptor. Nature 482(7386):552–556. doi:10.1038/Nature10867
11. Rollauer SE, Tarry MJ, Graham JE, Jaaskelainen M, Jager F, Johnson S, Krehenbrink M, Liu SM, Lukey MJ, Marcoux J, McDowell MA, Rodriguez F, Roversi P, Stansfeld PJ, Robinson CV, Sansom MS, Palmer T, Hogbom M, Berks BC, Lea SM (2012) Structure of the TatC core of the twin-arginine protein transport system. Nature 492(7428):210–214. doi:10.1038/nature11683
12. Hemdan ES, Porath J (1985) Development of immobilized metal affinity-chromatography. 2. Interaction of amino-acids with immobilized nickel iminodiacetate. J Chromatogr 323 (2):255–264. doi:10.1016/S0021-9673(01)90388-7
13. Hemdan ES, Porath J (1985) Development of immobilized metal affinity-chromatography. 3. Interaction of oligopeptides with immobilized nickel iminodiacetate. J Chromatogr 323 (2):265–272. doi:10.1016/S0021-9673(01)90389-9
14. Mohanty AK, Simmons CR, Wiener MC (2003) Inhibition of tobacco etch virus protease activity by detergents. Protein Expr Purif 27(1):109–114. doi:Pii S1046-5928(02)00589-2. doi:10.1016/S1046-5928(02)00589-2
15. Hunte C, Richers S (2008) Lipids and membrane protein structures. Curr Opin Struct Biol 18(4):406–411. doi:10.1016/j.sbi.2008.03.008
16. Weyand S, Shimamura T, Yajima S, Suzuki S, Mirza O, Krusong K, Carpenter EP, Rutherford NG, Hadden JM, O'Reilly J, Ma P, Saidijam M, Patching SG, Hope RJ, Norbertczak HT, Roach PCJ, Iwata S, Henderson PJF, Cameron AD (2008) Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. Science 322 (5902):709–713. doi:10.1126/Science.1164440
17. Shimamura T, Yajima S, Suzuki S, Rutherford NG, O'Reilly J, Henderson PJF, Iwata S (2008) Crystallization of the hydantoin transporter Mhp1 from microbacterium liquefaciens. Acta Crystallogr F 64:1172–1174. doi:10.1107/S1744309108036920
18. Shimamura T, Weyand S, Beckstein O, Rutherford NG, Hadden JM, Sharples D, Sansom

MSP, Iwata S, Henderson PJF, Cameron AD (2010) Molecular basis of alternating access membrane transport by the sodium-hydantoin transporter Mhp1. Science 328 (5977):470–473. doi:10.1126/Science.1186303

19. Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AG, Tate CG, Schertler GF (2008) Structure of a beta(1)-adrenergic G-protein-coupled receptor. Nature 454(7203):486–491
20. Gast P, Hemelrijk P, Hoff AJ (1994) Determination of the number of detergent molecules associated with the reaction-center protein isolated from the photosynthetic bacterium rhodopseudomonas-viridis – effects of the amphiphilic molecule 1,2,3-heptanetriol. FEBS Lett 337(1):39–42. doi:10.1016/0014-5793(94)80625-X
21. Sennhauser G, Amstutz P, Briand C, Storchenegger O, Grutter MG (2007) Drug export pathway of multidrug exporter AcrB revealed by DARPin inhibitors. PLoS Biol 5 (1):106–113. doi:10.1371/journal.pbio.0050007, ARTN e7
22. Lu M, Symersky J, Radchenko M, Koide A, Guo Y, Nie RX, Koide S (2013) Structures of a Na+-coupled, substrate-bound MATE multidrug transporter. Proc Natl Acad Sci U S A 110 (6):2099–2104. doi:10.1073/Pnas.1219901110
23. Iwata S, Ostermeier C, Ludwig B, Michel H (1995) Structure at 2.8-Angstrom resolution of cytochrome-C-oxidase from paracoccus-denitrificans. Nature 376(6542):660–669. doi:10.1038/376660a0
24. Ostermeier C, Iwata S, Ludwig B, Michel H (1995) F-V fragment mediated crystallization of the membrane-protein bacterial cytochrome-C-oxidase. Nat Struct Biol 2 (10):842–846. doi:10.1038/Nsb1095-842
25. Ostermeier C, Harrenga A, Ermler U, Michel H (1997) Structure at 2.7 angstrom resolution of the Paracoccus denitrificans two-subunit cytochrome c oxidase complexed with an antibody F-V fragment. Proc Natl Acad Sci U S A 94(20):10547–10553. doi:10.1073/Pnas.94.20.10547
26. Hunte C, Koepke J, Lange C, Rossmanith T, Michel H (2000) Structure at 2.3 angstrom resolution of the cytochrome bc(1) complex from the yeast Saccharomyces cerevisiae co-crystallized with an antibody Fv fragment. Struct Fold Des 8(6):669–684. doi:10.1016/S0969-2126(00)00152-0
27. Zhou YF, Morais-Cabral JH, Kaufman A, MacKinnon R (2001) Chemistry of ion coordination and hydration revealed by a K+ channel-Fab complex at 2.0 angstrom resolution. Nature 414(6859):43–48. doi:10.1038/35102009
28. Hino T, Arakawa T, Iwanari H, Yurugi-Kobayashi T, Ikeda-Suno C, Nakada-Nakura Y, Kusano-Arai O, Weyand S, Shimamura T, Nomura N, Cameron AD, Kobayashi T, Hamakubo T, Iwata S, Murata T (2012) G-protein-coupled receptor inactivation by an allosteric inverse-agonist antibody. Nature 482(7384):237–U130. doi:10.1038/Nature10750
29. Rasmussen SG, Choi HJ, Fung JJ, Pardon E, Casarosa P, Chae PS, Devree BT, Rosenbaum DM, Thian FS, Kobilka TS, Schnapp A, Konetzki I, Sunahara RK, Gellman SH, Pautsch A, Steyaert J, Weis WI, Kobilka BK (2011) Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. Nature 469(7329):175–180. doi:10.1038/nature09648, nature09648 [pii]
30. Kruse AC, Ring AM, Manglik A, Hu J, Hu K, Eitel K, Hubner H, Pardon E, Valant C, Sexton PM, Christopoulos A, Felder CC, Gmeiner P, Steyaert J, Weis WI, Garcia KC, Wess J, Kobilka BK (2013) Activation and allosteric modulation of a muscarinic acetylcholine receptor. Nature 504(7478):101–106. doi:10.1038/nature12735
31. Landau EM, Rosenbusch JP (1996) Lipidic cubic phases: a novel concept for the crystallization of membrane proteins. Proc Natl Acad Sci U S A 93(25):14532–14535. doi:10.1073/Pnas.93.25.14532
32. Liu W, Hanson MA, Stevens RC, Cherezov V (2010) LCP-Tm: an assay to measure and understand stability of membrane proteins in a membrane environment. Biophys J 98 (8):1539–1548. doi:10.1016/j.bpj.2009.12.4296, S0006-3495(09)06148-7 [pii]
33. Caffrey M (2009) Crystallizing membrane proteins for structure determination: use of lipidic mesophases. Annu Rev Biophys 38:29–51. doi:10.1146/Annurev.Biophys.050708.133655
34. Cherezov V (2011) Lipidic cubic phase technologies for membrane protein structural studies. Curr Opin Struct Biol 21(4):559–566. doi:10.1016/J.Sbi.2011.06.007
35. Wadsten P, Wohri AB, Snijder A, Katona G, Gardiner AT, Cogdell RJ, Neutze R, Engstrom S (2006) Lipidic sponge phase crystallization of membrane proteins. J Mol Biol 364(1):44–53. doi:10.1016/J.Jmb.2006.06.043
36. Cherezov V, Clogston J, Papiz MZ, Caffrey M (2006) Room to move: crystallizing membrane

proteins in swollen lipidic mesophases. J Mol Biol 357(5):1605–1618. doi:S0022-2836(06) 00078-7 [pii] 10.1016/j.jmb.2006.01.049

37. Shiroishi M, Kobayashi T, Ogasawara S, Tsujimoto H, Ikeda-Suno C, Iwata S, Shimamura T (2011) Production of the stable human histamine H(1) receptor in Pichia pastoris for structural determination. Methods 55(4):281–286. doi:10.1016/j.ymeth.2011.08.015

38. Shimamura T, Shiroishi M, Weyand S, Tsujimoto H, Winter G, Katritch V, Abagyan R, Cherezov V, Liu W, Han GW, Kobayashi T, Stevens RC, Iwata S (2011) Structure of the human histamine H1 receptor complex with doxepin. Nature 475(7354):65–70. doi:10.1038/nature10236. nature10236 [pii]

39. Newby ZER, O'Connell JD, Gruswitz F, Hays FA, Harries WEC, Harwood IM, Ho JD, Lee JK, Savage DF, Miercke LJW, Stroud RM (2009) A general protocol for the crystallization of membrane proteins for X-ray structural investigation. Nat Protoc 4(5):619–637. doi:10.1038/Nprot.2009.27

40. Caffrey M, Cherezov V (2009) Crystallizing membrane proteins using lipidic mesophases. Nat Protoc 4(5):706–731. doi:10.1038/Nprot.2009.31

41. Cherezov V http://cherezov.usc.edu

42. Caffrey M, Porter C (2010) Crystallizing membrane proteins for structure determination using lipidic mesophases. J Visualized Exp: JoVE (45) e1712. doi:10.3791/1712

43. Liu W, Cherezov V (2011) Crystallization of membrane proteins in lipidic mesophases. J Visualized Exp: JoVE (49) e2501. doi:10.3791/2501

44. QIAGEN (2003) A handbook for high-level expression and purification of 6xHis-tagged proteins. QIAGEN, Hilden

45. TALON Metal Affinity Resins User Manual

46. Affymetrix Anatrace Products catalog

47. Kors CA, Wallace E, Davies DR, Li L, Laible PD, Nollert P (2009) Effects of impurities on membrane-protein crystallization in different systems. Acta Crystallogr D 65:1062–1073. doi:10.1107/S0907444909029163

48. Lide DR (2008) CRC handbook of chemistry and physics, 89th edn. CRC Press, Boca Raton

Chapter 7

Cell-Free Synthesis of Membrane Proteins

Tomomi Kimura-Someya, Toshiaki Hosaka, Takehiro Shinoda, Kazumi Shimono, Mikako Shirouzu, and Shigeyuki Yokoyama

Abstract

Among the various membrane protein synthesis methods available today, the cell-free protein synthesis method is a relatively new tool. Recent technological advances in the lipid and detergent conditions for the cell-free synthesis of membrane proteins have enabled the robust production of various membrane proteins for structural biology. In this chapter, we describe representative conditions for the production of high quantities of membrane proteins by the cell-free method, with crystallization quality. We also discuss examples of membrane proteins that were successfully synthesized by the cell-free method. The crystals of these highly purified proteins resulted in solved crystal structures.

Keywords Cell-free protein synthesis, Membrane protein, Lipid, Crystal structure determination

1 Introduction

The cell-free protein synthesis method is used to synthesize proteins in vitro, with the cellular translation machinery. The reaction solution for cell-free protein synthesis contains a cell extract, the template DNA or messenger RNA, substrates such as amino acids, and other components (see Fig. 1 in Chap. 5). The original cell-free protein synthesis method was developed long before the advent of recombinant protein expression with host-vector systems, and synthesis systems with rabbit reticulocyte, wheat germ, *Escherichia coli* extracts, etc. were used mainly for the small-scale preparation of proteins, particularly those labeled with a radioactive amino acid. The cell-free protein synthesis method has been drastically improved over the past 20 years, and has become one of the standard methods of protein sample preparation for structural biology, because of a variety of advantages over other expression methods (see Chap. 5). For uses in structural biology analyses by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, which require large amounts of highly pure and homogeneous proteins, the cell-free protein synthesis method

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks, DOI 10.1007/978-4-431-56030-2_7, © Springer Japan 2016

using the *E. coli* cell extract is by far the most frequently chosen, among those using various cell extracts.

The cell-free protein synthesis method is recognized as one of the best methods for the preparation of integral membrane proteins [1, 2]. Unlike recombinant protein expression in host cells, the cell-free system for protein synthesis is not encapsulated in cells, by definition, and therefore lacks the cell membrane. The function of the cell-free synthesis system may be modified, simply by adding the necessary components. Additional components required for the cell-free synthesis of integral membrane proteins are lipids and/or detergents, as the membrane proteins synthesized in the absence of lipid-detergent aggregate and precipitate, due to the hydrophobicity of their transmembrane regions [3, 4]. The membrane proteins synthesized in the cell-free system may form micelles with detergents, and liposomes and nanodiscs with lipids, by embedding their transmembrane regions in the hydrophobic environments provided by the detergents/lipids. Furthermore, certain membrane proteins may require specific boundary lipids for the formation and maintenance of their proper functional structures in the lipid bilayer. Consequently, it is important to develop techniques to supply these materials, according to their roles, to the cell-free protein synthesis system.

In both bacterial and eukaryotic cells, many integral membrane proteins are inserted into the membrane by the translocon, a membrane protein complex that translocates polypeptides through membranes. However, the common cell-free methods do not require translocons for the integration of the synthesized membrane protein into the lipid-detergent complex. In our cell-free synthesis method with both detergent and lipid, the nascent polypeptides synthesized in the cell-free system are co-translationally integrated into the lipid bilayer environment of membrane fragments, which then assemble gradually into larger membranes, such as liposomes. This observation indicated that the edges of the membrane fragments, which are possibly covered with detergents, serve as the entrance for the polypeptides to become integrated into the lipid bilayer environment, instead of the translocon. On the other hand, in eukaryotic cells, certain types of receptors and related proteins are considered to be integrated, after synthesis, into membrane domains, such as lipid rafts and detergent-resistant membranes (DRMs), which are rich in cholesterol and glycolipids [5]. These membrane proteins may interact with specific lipids such as cholesterol, which is easily included in the cell-free protein synthesis system. These aspects of the cell-free protein synthesis system have facilitated the folding of many membrane proteins into their native structures.

When membrane proteins are expressed in host cells by recombinant DNA techniques, they are usually solubilized from the cellular membrane fractions with detergents. If the detergents are too harsh, then the native structures of the membrane proteins will be destroyed. In contrast, if the detergents are too mild, the recoveries of the expressed membrane proteins in the soluble fractions will be very low. Therefore, the detergent employed for the solubilization of the membrane proteins from the cell membrane should be carefully chosen, usually by extensive screening of a variety of detergents. In particular, when membrane proteins are integrated in DRMs, which are resistant to nonionic detergents such as Triton X-100, the choice of a detergent for solubilization without denaturation is quite difficult. On the other hand, the cell-free synthesis method does not require the use of harsh detergents for solubilization and therefore allows a broader choice of the optimal detergents specific to the individual target membrane protein. This is another strong advantage of the cell-free synthesis method.

The recombinant membrane proteins expressed in cells are usually purified by chromatography, after solubilization from the membrane fractions. However, it is often difficult to remove the contaminating membrane proteins derived from the host cells. On the other hand, in the cell-free synthesis method, the targeted membrane proteins are generally the only proteins synthesized and inserted into the lipid-detergent complexes, although a number of soluble proteins exist in the cell-free reaction solution. This is the reason why the cell-free synthesized membrane proteins can be highly purified. When membrane proteins are highly overexpressed in *E. coli* cells, inclusion bodies may be generated. In these cases, the purity of the product may be very high, as long as the membrane proteins recovered from the inclusion bodies with denaturants, such as urea, guanidine hydrochloride, and sodium dodecyl sulfate, are properly refolded into the native structures. In contrast, in the case of the cell-free synthesis method, the refolding of the synthesized membrane proteins or the solubilization of precipitants with harsh denaturants is usually unnecessary.

There are other advantages of cell-free protein synthesis, as compared to cell-based expression methods. The expression levels of membrane proteins in cells are often low, probably due to the limited capacity of the cell for membrane protein accommodation and the cytotoxicity due, for example, to improper cell membrane structures generated by an excess of inserted membrane proteins. The cytotoxicity of a membrane protein in recombinant expression often prevents not only its large-scale preparation but also biochemical analyses of its functions. On the other hand, the success rate of cell-free membrane protein synthesis is rather high, since cytotoxicity is not a problem. In addition, in the cell-free synthesis system, it is easy to regulate the subunit composition for multisubunit complexes of membrane proteins, simply by adjusting the concentrations of the DNA templates encoding the subunits to reproduce the proper subunit stoichiometry. The preparation of membrane protein samples with the full selenomethionine substitution of methionine residues, in order to determine the phase in crystallography, is a straightforward procedure using cell-free synthesis. PCR-amplified DNA fragments may be used directly as the templates in the cell-free protein synthesis system, without cloning into plasmid vector, which is useful for high-throughput screening of conditions and/or robotic automation of synthesis.

This chapter describes the practical aspects of the cell-free membrane protein synthesis methods and their application to the structure determinations of integral membrane proteins.

2 Cell-Free Synthesis Methods for Membrane Proteins

2.1 Cell-Free Protein Synthesis Reactions

The cell-free protein synthesis reaction can be performed in the batch, dialysis, or bilayer mode (see Fig. 1 in Chap. 5). In the batch mode, the synthesis is performed by incubating the reaction solution in a container until the reaction rate slows down, due to the exhaustion of substrates and/or energy sources. To prolong the reaction for higher yield, the dialysis (or continuous-exchange cell-free, CECF) mode [6, 7] and the bilayer mode [8] have been developed to supplement the reaction solution with components, such as substrate amino acids and energy sources, from the feeding solution. In the dialysis mode, the reaction and feeding solutions are separated by a dialysis membrane. In the case of the large-scale synthesis of a certain sample for structural biology, a dialysis bag containing the inner, reaction solution (e.g., 3–9 mL) is placed in the outer, feeding solution (tenfold larger volume than that of the reaction solution). On the other hand, the screening of a large number of constructs, with respect to checking the yield and/or the biochemical properties, can be performed by small-scale synthesis (ca. 30 μL) in a multi-well format. In the bilayer mode, the low-density, feeding solution is laid directly on the high-density, reaction solution, and the two solutions are gradually mixed with each other during the course of the reaction. Our group primarily uses the dialysis mode for the cell-free synthesis of membrane proteins.

In the cell-free protein synthesis system, the extract prepared from *E. coli* cells is practically superior, with respect to both quantity and quality, to other commercially available extracts, such as the animal [9, 10], plant [11], and reconstituted systems, for the preparation of membrane proteins from not only bacteria but also eukaryotes, e.g., humans. Usually, the S30 fraction of the *E. coli* extract (the supernatant fraction obtained after cell disruption and centrifugation at 30,000 × g) is used in the cell-free protein synthesis system. Translation is coupled with transcription by T7 phage RNA polymerase from the T7 promoter in the template DNA, encoding the target membrane protein without the signal peptide sequence, in the form of a plasmid. Detergents and/or lipids are added to the reaction solution and/or the feeding solution for membrane protein synthesis. In the following Sects. (2.2, 2.3, 2.4, and 2.5), more details of the methods are described, according to the different uses of detergents and lipids.

2.2 Membrane Protein Synthesis in the Presence of Detergents

In this method, membrane protein synthesis is performed in the presence of a detergent at a concentration higher than the critical micelle concentration (CMC). In the dialysis mode, the detergent is added to both the reaction and feeding solutions. As the hydrophobic domains of the membrane protein are synthesized, they interact immediately with the hydrophobic portion of the detergent molecules. Thus, the detergent molecules surround the protein, exposing the hydrophilic portions of the molecules and generating "solubilized" membrane proteins (Fig. 1(1)). After ultracentrifugation, the protein is recovered from the supernatant. We simply refer to this synthesis method as "the detergent method."

Since the type and the concentration of the detergent greatly affect the yield and folding of proteins, a systematic screening of a panel of detergents should be performed to determine the

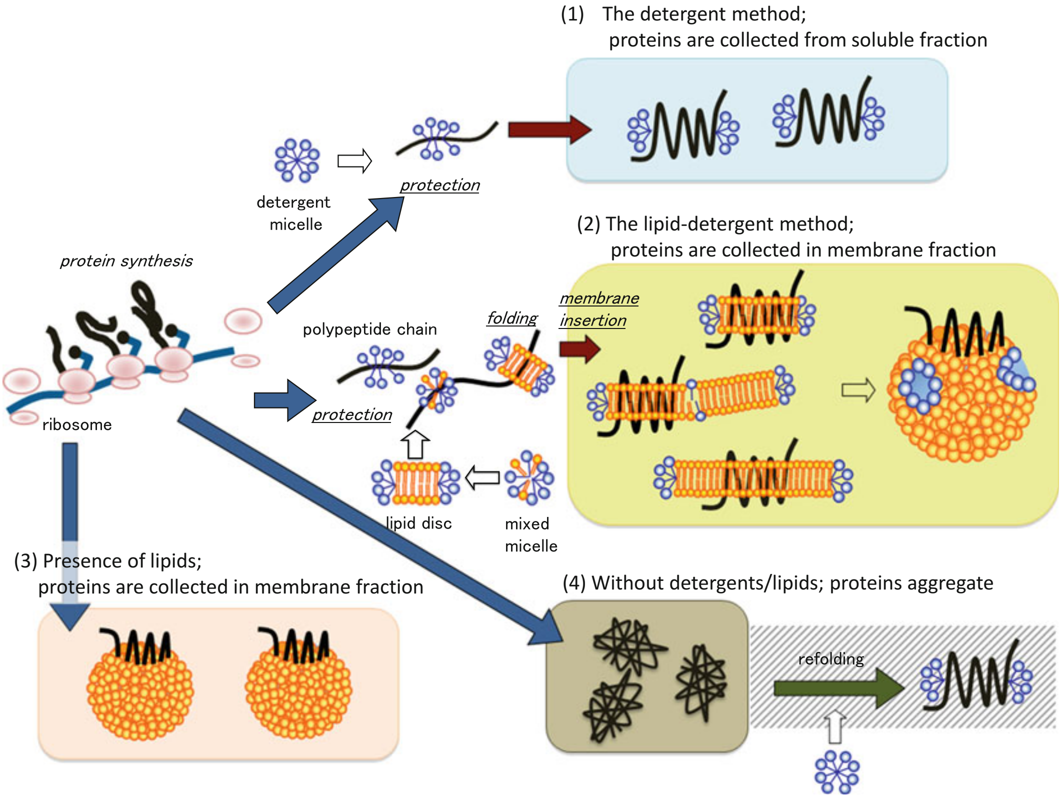

(1) The detergent method;
proteins are collected from soluble fraction
detergent micelle
protection
protein synthesis
ribosome
(2) The lipid-detergent method;
proteins are collected in membrane fraction
polypeptide chain
folding
membrane insertion
protection
lipid disc
mixed micelle
(3) Presence of lipids;
proteins are collected in membrane fraction
(4) Without detergents/lipids; proteins aggregate
refolding

Fig. 1 Illustration of the detergent method (*1*), the lipid-detergent method (*2*), and the conditions in the presence of lipids (*3*) and without detergents/lipids (*4*) for cell-free membrane protein synthesis

appropriate experimental conditions. Any detergents may be tested, provided they do not interfere with protein synthesis [3]. Each detergent, at a concentration above its CMC, is mixed with ~30 μL of a small-scale reaction solution, to assess the yield and precipitation. After suitable detergent types are identified, the optimum concentrations are investigated. The detergent used in cell-free membrane protein synthesis can be replaced by another detergent in the subsequent purification process. However, it is not always possible to completely exchange the detergents, and consequently, the residual detergent from the synthesis reaction may affect the stability of the protein during crystallography. Therefore, even in the cell-free synthesis process, it is better to select detergents that can be used in the subsequent purification and crystallization processes. Thus, we usually test digitonin, Brij-78, Brij-35, and *n*-dodecyl-β-*D*-maltopyranoside (DDM).

After optimization of the synthesis conditions, the scale of the cell-free synthesis may be increased to 9/90 mL (reaction solution/feeding solution). The yield is usually sufficient to obtain milligram quantities of the membrane protein, if the amino acid sequence lacks problematic regions that affect the protein synthesis machinery. The specific activity of the synthesized protein should be examined, if a standard protein is available. In addition, the correctly folded proteins may be purified by ligand or substrate affinity chromatography.

2.3 Membrane Protein Synthesis in the Presence of Detergents and Lipids

To synthesize membrane proteins that require either a lipid bilayer environment or a particular lipid to maintain their structure, activity, and stability, we developed a cell-free synthesis system that uses lipids with detergents [4, 12]. A suspension of lipids and one or more detergents, at concentrations above the CMCs, is added to the reaction solution, but not to the external feeding solution, at the initiation of protein synthesis. First, the synthesized membrane protein molecules form mixed micelles with the detergent and lipid molecules. The detergent in the reaction solution diffuses through the dialysis membrane to the external feeding solution, thereby reducing its concentration in the reaction solution. Therefore, the number of detergent molecules present in the detergent/lipid micelles gradually decreases, and fragments of lipid bilayer membrane (or lipid bilayer discs surrounded by a ring of detergent) are formed. Concurrently, the polypeptide chains synthesized and surrounded by micelles are incorporated into the membrane fragments, which fuse together to form larger fragments and eventually liposomes (lipid bilayer vesicles) containing the membrane proteins. The membrane proteins in this form are collected by ultracentrifugation as pellets (Fig. 1(2)). We here designate this method as "the lipid-detergent method." In principle, it is possible to separate the target membrane proteins from the soluble proteins, including those required for transcription and translation and those derived from the S30 fraction of the *E. coli* extract. In reality, some other lipophilic proteins bound to the membranous lipid structures are also collected, as protein contaminants. Nonetheless, the degree of contamination is markedly smaller than that of the pre-ultracentrifugation sample, and therefore, the sample is a good starting material for purification and crystallization.

Detergents should be selected based on their effects on both protein synthesis and liposome formation. In many cases, detergents derived from steroids, e.g., digitonin, sodium cholate, and CHAPS (3-[(3-cholamidopropyl)-dimethylammonio]-1-propane sulfonate), are appropriate for use in the lipid-detergent method [12]. However, systematic screening is still necessary to determine the optimal concentrations.

Next, the most appropriate lipid components are selected, to prevent the target protein from losing its activity or folding incorrectly. The crystal structures of membrane proteins purified from natural materials have revealed the presence of lipids at specific locations [13]. The stability of such proteins can be improved when lipids are supplied during synthesis. If it is unclear whether lipids are needed, or which type of lipid is needed, then the best strategy is to test natural lipid extracts containing different lipid species. The first choice may be an organ extract from the organism from which the target protein was isolated. If applicable, thin-layer chromatography and mass spectrometry can be used to analyze the lipid content in the target protein isolated from the natural membrane, to identify the required lipid components. Heterologous expression systems, such as *E. coli* and insect cells, may not contain all of the lipid components needed to stabilize the target membrane protein. The limited variety of the lipid compositions in such expression systems often makes expression and purification difficult; however, the cell-free synthesis method allows the addition and assessment of various lipid compositions to mimic natural environments.

The state of the lipids added to the reaction solution is also important. To form a finely dispersed lipid suspension, the lipids must be ultrasonically dispersed and mixed with the optimal detergent. If the lipid suspension is poorly prepared, then the synthesized membrane protein will not be properly integrated within the membrane bilayer. The simple addition of liposomes into the cell-free reaction solution will not allow the proper and efficient integration of the synthesized membrane protein into the lipid bilayer. In the lipid-detergent method, the efficiencies of protein folding and integration into the membrane are drastically increased, due to liposome formation concurrent with protein synthesis.

2.4 Membrane Protein Synthesis in the Presence of Lipids

Inverted membrane vesicles and natural membrane vesicles, such as microsomes, have been used in place of liposomes (Fig. 1(3)). This method is different from that described in Sect. 2.3, because no detergent is used. In a previous study, inverted membrane vesicles from *E. coli* were used to synthesize a functional form of tetracycline transporter, a membrane protein derived from *E. coli*, under cell-free conditions [14]. To prepare a protein sample for crystallography, the use of lipid fractions with no protein contaminants, such as artificial liposomes, is advantageous for subsequent purification. The PURE system, a reconstituted cell-free protein synthesis system containing liposomes and factors that promote membrane insertion, such as signal recognition particle (SRP), is also available [15].

Although it is not a lipid-only system, nanodiscs (also known as nanolipoprotein particles) have been attracting attention. The synthesis of many functional G protein-coupled receptors (GPCRs) by the nanodisc method has been reported [16]. The large amount of apolipoproteins, which comprise nanodiscs, is problematic for crystallography. However, this nanodisc-based cell-free synthesis method can be used for some structural analyses of membrane proteins, e.g., nuclear magnetic resonance imaging [17].

2.5 Membrane Protein Synthesis Without Detergents/Lipids

This method synthesizes membrane proteins in aggregates without using any detergents or lipids, and the synthesized proteins are obtained from a pellet after low-speed centrifugation (Fig. 1(4)). It is unlikely that the proteins collected in the pellet fraction are correctly folded, because of nonspecific hydrophobic interactions between the hydrophobic domains of the proteins. However, the original protein activity can be reconstructed if an appropriate detergent is used to solubilize the protein and form liposomes [18], indicating that the correct protein folding depends on the experimental conditions. According to the review by Katzen et al., unlike inclusion bodies in *E. coli* [19], the protein aggregates formed by cell-free synthesis can be easily solubilized by detergent, because the protein-protein interactions are relatively weak. However, when a structural analysis is planned, the detergent method and the lipid-detergent method, which both enhance correct protein folding during synthesis, are more advantageous for synthesizing the proteins than the synthesis without detergent/lipid method, as the latter requires an extensive search for the optimal refolding conditions.

3 Crystallography of Proteins Synthesized by the Cell-Free Synthesis Methods

3.1 Acetabularia Rhodopsins I and II

Acetabularia rhodopsins I (ARI) and II (ARII) are microbial-type rhodopsins, membrane proteins with seven α-helical transmembrane domains, derived from a eukaryotic unicellular organism, the marine alga *Acetabularia acetabulum*. It was difficult to overexpress ARI and ARII in *E. coli*. However, we used the lipid-

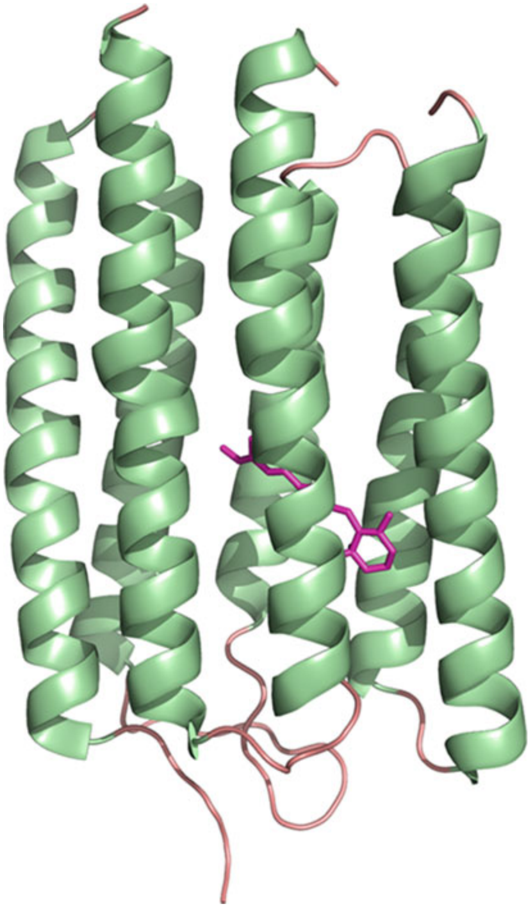


Fig. 2 Model of ARII from *Acetabularia acetabulum*, based on the crystal structure. ARII, synthesized by the cell-free method, was purified and its structure was solved by crystallography (PDB ID: 3AM6)

detergent method containing the essential pigment all-*trans* retinal, to achieve large-scale cell-free synthesis, biochemical and biophysical analyses, and protein crystallography [20]. The synthesis was performed in the presence of 0.05–0.8 % digitonin as the detergent source and egg yolk lecithin (L-α-phosphatidylcholine, 6.7 mg/mL) as the lipid source. The ARII protein was isolated from the pellet fraction after ultracentrifugation, because the majority of ARII was present in the membrane faction, rather than the soluble fraction. In addition, with 0.4 % digitonin, the largest fraction of the synthesized protein was incorporated into the liposomes. At lower detergent concentrations, the protein production was high, but the synthesized proteins precipitated without being incorporated into the liposomes. The membrane fraction was solubilized using DDM, and ARII was purified and crystallized in the presence of DDM. Although the crystal structure (Fig. 2, PDB ID: 3AM6) was similar to the previously determined structure of bacteriorhodopsin, we observed several structural features specific to ARII. This is the first crystal structure of a membrane protein that was synthesized by our lipid-detergent method. We tested the ARII protein in biochemical experiments and confirmed its proton transport activity [21].

We also constructed a system to overproduce correctly folded ARI by the cell-free protein synthesis method, using very similar conditions to those for ARII synthesis. We were able to obtain a large amount of the highly purified protein by the above-described simple purification methods. We performed the biophysical analysis of the light-driven proton pump mechanism during the photochemical reaction of ARI, using the cell-free synthesis product, and also crystallized ARI by the lipidic mesophase method. As the result of the X-ray crystallography analysis, the structure has been determined at 1.52–1.80 Å resolution [22] (PDB IDs: 5AWZ, 5AX0, 5AX1), which was the third highest-resolution structure, among the numerous structures of microbial rhodopsins in the dark state. The existence of abundant water molecules was confirmed in the large cavity on the proton-releasing side, which explained the relatively low pK_a of the proton-releasing residue. These results indicated that for membrane proteins, the cell-free protein synthesis methods, and particularly the lipid-detergent method, provide large amounts of high-quality samples. Thus, we were able to obtain the high-resolution crystal structure. This is a good example of the utility of the cell-free synthesis methods for structural-functional studies of membrane proteins.

3.2 Proteorhodopsin

Proteorhodopsin (PR), which was first discovered in the metagenomic uncultivated SAR86 group prokaryotes (*γ-proteobacteria*) in a DNA library from Monterey Bay, California, contains at least seven transmembrane α-helices and a retinal molecule that is covalently bound via a Schiff base to the side chain of a lysine residue [23, 24]. Currently, over 4,000 PR gene sequence variants have been deposited in the GenBank database [25–29]. The sequence identity between PR and bacteriorhodopsin is approximately 30 % [23]. Proteorhodopsin is a light-harvesting proton pump and thus could play an important role in solar energy transduction in the biosphere [24, 30–35]. However, biochemical and electrophysiological investigations have progressed slowly. Moreover, the structural analysis of PR was not performed until recently.

We have embarked on research toward the structural analysis of the marine γ-proteobacterium PR protein, from an ocean isolate, by the *E. coli* cell-free synthesis method. The lipid-detergent method was used for the cell-free synthesis of the PR protein. Almost all of the PR protein was embedded within liposomes. PR was purified by affinity chromatography, protease digestion of the affinity tag and gel filtration. We thus obtained about 10 mg of purified PR from a 9/90 mL cell-free synthesis reaction. Crystals

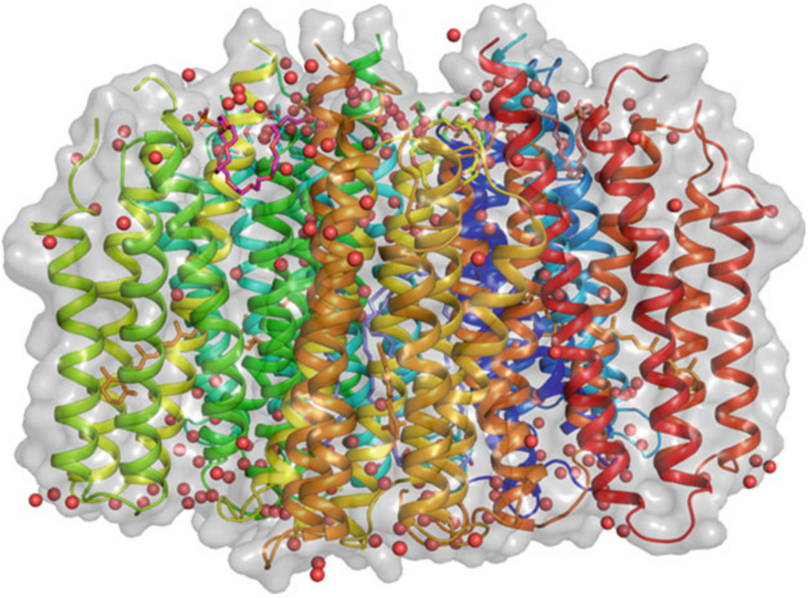


Fig. 3 Pentameric structure of proteorhodopsin determined at 2.0 Å resolution

suitable for X-ray diffraction were obtained from the purified PR samples, and the crystal structure was solved at 2.0 Å resolution (Fig. 3, Hosaka et al. manuscript in preparation).

3.3 Microbial Multidrug Efflux Protein EmrE, Purified from the Insoluble Fraction

Chen et al., of the Scripps Research Institute in the United States, determined the crystal structure of EmrE, a four-transmembrane multidrug transporter from *E. coli*, using a cell-free expression system that enabled the facile labeling of proteins with selenomethionine. The protein was synthesized in the absence of detergents, but was solubilized using *n*-nonyl-β-*D*-glucopyranoside (NG), followed by purification, crystallization, and crystallography [36]. They also used an *E. coli* cell-based expression system to synthesize non-labeled EmrE, which was purified and crystallized in a similar manner. Their study revealed that the crystal structures of the cell-free and cell-based expressed EmrE are nearly identical. Furthermore, the substrate-binding activity and affinity are similar between the two proteins, suggesting that the EmrE solubilized in NG folded correctly.

4 Conclusion

We have reviewed the methods currently used for the cell-free synthesis of integral membrane proteins for structural biology. We also described some of the successful crystallographic studies performed with proteins generated by the cell-free system derived from *E. coli* cells. Several *E. coli* cell-free protein synthesis kits are now commercially available, including the Remarkable Yield Translation System Kit (ProteinExpress, Chiba, Japan). The cell-free protein synthesis methods provide a variety of advantages in terms of both quantity and quality for structural biology, as compared to the conventional recombinant expression in host cells, including *E. coli*, insect, and mammalian cells. Therefore, we believe that the cell-free protein synthesis methods will be more extensively used for the preparation and crystallography of integral membrane proteins, including human GPCRs and channels.

Acknowledgments

This work was supported by the National Project on Protein Structural and Functional Analyses, the Targeted Proteins Research Program, and the Platform for Drug Discovery, Informatics, and Structural Life Science from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan (to S.Y.).

References

1. Junge F, Schneider B, Reckel S et al (2008) Large-scale production of functional membrane proteins. Cell Mol Life Sci 65 (11):1729–1755
2. Bernhard F, Tozawa Y (2013) Cell-free expression-making a mark. Curr Opin Struct Biol 23(3):374–380
3. Ishihara G, Goto M, Saeki M et al (2005) Expression of G protein coupled receptors in a cell-free translational system using detergents and thioredoxin-fusion vectors. Protein Expr Purif 41:27–37
4. Shimono K, Goto M, Kikukawa T et al (2009) Production of functional bacteriorhodopsin by an *Escherichia coli* cell-free protein synthesis system supplemented with steroid detergent and lipid. Protein Sci 18:2160–2171
5. Pike LJ (2003) Lipid rafts: bringing order to chaos. J Lipid Res 44(4):655–667
6. Kigawa T, Yabuki T, Yoshida Y et al (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett 442(1):15–19
7. Kim DM, Choi CY (1996) A semi-continuous prokaryotic coupled transcription/translation system using a dialysis membrane. Biotechnol Prog 12(5):645–649
8. Sawasaki T, Hasegawa Y, Tsuchimochi M et al (2002) A bilayer cell-free protein synthesis system for high-throughput screening of gene products. FEBS Lett 514(1):102–105
9. Mikami S, Kobayashi T, Masutani M et al (2008) A human cell-derived *in vitro* coupled transcription/translation system optimized for production of recombinant proteins. Protein Expr Purif 62(2):190–198
10. Mikami S, Masutani M, Sonenberg N et al (2006) An efficient mammalian cell-free translation system supplemented with translation factors. Protein Expr Purif 46(2):348–357
11. Takai K, Sawasaki T, Endo Y (2010) Practical cell-free protein synthesis system using purified wheat embryos. Nat Protoc 5:227–238
12. Kimura-Soyema T, Shirouzu M, Yokoyama S (2014) Cell-free membrane protein expression. Methods Mol Biol 1118:267–273
13. Tsukihara T, Aoyama H, Yamashita E et al (1996) The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. Science 272:1136–1144
14. Wuu JJ, Swartz JR (2008) High yield cell-free production of integral membrane proteins without refolding or detergents. Biochim Biophys Acta 1778:1237–1250
15. Kuruma Y, Nishiyama K, Shimizu Y et al (2005) Development of a minimal cell-free translation system for the synthesis of presecretory and integral membrane proteins. Biotechnol Prog 21:1243–1251
16. Katzen F, Fletcher JE, Yang J-P et al (2008) Insertion of membrane proteins into discoidal membranes using a cell-free protein expression approach. J Proteome Res 7:3535–3542

17. Raschle T, Hiller S, Yu T-Y et al (2009) Structural and functional characterization of the integral membrane protein VDAC-1 in lipid bilayer nanodiscs. J Am Chem Soc 131:17777–17779
18. Keller T, Schwarz D, Bernhard F et al (2008) Cell free expression and functional reconstitution of eukaryotic drug transporters. Biochemistry 47:4552–4564
19. Katzen F, Peterson TC, Kudlicki W (2009) Membrane protein expression: no cells required. Trends Biotechnol 27:455–460
20. Wada T, Shimono K, Kikukawa T et al (2011) Crystal structure of the eukaryotic light-driven proton-pumping rhodopsin, *Acetabularia* rhodopsin II, from marine alga. J Mol Biol 411:986–998
21. Kikukawa T, Shimono K, Tamogami J et al (2011) Photochemistry of *Acetabularia* rhodopsin II from a marine plant, *Acetabularia acetabulum*. Biochemistry 50:8888–8898
22. Furuse M, Tamogami J, Hosaka T et al (2015) Structural basis for the slow photocycle and late proton release in Acetabularia rhodopsin I from the marine plant Acetabularia acetabulum. Acta Crystallogr D Biol Crystallogr 71:2203–2216
23. Béjà O, Aravind L, Koonin EV et al (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289:1902–1906
24. Béjà O, Spudich EN, Spudich JL et al (2001) Proteorhodopsin phototrophy in the ocean. Nature 411:786–789
25. Venter JC, Remington K, Heidelberg JF et al (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74
26. Sabehi G, Massana R, Bielawski JP et al (2003) Novel proteorhodopsin variants from the Mediterranean and Red Seas. Environ Microbiol 5:842–849
27. Rusch DB, Halpern AL, Sutton G et al (2007) The sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical pacific. PLoS Biol 5:e77
28. de la Torre JR, Christianson LM, Béjà O et al (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proc Natl Acad Sci U S A 100:12830–12835
29. Sabehi G, Béjà O, Suzuki MT et al (2004) Different SAR86 subgroups harbour divergent proteorhodopsins. Environ Microbiol 6:903–910
30. Fuhrman JA, Schwalbach MS, Stingl U (2008) Proteorhodopsins: an array of physiological roles? Nat Rev Microbiol 6:488–494
31. Yoshizawa S, Kawanabe A, Ito H et al (2012) Diversity and functional analysis of proteorhodopsin in marine Flavobacteria. Environ Microbiol 14:1240–1248
32. Gómez-Consarnau L, González JM, Coll-Lladó M et al (2007) Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. Nature 445:210–213
33. González JM, Fernández-Gómez B, Fernàndez-Guerra A et al (2008) Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria). Proc Natl Acad Sci U S A 105:8724–8729
34. Martinez A, Bradley AS, Waldbauer JR et al (2007) Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. Proc Natl Acad Sci U S A 104:5590–5595
35. Kralj JM, Spudich EN, Spudich JL et al (2008) Raman spectroscopy reveals direct chromophore interactions in the Leu/Gln105 spectral tuning switch of proteorhodopsins. J Phys Chem B 18(37):11770–11706
36. Chen YJ, Pornillos O, Lieu S et al (2007) X-ray structure of EmrE supports dual topology model. Proc Natl Acad Sci U S A 104:18999–19004

Part III

Crystallization and Crystal Engineering

Chapter 8

Screening of Cryoprotectants and the Multistep Soaking Method

Miki Senda and Toshiya Senda

Abstract

Crystals obtained from an initial crystallization screening are not always of sufficient quality for structural determination at atomic resolution. For this reason, post-crystallization treatments such as cryoprotection and dehydration have frequently been utilized to improve the crystal quality. In addition, several recent studies have shown that cryoprotectants can interact with the proteins in the obtained crystal and further stabilize them, leading to further improvement of the crystal quality. In this chapter, we propose a multistep soaking method in which crystals are sequentially soaked in two to three cryoprotectant solutions. This method was found to be effective for improving the crystal quality. However, since the screening of cryoprotectants for use in this method involves much trial and error, it is important to record each step of the screening in a systematic manner.

Keywords Cryoprotectants, Post-crystallization treatment, Artificial mother liquor

1 Introduction

1.1 Post-crystallization Treatment

While X-ray crystallography is a powerful tool to determine the tertiary structure of biological macromolecules at atomic resolution, it generally requires high-quality crystals that diffract to better than 3 Å resolution. Therefore, numerous attempts have been made to establish a method for obtaining crystals of sufficient quality. In particular, the development of crystallization screening kits and crystallization robots has dramatically increased the success rate and efficiency of the initial crystallization screening of biological macromolecules [1–5]. Nonetheless, the obtained crystals are not always of sufficient quality for crystal structure determination at atomic resolution. Thus a reliable method for improving the quality of the crystals is needed.

One of the frequently utilized methods to improve the crystal quality is changing the target protein to its homologue. Crystallization of a homologue of the target protein sometimes yields a high-quality crystal [6–10]. Deletion of intrinsically disordered

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks,
DOI 10.1007/978-4-431-56030-2_8, © Springer Japan 2016

region(s) from the target protein is also effective for improving the crystal quality [11]. However, in some cases, we cannot adopt these strategies. When our target protein is not a recombinant protein, for example, preparation of a deletion mutant is impossible. If homologues of the target protein cannot yield crystals better than the original one, we need to use the original crystal. In these cases, another strategy for crystal quality improvement is required. Post-crystallization treatment [12, 13] can be applied even in these difficult cases. Post-crystallization treatment improves the crystal quality by subjecting the crystal to various combinations and levels of soaking, freezing, and humidity [14–21].

1.2 Cryo-crystallography

In recent years, most diffraction data have been collected at cryogenic temperature to avoid radiation damage by the synchrotron X-ray radiation [22, 23]. Crystals are flash-cooled before data collection and kept frozen in a cold N_2 flow, typically at 100 K, during the diffraction data collection. The techniques of diffraction data collection at cryogenic temperature were developed in the early 1990s, when the use of synchrotron radiation became common in the field of protein crystallography [23, 24]. Protein crystals typically consist of approximately 50 % water molecules by volume [25], and the water molecules surrounding and/or inside the protein crystal form crystalline ice upon freezing, damaging the protein crystal. Furthermore, the crystalline ice causes strong circular diffractions known as ice rings at around 3.7 Å resolution, which hampers the collection of high-quality data from the frozen crystal (Fig. 1a) [23]. For this reason, the water molecules should be frozen in an amorphous state to avoid crystalline ice formation (Fig. 1b). In the

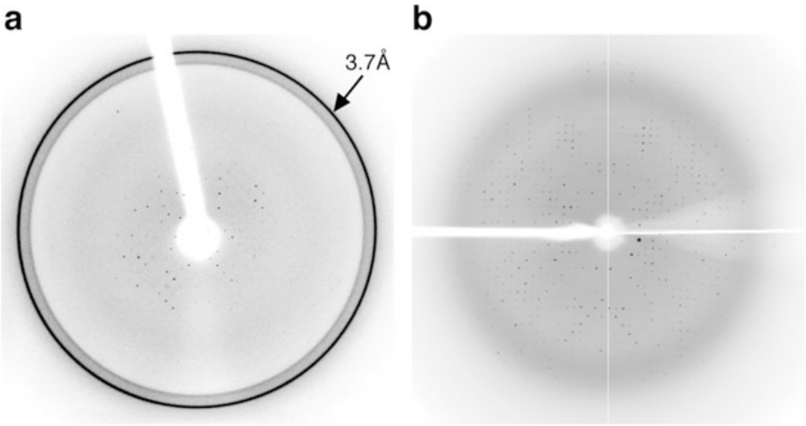


Fig. 1 A diffraction pattern with ice rings. (**a**) A TAF-IβΔC crystal was flash-cooled without cryoprotectant solution. Ice rings were observed at around 3.7 Å resolution. (**b**) A TAF-IβΔC mutant (Leu104Met/Leu145Met/Leu166Met) crystal was flash-cooled using artificial mother liquor containing 30 %(w/v) trehalose. No ice rings were observed because the level of cryoprotection was appropriate

1960s, it was discovered that small compounds such as sucrose could be utilized to freeze water molecules in an amorphous state [26]. In 1975, the replacement of crystallization solution with an organic compound was reported to be an effective method for crystal freezing [27]. The method of crystal freezing was rapidly improved in the 1990s, when protein crystallographers began to utilize many organic compounds and sugars as cryoprotectants [23, 24]. Today, kits for screening the cryoprotectants to be used in crystal freezing are commercially available.

1.3 Cryo-conditions

While cryoprotectants have been widely utilized in crystal freezing, most crystallographic investigations have been carried out without intensive screening of the cryoprotectants. Since inappropriate selection of a cryoprotectant leads to poor diffractions, optimization of the cryo-conditions is critical to obtaining high-quality diffraction data. Interestingly, several analyses have revealed that cryoprotectants occasionally interact with protein molecules in the crystal, resulting in the improvement of the crystal quality by stabilizing the target protein [28–33]. Therefore, cryoprotectants can also be used to improve the crystal quality by stabilizing proteins in the crystal. In this chapter, we describe methods to improve the crystal quality using cryoprotectants. We begin by explaining the basic soaking technique. We then describe the methods used to screen for optimal cryo-conditions and the protocol of the multistep soaking.

1.4 Basics of Soaking Experiments

Since the cryoprotection of the protein crystal is achieved by soaking, it is essential to prepare artificial mother liquor (or standard buffer) before the soaking experiment. Otherwise, the crystals will be damaged during soaking and the results will be poor. It has been established that artificial mother liquor can maintain the crystal of a target protein for at least 2–3 days without damaging it [34]. An artificial mother liquor can be prepared based on the conditions of the crystallization solution (reservoir solution). There are two critical conditions for the artificial mother liquor. First, the pH of the artificial mother liquor should be adjusted to that of the droplet solution that produces the crystal. The difference of pH should be less than 0.1 in order to avoid pH shock when transferring a crystal from a crystallization droplet to the artificial mother liquor. Second, the concentration of the precipitant(s) should be optimized; the concentration needs to be increased by 10–20 % from that of the crystallization solution to avoid dissolution of crystals. Once the conditions of the artificial mother liquor are fixed, the artificial mother liquor can be utilized for a variety of soaking experiments. The artificial mother liquor, of course, can be utilized for typical soaking experiments to prepare crystals of the protein-small compound complexes.

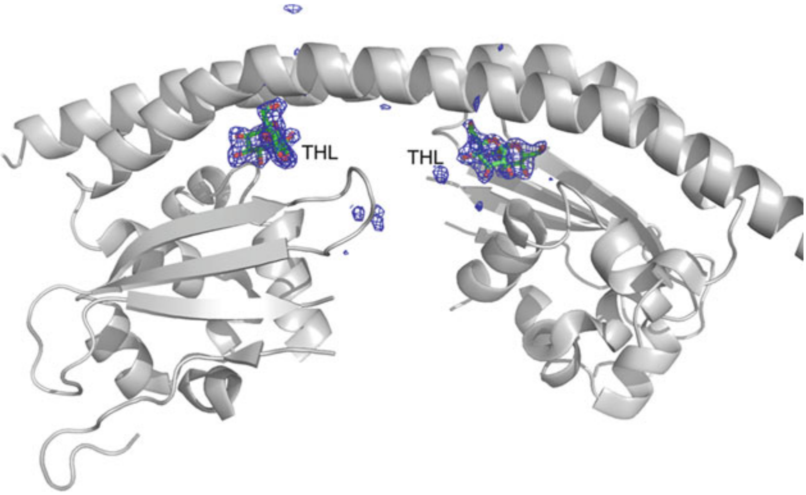


Fig. 2 Trehalose molecules bound to TAF-Iβ. Trehalose (*THL*) molecules can be seen located between the two domains. The mFo-DFc densities are contoured at about 3.0 σ

After fixing the conditions of the artificial mother liquor, we can start the screening of cryoprotectants. It is convenient to use commercially available screening kits for the initial screening. This first screening of cryoprotectants should be done with 20–30 regents. The effects of the cryoprotectant will differ under different soaking times and/or soaking temperatures, and these parameters are generally optimized after the initial screening.

As described above, some cryoprotectants interact with the target protein in the crystal. This type of interaction has been suggested to stabilize the target protein in the crystal and improve the crystal quality [28–33]. In the case of TAF-Iβ, which is a histone chaperone that interacts preferentially with the histone H3-H4 complex, molecules of trehalose, which was used as a cryoprotectant, were found in the crystal structure [31, 32]. These trehalose molecules were located between two domains and interacted with these domains by forming intensive hydrogen bonds. These interactions seemed to stabilize the protein structure and improve the crystal quality (Fig. 2). Since most of cryoprotectants have several polar groups, such as hydroxyl and carbonyl groups, cryoprotectants are likely to interact with protein molecules via hydrogen bonds. It is therefore reasonable to utilize more than one cryoprotectant to stabilize the protein structure. Indeed, a combination of several cryoprotectants has been proven effective for crystal quality improvement [35].

1.5 Crystal Annealing

One of the common problems in crystal freezing is the increase of crystal mosaicity [24]. This seems to happen even when the water molecules are frozen in an amorphous state. The high mosaicity can sometimes be improved by a crystal annealing procedure, in which a frozen crystal is kept at room temperature to thaw it and then refrozen by a cold N_2 flow after several seconds. Since this annealing procedure sometimes dramatically improves the crystal mosaicity [16, 17], crystal annealing has frequently been applied in conjunction with cryoprotection.

1.6 Evaluation of Diffraction Images

In the screening of the cryoprotectants, particularly for the multistep soaking method, we need to examine numerous diffraction images to select suitable soaking conditions. It is critical that the diffraction images should be evaluated using the same criteria. In this manuscript, we utilize two measures for crystal resolution, the *maximum resolution* and *resolution limit*. The maximum resolution is defined on the basis of the statistics of diffraction data processing/scaling. We define the maximum resolution as the resolution that satisfies *R*merge < 0.5 and I/σ(I) > 3 at the outermost resolution shell. The second measure is the resolution limit, which is determined by the visual inspection of diffraction images. Due to the lack of numerical criteria on the maximum resolution, the resolution limit may show some degree of deviation. Our analysis, however, revealed that these two values show a correlation if the resolution limit is determined carefully (Fig. 3). In the screening process, we need to evaluate the quality of each crystal with a few snapshot images. Since we use only a limited number of crystals for the full data collection, it is not practical to compare the crystal quality using the maximum resolution. We therefore need to utilize the resolution limit of the crystal to compare the crystal quality. It is of course possible to use other criteria; some programs have been developed to estimate the resolution of the crystal from a diffraction image. The most important point is, however, that the criteria should be the same throughout the screening process.

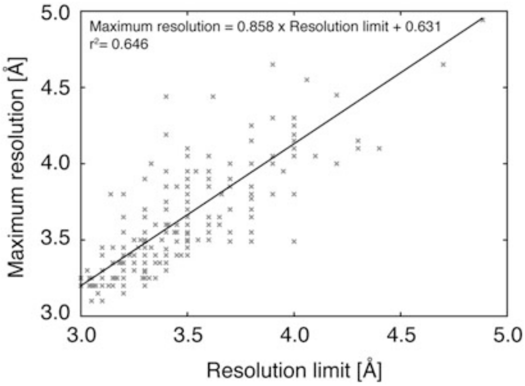


Fig. 3 Correlation between the maximum resolution and the resolution limit. The resolution limit showed reasonable correlation with the maximum resolution. The resolution limit, which is determined by visual inspection of a few diffraction images, could be used in the cryoprotectant screening

2 Materials

2.1 Chemicals

Sugars, alcohols, PEGs, and organic compounds (glycerol; ethylene glycol; PEG200; PEG400; PEG600; PEG4000; polyvinylpyrrolidone K 15, (+/−)-2-methyl-2,4-pentanediol (MPD); 1,6-hexanediol; 1,2-propanediol; dimethyl sulfoxide (DMSO); 2-propanol; ethanol; methanol; D-(+)-sucrose; meso-erythritol; xylitol; D-(+)-raffinose; D-(+)-trehalose dihydrate (THL); D-(+)-glucose, etc.) have frequently been utilized as cryoprotectants. Screening kits of cryoprotectants are commercially available from several companies. Special care must be taken when using PEG as a cryoprotectant. While PEG is available from various suppliers, the quality of PEG will differ widely among these sources. When obtaining PEG, therefore, check the pH of the PEG solution; again, the pH values of the PEG solutions will differ among vendors.

2.2 Glassware

For the soaking experiments, a depression glass is useful (Fig. 4). The glass should be siliconized before use. A cryoloop is required to transfer crystals between the solutions. A stereomicroscope is utilized for observation of the soaked crystals. It is convenient to use the stereomicroscope in a cold room, because soaking at 4 °C is frequently effective.

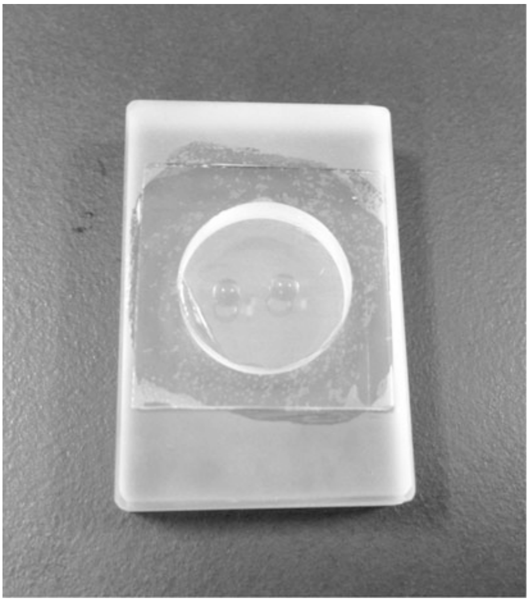


Fig. 4 Depression glass. Two droplets are placed at the bottom of the well. The well should be sealed with a cover glass using grease

3 Methods

3.1 Preparation of Artificial Mother Liquor

1. Measure the pH of the crystallization solution that harbors crystals of your target protein.
2. Prepare some candidates of the artificial mother liquor on the basis of the crystallization (reservoir) solution (Table 1). Their pH should be adjusted to that of the crystallization

Table 1
Examples of artificial mother liquor and cryoprotectant solution [31, 32, 35–39]

Protein name	Crystallization conditions	Artificial mother liquor	Cryoprotectant solution
BphA4	1.7–2.1 M sodium formate	2.5 M sodium formate	27.5 %(v/v) glycerol
	100 mM sodium acetate pH 5.3–5.4	0.1 M sodium acetate pH 5.4	2.5 M sodium formate
			0.1 M sodium acetate pH 5.4
CagA(1–876)	7–10 %(v/v) ethanol	20 %(v/v) ethanol	28 %(w/v) trehalose
	50 mM Tris-HCl pH 7.0–7.1	50 mM Tris-HCl pH 7.0	20 %(v/v) ethanol
			50 mM Tris-HCl pH 7.0
DDO (D-aspartate oxidase)	5–10 %(w/v) PEG8000	20 %(w/v) PEG8000	20 %(v/v) glycerol
	100 mM potassium dihydrogen phosphate	100 mM potassium dihydrogen phosphate	20 %(w/v) PEG8000
	100 mM sodium acetate pH 4.7	100 mM sodium acetate pH 4.7	100 mM potassium dihydrogen phosphate
			100 mM sodium acetate pH 4.7
DSD (D-serine dehydratase)	12–15 %(w/v) PEG4000	30 %(w/v) PEG4000	30 %(v/v) glycerol
	10 %(v/v) 2-propanol	10 %(v/v) 2-propanol	20 %(w/v) PEG4000
	100 mM MES pH 6.5	100 mM MES pH 6.5	10 %(v/v) 2-propanol
			100 mM MES pH 6.5
TAF-Iβ	2.4–2.85 M ammonium sulfate	2.75 M ammonium sulfate	30 %(w/v) trehalose
	200 mM potassium sodium tartrate	200 mM K/Na tartrate	2.75 M ammonium sulfate
	30 mM magnesium chloride	30 mM magnesium chloride	200 mM K/Na tartrate
	100 mM sodium citrate pH 5.4–5.5	100 mM sodium citrate pH 5.4	30 mM magnesium chloride
			100 mM sodium citrate pH 5.4

solution (measured in step 1). The concentration of precipitant (s) should be increased by 10–20 %.

3. Soak crystals in each of the prepared solutions. Crystals can be transferred from a crystallization droplet to the prepared solution by a cryoloop. Do not damage the crystal when transferring it. In particular, avoid touching the crystal with the cryoloop. A depression glass (Fig. 4) is useful for the soaking experiment. After the crystal transfer, the well of the depression glass should be sealed to avoid evaporation of the artificial mother liquor. A typical volume of the artificial mother liquor for the soaking experiment is 5–20 μL.
4. Observe the soaked crystals with a stereomicroscope. Check for cracks on the crystal surface just after soaking. If the crystals crack immediately after soaking, the concentration of the precipitant should be changed. Also, it is useful to check the pH of the solution. Finally, sometimes low-temperature soaking (e.g., soaking at 4 °C) may prevent the crystal damage.
5. After the observation, the depression glass should be stored in an incubator. The temperature will usually be the same as that of the crystallization.
6. Crystals should be observed each day, with careful monitoring for cracks on the surfaces and change in the crystal size due to dissolution. The goal is to identify conditions that do not crack and dissolve the soaked crystals. The best means of accomplishing this is to take a photo of the soaked crystals each day in order to monitor their status.
7. If possible, it is better to check diffractions from a crystal soaked in the artificial mother liquor.
8. When you cannot stop the dissolution of the crystal in the artificial mother liquor, try to add your target protein in the artificial mother liquor; the concentration of the protein is usually less than that of the crystallization conditions.
9. Table 1 shows examples of the artificial mother liquor. Compare the conditions of the crystallization solution and artificial mother liquor.

3.2 Screening of Cryoprotectants

1. Prepare a cryoprotectant solution on the basis of the conditions of the artificial mother liquor. The concentration of a protectant is approximately 15–30 % (w/v, v/v). It is convenient to prepare a 2× artificial mother liquor and 30–60 % solution of a

cryoprotectant and mix them in a 1:1 ratio. After mixing the solutions, the pH of the mixture should be adjusted to that of the artificial mother liquor.

2. Several crystals should be soaked in one cryoprotectant solution to check for reproducibility. It is highly recommended that the size of each crystal be recorded. Just after soaking, check the appearance of the crystal. Here again, for this purpose, it is best to take a photo of the soaked crystals. In some cases, the crystals will be damaged immediately after soaking in the cryoprotectant solution.
3. A soaking time of 30 s to 3 min is recommended for the initial screening.
4. After the soaking, the crystal is mounted on a cryoloop and frozen. Please note that freezing in a cold N_2 flow (*ca.* 100 K) and freezing with liquid nitrogen may result in different quality of crystal diffractions. Try both methods of freezing.
5. Take diffraction patterns (snapshots) from several directions ($\varphi = 0°, 45°, 90°$, etc.) and analyze the crystal quality. Check the resolution (resolution limit), mosaicity, shape of diffraction spots, and anisotropy carefully. Handling of the crystal frequently damages the crystal quality. To avoid this type of artifact, it is important to check the crystal quality with two to three crystals. While it is possible to statistically judge the significance of the difference in resolution, it is better to detect obvious differences in the crystal quality at the initial stage of the screening.
6. Try 20–30 cryoprotectants and select cryoprotectant solutions that give high-resolution diffractions.
7. Optimize the soaking time and temperature of the selected cryoprotectant solution(s).
8. When the crystals are damaged by soaking, try soaking at a low temperature (e.g., 4 °C). Sometimes low-temperature soaking dramatically improves the situation.

3.3 Multistep Soaking Method

1. Prepare the artificial mother liquor as described in Sect. 3.1.
2. Perform cryoprotectant screening as described in Sect. 3.2. Make a list of cryoprotectants that improve the crystal quality. Even if the effect of a cryoprotectant is marginal, it should be included in the list of the cryoprotectants that will be used in the multistep soaking. Even if the effect of a given

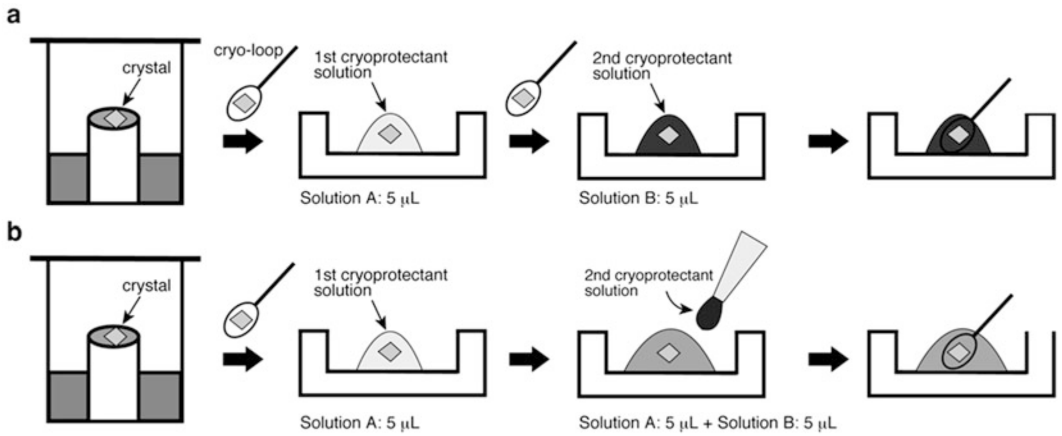

a
crystal
cryo-loop
1st cryoprotectant solution
Solution A: 5 µL
2nd cryoprotectant solution
Solution B: 5 µL
b
crystal
1st cryoprotectant solution
Solution A: 5 µL
2nd cryoprotectant solution
Solution A: 5 µL + Solution B: 5 µL

Fig. 5 Two methods for multistep soaking. (**a**) In the first method, a crystal is transferred from a crystallization droplet to the first cryoprotectant solution and then transferred to the second cryoprotectant solution. (**b**) In the second method, the crystal is soaked in the first cryoprotectant solution, and then a second cryoprotectant solution is added to the first

cryoprotectant is small, in combination with other cryoprotectants, it may dramatically improve the crystal quality.

3. Try the combinations of cryoprotectants listed above. When combining cryoprotectants, the order of soaking affects the results. There are two methods for the multistep soaking. In the first method, a crystal soaked in the first cryoprotectant solution is transferred into another cryoprotectant solution using a cryoloop (Fig. 5a). In the second method, the second cryoprotectant solution is added to the first cryoprotectant solution containing a crystal (Fig. 5b). A combination of two cryoprotectants is most typically used for the multistep soaking.
4. Examine the quality of the diffraction images. As described above, take the diffraction patterns from several different directions ($\varphi = 0°, 45°, 90°$, etc.) and analyze the crystal quality. Handling of the crystal frequently damages the crystal quality. To avoid this type of artifact, it is important to check the crystal quality with two to three crystals.
5. When you find a good combination of cryoprotectants, optimize the soaking time and temperature of each soaking.

When using the multistep soaking method, it is highly recommended that a spreadsheet program such as MS Excel be used to prepare a table (Table 2). This type of table is useful to systematically compare the quality of the crystals treated under various conditions.

Table 2
Data sheet for screening of cryoprotectant solution

				First step soaking		Second step soaking		Resolution	
ID	Crystal ID	Plate ID	Size [mm]	Cryoprotectant 1	Time	Cryoprotectant 2	Time	Max [Å]	Limit [Å]
###	Protein-yymmdd-###	###	##	##% ### ##% ## ## mM ## pH ##	## min	##% ### ##% ## ## mM ## pH ##	## min	###	###
001	Caga-091218-006	001	0.2	30 % trehalose 20 % ethanol 50 mM Tris-HCl pH 8.8	3 min	–	–	4.7	4.65
002	Caga-100528-001	058	0.5	17.5 % erythritol 40 % ethanol 50 mM Tris-HCl pH 9.0	0.3 min	–	–	3.4	3.70
003	Caga-100627-015	117	0.4	30 % trehalose 20 % ethanol 50 mM Tris-HCl pH 8.8	14 h	17.5 % erythritol 40 % ethanol 50 mM Tris-HCl pH 9.0	2.3 min	3.3	3.49
004	Caga-120128-034	471	0.3	30 % trehalose 20 % ethanol 50 mM Tris-HCl pH 7.0	30 min	30 % PEG1000 20 % ethanol 50 mM Tris-HCl pH 7.0	1.0 min	3.4	3.30

Acknowledgments

This study was supported by the New Energy and Industrial Technology Development Organization (NEDO), CREST (JST), Grants-in-Aid for Scientific Research on an Innovative Area (MEXT), and the Platform for Drug Discovery, Informatics, and Structural Life Science (MEXT) of Japan.

References

1. Jancarik J, Kim S (1991) Sparse matrix sampling: a screening method for crystallization of proteins. J Appl Crystallogr 24:409–411
2. Morris DW, Kim CY, McPherson A (1989) Automation of protein crystallization trials: use of a robot to deliver reagents to a novel multi-chamber vapor diffusion method. Biotechniques 7:522–527
3. Luft JR, Collins RJ, Fehrman NA, Lauricella AM, Veatch CK, DeTitta GT (2003) A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J Struct Biol 142:170–179
4. Hosfield D, Palan J, Hilgers M et al (2003) A fully integrated protein crystallization platform for small-molecule drug discovery. J Struct Biol 142:207–217
5. Hiraki M, Kato R, Nagai M et al (2006) Development of an automated large-scale protein-crystallization and monitoring system for high-throughput protein-structure analyses. Acta Crystallogr D 62:1058–1065
6. Kendrew JC, Parrish RG, Marrack JR, Orlans ES (1954) The species specificity of myoglobin. Nature 174:946–949
7. Lawson DM, Artymiuk PJ, Yewdall SJ et al (1991) Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts. Nature 349:541–544
8. Longenecker KL, Garrard SM, Sheffield PJ et al (2001) Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoGDI. Acta Crystallogr D 57:679–688
9. Schlatter D, Thoma R, Küng E et al (2005) Crystal engineering yields crystals of cyclophilin D diffracting to 1.7 Å resolution. Acta Crystallogr D 61:513–519
10. Derewenda ZS, Vekilov PG (2006) Entropy and surface engineering in protein crystallization. Acta Crystallogr D 62:116–124
11. Chait BT (1994) Mass spectrometry–a useful tool for the protein X-ray crystallographer and NMR spectroscopist. Structure 2:465–468
12. Heras B, Martin JL (2005) Post-crystallization treatments for improving diffraction quality of protein crystals. Acta Crystallogr D 61:1173–1180
13. Newman J (2006) A review of techniques for maximizing diffraction from a protein crystal in stilla. Acta Crystallogr D 62:27–31
14. Tong L, Qian C, Davidson W et al (1997) Experiences from the structure determination of human cytomegalovirus protease. Acta Crystallogr D 53:682–690
15. Fu Z, DuBois GC, Song SP et al (1999) Improving the diffraction quality of MTCP-1 crystals by post-crystallization soaking. Acta Crystallogr D 55:5–7
16. Harp JM, Timm DE, Bunick GJ (1998) Macromolecular crystal annealing: overcoming increased mosaicity associated with cryocrystallography. Acta Crystallogr D 54:622–628
17. Yeh JI, Hol WGJ (1998) A flash-annealing technique to improve diffraction limits and lower mosaicity in crystals of glycerol kinase. Acta Crystallogr D 54:479–480
18. Samygina VR, Antonyuk SV, Lamzin VS, Popov AN (2000) Improving the X-ray resolution by reversible flash-cooling combined with concentration screening, as exemplified with PPase. Acta Crystallogr D 56:595–603
19. Petock JM, Wang Y, DuBois GC, Harrison RW, Weber IT (2001) Effect of different post-crystallization soaking conditions on the diffraction of Mtcp1 crystals. Acta Crystallogr D 57:763–765
20. Green TJ, Luo M (2006) Resolution improvement of X-ray diffraction data of crystals of a vesicular stomatitis virus nucleocapsid protein oligomer complexed with RNA. Acta Crystallogr D 62:498–504
21. Abad-Zapatero C, Oliete R, Rodriguez-Puente S et al (2011) Humidity control can compensate for the damage induced in protein crystals by alien solutions. Acta Crystallogr F 67:1300–1308
22. Watenpaugh KD (1991) Macromolecular crystallography at cryogenic temperatures. Curr Opin Struct Biol 13:434–551
23. Garman EF, Schneider TR (1997) Macromolecular cryocrystallography. J Appl Crystallogr 30:211–237
24. Rodgers DW (1994) Cryocrystallography. Structure 2:1135–1140
25. Matthews BW (1968) Solvent content of protein crystals. J Mol Biol 33:491–497
26. Haas DJ, Rossmann MG (1970) Crystallographic studies on lactate dehydrogenase at -75 °C. Acta Crystallogr B 26:998–1004
27. Petsko GA (1975) Protein crystallography at sub-zero temperatures: cryo-protective mother liquors for protein crystals. J Mol Biol 96:381–392
28. Sousa R (1996) Use of glycerol, polyols and other protein structure stabilizing agents in protein crystallization. Acta Crystallogr D 51:271–277

29. Wimberly BT, Brodersen DE, Clemons WM Jr et al (2000) Structure of the 30S ribosomal subunit. Nature 407:327–339
30. Clemons WM Jr, Brodersen DE, McCutcheon JP et al (2001) Crystal structure of the 30 S ribosomal subunit from *Thermus thermophilus*: purification, crystallization and structure determination. J Mol Biol 310:827–843
31. Muto S, Senda M, Akai Y et al (2007) Relationship between the structure of SET/TAF-Iβ/INHAT and its histone chaperone activity. Proc Natl Acad Sci U S A 104:4285–4290
32. Senda M, Muto S, Horikoshi M, Senda T (2008) Effect of leucine-to-methionine substitutions on the diffraction quality of histone chaperone SET/TAF-Iβ/INHAT crystals. Acta Crystallogr F 64:960–965
33. Bertheleme N, Chae PS, Singh S, Mossakowska D, Hann MM, Smith KJ, Hubbard JA, Dowell SJ (2013) Unlocking the secrets of the gatekeeper: methods for stabilizing and crystallizing GPCRs. Biochim Biophys Acta 1828:2583–2591
34. Blundell TL, Johnson LN (1976) Protein crystallography. Academic, New York
35. Hayashi T, Senda M, Morohashi H et al (2012) Tertiary structure and functional analysis of the *Helicobacter pylori* CagA oncoprotein. Cell Host Microbe 12:20–33
36. Senda M, Kishigami S, Kimura S et al (2007) Molecular mechanism of the redox-dependent interaction between NADH-dependent ferredoxin reductase and Rieske-type [2Fe-2S] ferredoxin. J Mol Biol 373:382–400
37. Senda M, Yamamoto A, Tanaka H et al (2012) Crystallization and preliminary crystallographic analysis of D-aspartate oxidase from porcine kidney. Acta Crystallogr F 68:644–646
38. Senda M, Tanaka H, Ishida T, Horiike K, Senda T (2011) Crystallization and preliminary crystallographic analysis of D-serine dehydratase from chicken kidney. Acta Crystallogr F 67:147–149
39. Tanaka H, Senda M, Venugopalan N et al (2011) Crystal structure of a zinc-dependent D-serine dehydratase from chicken kidney. J Biol Chem 286:27548–27558

Chapter 9

Protein Modification for Crystallization

Toshio Hakoshima

Abstract

Technological advances in data collection with synchrotron radiation sources and phasing methods including automated model building and validation have highlighted crystallization as the rate-limiting step in X-ray diffraction studies of macromolecular structures. Although protein crystallization remains a stochastic event, protein engineering with the advent of recombinant methods enables us to generate target proteins possessing a higher propensity to form crystals suitable for X-ray diffraction data collection. This chapter presents an overview of protein engineering methods designed to enhance crystallizability and discusses examples of their successful application.

Keywords Protein engineering, Recombinant DNA technology, Flexible loops, Non-conserved insertion, Secondary structure prediction, Natural variation, Artificial linker

1 Introduction

Advanced recombinant technology and biochemical installations significantly reduce efforts required for protein production and purification. Moreover, the use of superb crystallization screening kits coupled with high-performance crystallization robots has changed previously laborious trial-and-error crystallization experiments to routine laboratory work that can be executed by specialist and nonspecialist researchers alike. However, the preparation of single well-diffracting crystals of the target proteins remains a time-consuming challenge. Once single well-diffracting crystals have been obtained, however, X-ray data collection using synchrotron radiation and phasing of the intensity data followed by structural refinement are relatively straightforward tasks.

Two approaches have been employed to improve protein crystal quality and size. Firstly, natural variation in amino acid sequences of homologues or homologues from different species can be exploited to identify a target with suitable crystallization properties. Alternatively, artificial modification of target proteins by the use of recombinant techniques can be employed to enhance the

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks,
DOI 10.1007/978-4-431-56030-2_9, © Springer Japan 2016

target protein's propensity to crystallize or to improve the diffraction quality of the resulting crystals. In this review, we firstly introduce some examples of the use of homologue proteins to demonstrate the impact of natural sequence variations on crystallizability and crystal lattices and then discuss current progress in protein engineering methodologies used to improve the crystal quality of target proteins that are recalcitrant to crystallization in their wild-type form. Protein engineering methodologies can execute internal deletion of non-conserved flexible loops in addition to frequently used N- and C-terminal truncations, in addition to the use of fusion proteins between tags and target proteins and between ligands and target proteins. Although these approaches require preliminary optimization screening, the screening procedures are fairly well established and therefore can be routinely performed to obtain diffraction-quality crystals.

2 Methods

2.1 Homologous Proteins

Historically, natural variations in the amino acid sequences of homologues were exploited to identify targets with suitable crystallization properties during the purification procedure [1, 2]. Extensive application of this approach resulted in a scramble to report on the first structure determination of transcription factors in 1990s. How much variation in the sequences is needed to enhance the propensity to crystallize or to improve the diffraction quality? Human and mouse genomes share well-conserved sequences of their homologues with high amino acid identity (>90 %), and their conserved functionally important domains display high identity (>95 %). These high sequence identities significantly decrease the possibility of significant improvements in crystallization or crystal quality. In practice, homologues with less than 80 % identity are potential targets for improvement trials. Recent examples are mammalian T-lymphoma invasion and metastases 1 and 2 (Tiam1 and Tiam2), which are Rac-specific guanine exchange factors (Rac-GEFs) [3, 4], and dwarf 14 (D14) and related proteins, which are plant hormone receptor candidates [5].

Tiam1 possesses a novel functional PHCCEx domain (~30 kDa) for plasma membrane association and specific binding to a class of membrane proteins. The domain boundary of the mouse Tiam1 PHCCEx domain was delineated following extensive screening of expression constructs, since several constructs produced proteins that were easily degraded during the protein purification steps. The optimized construct produced a stable protein sample that was successfully crystallized in the form of needlelike crystals of a hexagonal lattice ($P6_422$, $a = b = 113.5$ Å, $c = 113.8$ Å, $\gamma = 120°$), although the crystals diffracted poorly up to 4.5 Å using synchrotron radiation at SPring-8. Several trials to improve the diffraction by changing conditions or using additives were unsuccessful. Human and mouse Tiam1 PHCCEx domains share high sequence identity (>90 %). Thus, focus was then set on Tiam2, which is a functional homologue of Tiam1 with 65 % sequence identity. Our sequence alignment showed that the Tiam2 PHCCEx domain possesses no large insertion or deletion compared with Tiam1, suggesting that the sequence variation is relatively high but suitable for possible improvement. With this sequence variation, protein samples of the Tiam2 PHCCEx domain were purified in a similar manner to that of Tiam1. Crystallization screening of the Tiam2 PHCCEx domain yielded two crystal forms, chunky crystals of tetragonal ($P4_32_12$, $a = b = 105.6$ Å, $c = 287.6$ Å) and monoclinic ($P2_1$, $a = 46.7$ Å, $b = 104.8$ Å, $c = 116.0$ Å, $\beta = 80.6°$) lattices. The tetragonal crystals diffracted at 3.2 Å and the monoclinic up to 2.08 Å, which are sufficient for structure determination and detailed characterization of the molecular structure.

D14 and related D14-like (D14L) proteins belong to an α/β hydrolase family based on amino acid sequences and are candidates for strigolactone and karrikin receptors, respectively. *Arabidopsis thaliana* (*At*) and *Oryza sativa* (*Os*, rice) D14 share 74 % amino acid identity. The recombinant protein of *At*D14 was easily prepared as a soluble protein and concentrated to 20 mg/mL to yield crystals following conventional crystallization screening. However, the diffraction limit of these crystals was around 4 Å, the mosaicity was large, and the diffraction spots appeared as streaks. Compared with *At*D14, *Os*D14 possesses an additional non-conserved sequence of 54 residues at the N-terminus. This N-terminal non-conserved extension contains many Gly and Ser residues and was predicted as an intrinsically disordered random coil. In general, *Os* proteins often possess such additional sequences predicted to form random coils. N-terminal truncated *Os*D14 (Δ54) could be prepared as a soluble protein, although its solubility was poor and the maximum concentration was 3 mg/mL. Despite the limited suitability for structural work, the orthorhombic crystals ($P2_12_12_1$, $a = 48.0$ Å, $b = 88.2$ Å, $c = 121.2$ Å) of *Os*D14 (Δ54) diffracted at 1.45 Å. D14L is also referred to as KARRIKIN INSENSITIVE 2 (KAI2) and shares about 50 % amino acid identity with D14. The recombinant protein of *At*D14L was efficiently expressed, easily purified, concentrated to 20 mg/mL, and crystallized in monoclinic crystals ($P2_1$, $a = 51.0$ Å, $b = 55.6$ Å, $c = 53.1$ Å, $\beta = 115.8°$) that diffracted up to 1.15 Å.

2.2 Internally Truncated Proteins

As already mentioned in the case of *Os*D14, N- and/or C-terminal truncation(s) can be frequently implemented in an effort to improve protein properties such as stability, solubility, and crystallizability. This approach could be extended to internal loop regions that may prevent crystallization of the target proteins. One recent example is the C-terminal cargo-recognition domain comprising myosin tail homology 4 (MyTH4) and 4.1/ezrin/radixin/moesin (FERM) subdomains, the so-called MyTH4–FERM cassette, found in nonconventional myosins [6]. A DNA fragment encoding the MyTH4–FERM cassette (residues 1486–2058) of human myosin-X cloned into the pET47b [+] vector (Novagen) produced a soluble protein, although this protein was unstable and suffered partial degradation during purification. No crystals were obtained from the purified sample. A protease labile region was found in the FERM domain. Compared with the canonical FERM domain from ezrin/radixin/moesin (ERM) proteins, the FERM domain of the myosin-X MyTH4–FERM cassette contains a non-conserved insertion of ca. 60 residues (1850–1910) located between α2B and α3B helices (Fig. 1). Using time-of-flight mass spectroscopy (TOF-MS), we identified the cleavage site at S1892-F1893, which was located within the non-conserved insertion. We designed S1892A and F1893A mutants to prevent this partial degradation. However, these mutant proteins were still degraded during purification. Next, we designed truncated proteins comprising deletion of residues forming the internal non-conserved insertion. Nucleotides encoding residues 1845–1891 (Δ47), 1872–1891 (Δ20), or 1882–1891 (Δ10) of the non-conserved insertion were deleted from the plasmid using inverse PCR. To test the cargo-binding affinity of these internally truncated MyTH4–FERM cassettes, we performed pull-down assays with a GST-fused netrin receptor, deleted in colorectal cancer (DCC), which is a myosin-X cargo protein. We found that Δ47 possessed reduced affinity, while Δ20 and Δ10 showed retained affinity. The MyTH4–FERM cassettes (Δ20) were successfully purified without degradation, crystallized in a monoclinic lattice ($P2_1$, $a = 185.2$ Å, $b = 49.6$ Å, $c = 94.0$ Å, $\beta = 116.7°$), and diffracted at 1.9 Å resolution. The complex between the MyTH4–FERM cassettes (Δ20) and DCC was also crystallized in a related lattice ($P2_1$, $a = 85.4$ Å, $b = 49.5$ Å, $c = 93.4$ Å, $\beta = 117.1°$) and diffracted at 2.2 Å resolution.

Interestingly, an independent structural work of a fusion protein between the myosin-X MyTH4–FERM cassettes and DCC showed that, after extensive trials, deletion of a 36-residue fragment (residues 1871–1906) within the non-conserved insertion was necessary for crystallization [7]. Moreover, another structural work of the MyTH4–FERM cassettes of myosin VIIA, which is a close homologue of myosin-X, showed that a 30-residue deletion (residues 1037–1066) in the MyTH4 domain but not in the FERM domain was necessary for crystallization of the cassette bound to Sans [8]. This 30-residue deletion is part of the non-conserved long insertion (residues 1030–1080) between helices α1M and α2M, compared with the myosin-X MyTH4 domain.

Internal deletion was also explored in recent structural work of the yeast Ire kinase-nuclease domain [9]. Ire1 is an ancient

```
mRadixin      MPKPINVRVTTMDAELEFAIQ-PNTTGKQLFDQVVKTVGLREVW---FFGLQYVDSKGYSTW-- 52
sfMoesin      MPKSMNVRVTTMDAELEFAIQ-QTTTGKQLFDQVVKTIGLREVW---FFGLQYTDSKGDLTW--
hsMyoX        ---MTSTVYCHGGGSCKITIN-SHTTAGEVVEKLIRGLAMEDSR--NMFALFEY--NGHVDKAIE 1756
XeMyoX        -SHMTTSVYCHGGGSCQISIN-SHTTAGEVVEKLIRGLSMDNSR--NMFALFEH--NKHTDRAVE
DmMyoX        -RSARRQIYRLPGGAERVVNTRCSTVVADVIAELCALLGVESEAEQQEFSLYCIVQGDAFTMPLA

mRadixin      LKLNKKVTQQDVKKEN----------PLQFKFRAKFFP-EDVSEELIQEITQRLFFLQVKEAI 110
sfMoesin      IKLYKKVMQQDVKKEN----------PLQFKFRAKFYP-EDVADELIQEITLKLFYLQVKNAI
hsMyoX        SRTVVADVLAKFEKLAATSEVGDL--PWKFYFKLYCF--LDTDNVPKDSVEFAFMFEQAHEAV 1815
XeMyoX        SRVIVADVLAKFERLAGTGDEEDDLGPWNLYFKLYCF--LDVQSVPKEGIEFAFMFEQAHESL
DmMyoX        ADEYILDVTTELLKSGQ---------PFYLIFCRSVWHFALKREPAPMPLYVEVLFNQVAPDYLEGLLLELPGN

mRadixin      LNDEIY-------CPPETAVLLASYAVQAKYGDYN -------------------------------
sfMoesin      LSDEIY-------CPPETSVLLASYAVQARHGDHN -------------------------------
hsMyoX        IHGHHP-------APEENLQVLAALRLQYLQGDYT HAAIPPLEEVYSLQRLKARISQSTKTFTPCE  1875
XeMyoX        TSGHFP-------APEETLQHLAALRLQYQHGDFS V--TWSLDTVYPVQRLKAKILQATKSSTSGH
DmMyoX        LEGLLLELPGNGVPVPEMVRDMARIAALLHRAADL -------------------------------

mRadixin      --------------- YLANDKEIHKPRLLPQRVLEQHKLTKEQWEERIQNWHEEHRGML 183
sfMoesin      --------------- FLANDPAVHGPRLLPQRVTDQHKMSREEWEQSITNWWQEHRGML
hsMyoX        RLEKRR SF LEGTLR SFRTGSVVRQKVEEEQMLDMWIKEEVSSARASIIDKWRKFQGMN 1935
XeMyoX        TLERRRTSFLEGTLK GFKVGSMRKQKVEEEQMMEMWVKEELSAARTSIAEKWSRLQGVS
DmMyoX        --------------- -SHVPAMKEIKFLLPKPALGIREIRPAQWVGLVQSAWPQVANLS

mRadixin      REDSMMEYLKIAQDLEMYGVNYFEIKNKKG-------------TELWLGVDALGLNIYEHDDKLTPKIGF 240
sfMoesin      REDAMMEYLKIAQDLEMYGVNYFEIRNKKN-------------TELWLGVDALGLNIYEKDDKLTPKIGF
hsMyoX        QEQAMAKYMALIKEWPGYGSTLFDVECKEGGFP-----------QELWLGVSADAVSVYKRGEG-RPLEVF 1994
XeMyoX        QHQAMVKYMAIVSEWPGYGPTLFDVEYKEGGFP-----------NDLWLGVSAENVSVYKRGDA-KPLETF
DmMyoX        PGQVKAQFLNVLATWPLFGSSFFAVKRIWAEEGPHVEDNHSPMWRDLILALNRRGVLFLDPNTH-ETLQHW


mRadixin      PWSEIRNISFND--------KKFVIKPIDKKAPDFVFYAPRLRINKRILALCMGNHELYMRRRKPDTI 300
sfMoesin      PWSEIRNISFN--------DRKFIIKPIDKKAPDFVFFAPRVRVNKRILALCMGNHELYMRRRKPDTI
hsMyoX        QYEHILSFGAP---------ANTYKIVVDERELLFETSEVVDVAKLMKAYISMIVKKRYSTTRSASSQ 2054
XeMyoX        QYEHIIFFGAP--------QPNTFKITVDDRELFFETTQVGEITKIMRAYINMIVKKRCSVRSVTSQD
DmMyoX        SFMEVISTRKVRSEDGALFLDMKVGNLMQQRVIRVQTEQAHEISRLVRQYITMAQISQRDKRELN

mRadixin      EVQQMKAQARVDSSGAA
hsMyoX        GSSR     2058
XeMyoX        SQSSNWAR
```

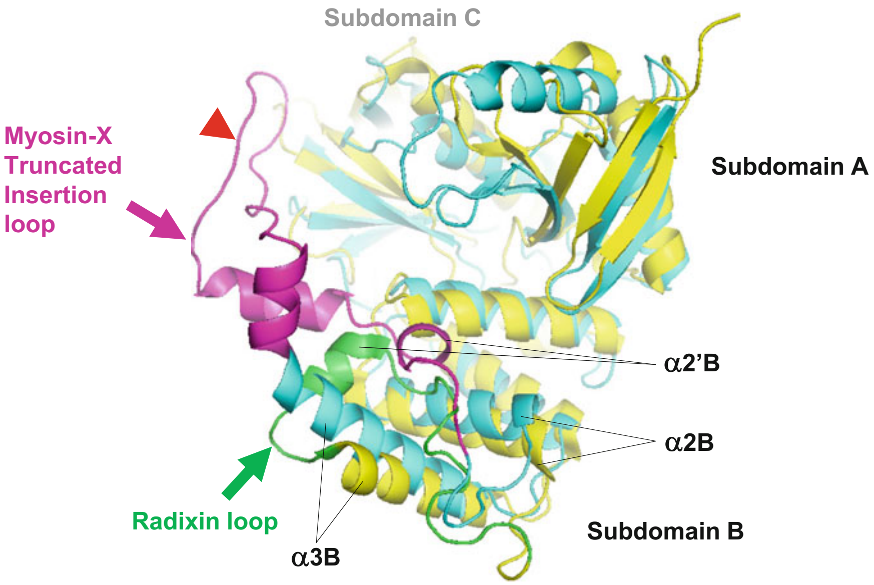


Fig. 1 Detection of non-conserved insertion of the FERM domain in the MyTH4–FERM cassette of nonconventional myosin-X in comparison with ERM proteins. *Top*: Sequence alignment of nonconventional myosin-X

transmembrane sensor of endoplasmic reticulum (ER) stress with dual protein kinase and ribonuclease activities of the cytoplasmic domain. Crystals derived from the Ire1 cytoplasmic domain were not suitable for structure determination. To improve crystallizability, the cytoplasmic domain was engineered to produce a variant containing a 24-residue (C869–F892) internal deletion within the kinase domain that removes a protease labile loop. This loop is part of 30-residue (863–898) insertion between αE and αEF and located at the C-terminal flanking region of the activation loop. It is well known that protein kinase domains frequently possess insertion at the activation loop and flanking regions, which contributes to the uniqueness of each kinase. Although the activation loop is a critical segment for kinase activity, the 24-residue deletion of Ire1 did not affect the enzymatic properties of the protein in vitro. The variant formed crystals that facilitated structural determination at 2.4 Å resolution.

In conclusion, deletion of non-conserved insertions represents one promising approach to improve protein stability and crystallizability. Careful sequence alignment of target proteins is essential for this approach. Long stretches of non-conserved insertions are primary candidates for deletion. In the case of the aforementioned examples, target proteins contained non-conserved insertions comprising more than 30 residues. The position and length of the peptide stretch to be deleted should be optimized by trial-and-error experiments followed by appropriate activity assays. The consideration of known crystal structures of homologues to the target proteins would greatly assist the design of the deletion.

2.3 Fusion and Chimera Proteins

Protein tags are routinely used in recombinant protein expression in order to facilitate the purification of target proteins [10, 11]. Other than short oligopeptides such as a hexahistidine, the use of highly soluble stable proteins, such as GST (glutathione S-transferase), MBP (maltose-binding protein), or thioredoxin, in the preparation and expression of fusion proteins can improve crystallizability and/or diffraction quality by modifying crystal

Fig. 1 (continued) (MyoX) from different sources and ERM proteins, radixin and moesin. Conserved or semi-conserved nonpolar residues are in red or orange. Non-conserved insertions are marked with *blue boxes*. The cleavage site of myosin-X during purification is in the *blue box* and highlighted with a *red circle*. The sources are mouse (m), *Spodoptera frugiperda* (sf), *Homo sapiens* (hs), *Xenopus laevis* (xl), and *Drosophila melanogaster* (dm). *Bottom*: Structural comparison between the obtained myosin-X FERM domain (cyan) with internal truncation (Δ20, see text) and the radixin FERM domain bound to the ICAM-2 peptide (*yellow*) (PDB accession code: 1J19). The structure of the radixin FERM domain represents the canonical FERM domain structures. The FERM domain contains subdomains *A*, *B*, and *C*. The non-conserved insertion found in the myosin-X FERM domain is inserted between α2B and α 3B helices of the subdomain *B*. The truncated insertion of myosin-X (*magenta*) displays a distinct conformation from that of radixin (*green*). Protease labile site is indicated with a *red arrow* contacts. This strategy was originally applied to the DNA-binding domain of DNA replication-related element-binding factor (DREF), which was crystallized as a fusion protein with *Escherichia coli* GST [12]. The key point of the utility of this technique lies in the design of the linker between the tag and the target protein, which is a determinant factor affecting crystallizability. Long linkers containing a protease cutting site and residues from multicloning sites found in commercial MBP-fusion expression kits should be converted to a shorter linker to limit conformational flexibility. Since the C-terminal end of MBP contains an α-helix, an oligo-alanine stretch was repeatedly employed for this linker [13]. This approach has successfully been applied to a variety of target proteins [14–17]. The alanine stretch of the linker is expected to form an α-helix that reduces the flexibility between the tag and target proteins, and in some cases the alanine linker exists as a loop and produces no direct contact between the tag and target proteins [17]. Generally, linkers comprising three or five alanines have been frequently tested for optimization of crystallizability of the fusion proteins. Some mutations to reduce surface entropy have also been applied to MBP in the fusion protein approach.

Another application concerning the use of fusion proteins relates to stabilization of protein–ligand or protein–protein complexes by increasing the local concentration in an effort to overcome the relatively weak affinity of ligand binding to form a complex. In this case, the ligand protein (or peptide) and its binding protein (or receptor) are fused by a linker peptide. Unlike the linker employed for fusion proteins of tag and target proteins described above, the linker in this case should be sufficiently flexible to facilitate ligand approach and direct binding to the binding site. The choice of linker length is dependent on the distance between the N- and C-terminal ends of the ligand and the binding protein. If the structure of the binding protein is known, extensive modeling could provide sufficient information for design of the linker length and connection to the N- or C-terminal end of the binding protein. If the structure of the binding protein is unknown, fusion proteins with the ligand linked to the N- or C-terminal end of the binding protein should be produced to determine which is most suitable for complex formation. Since the linker is designed to possess flexibility, small residues are employed such as glycine or a mixture of alanine and serine. For example, the structure of the complex between α-catenin and β-catenin was successfully determined using a fusion protein comprising the α-catenin-binding segment of β-catenin (residues 118–151) linked to the N-terminus of the D1 domain of α-catenin via a linker comprising five glycine residues [18]. In this fusion protein, the N-terminal 55 residues of the α-catenin D1 domain were removed since the N-terminal residues inhibit β-catenin binding to the D1 domain. Another example is a fusion protein between the myosin-X MyTH4–FERM cassette and the DCC peptide, as described above [7]. To improve the quality of the complex crystals, fusion proteins for crystallization were tested, and fusion of the DCC peptide to the C- but not N-terminal end of the myosin-X MyTH4–FERM cassette yielded high-quality crystals of the complex. This C-terminal fusion protein contained two linker residues (Ser and His) between the MyTH4–-FERM cassette and the DCC peptide as a result of the cloning process. Fortunately, the C-terminal very end of the cassette and the N-terminal very end of the DCC peptide were sufficiently flexible to form the complex. However, compared with the non-fused 1:1 complex [6], the conformation of the DCC peptide and its binding mode to the cassette was altered somewhat, probably due to the fusion. Thus, the application of fusion proteins to ligand–protein complexes should be accompanied with additional experimental tests to verify the binding mode and ligand conformation.

References

1. Kendrew JC, Parrish RG, Marrack JR, Orlans ES (1954) The species specificity of myoglobin. Nature (London) 174:946–949
2. Campbell JW, Due'e E, Hodgson G, Mercer WD, Stammers DK, Wendell PL, Muirhead H, Watson HC (1972) X-ray diffraction studies on enzymes in the glycolytic pathway. Cold Spring Harb Symp Quant Biol 36:165–170
3. Terawaki S, Kitano K, Hakoshima T (2008) Crystallographic characterization of the membrane-targeting domain of Rac-specific guanine nucleotide exchange factors Tiam1 and 2. Acta Crystallogr F64:1039–1042
4. Terawaki S, Kitano K, Mori T, Zhai Y, Higuchi Y, Itoh N, Watanabe T, Kaibuchi K, Hakoshima T (2010) The PHCCEx domain of Tiam1/2 is a novel protein- and membrane-targeting module. EMBO J 29:236–250
5. Kagiyama M, Hirano Y, Mori T, Kim S-Y, Kyozuka J, Seto Y, Yamaguchi S, Hakoshima T (2013) Structures of D14 and D14L in the strigolactone and karrikin signaling pathways. Genes Cells 18:147–160
6. Hirano Y, Hatano D, Takahashi A, Toriyama M, Inagaki N, Hakoshima T (2011) Structural basis of cargo recognition by the myosin-X MyTH4-FERM domain. EMBO J 30:2734–2747
7. Wei Z, Yan J, Lu Q, Pan L, Zhang M (2011) Cargo recognition mechanism of myosin X revealed by the structure of its tail MyTH4-FERM tandem in complex with the DCC P3 domain. Proc Natl Acad Sci U S A 108:3572–3577
8. Wu L, Pan L, Wei Z, Zhang M (2011) Structure of MyTH4-FERM domains in myosin VIIa tail bound to cargo. Science 331:757–760
9. Lee KP, Dey M, Neculai D, Cao C, Dever TE, Sicheri F (2008) Structure of the dual enzyme Ire1 reveals the basis for catalysis and regulation in nonconventional RNA splicing. Cell 132:89–100
10. Uhlen M, Forsberg G, Moks T, Hartmanis M, Nilsson B (1992) Fusion proteins in biotechnology. Curr Opin Biotechnol 3:363–369
11. Malhotra A (2009) Tagging for protein expression. Methods Enzymol 463:239–258
12. Kuge S, Fujii Y, Shimizu T, Hirose F, Matsukage A, Hakoshima T (1997) Use of a fusion protein to obtain crystals suitable for X-ray analysis: crystallization of a GST-fused protein containing the DNA-binding domain of DNA replication-related element-binding factor, DREF. Protein Sci 6:1783–1786
13. Smyth DR, Mrozkiewicz MK, McGrath WJ, Listwan P, Kobe B (2003) Crystal structures of fusion proteins with large-affinity tags. Protein Sci 12:1313–1322
14. Kobe B, Center RJ, Kemp BE, Poumbourios P (1999) Crystal structure of human T cell leukemia virus type 1 gp21 ectodomain crystallized as a maltose-binding protein chimera reveals structural evolution of retroviral transmembrane proteins. Proc Natl Acad Sci 96:4319–4324
15. Ke A, Wolberger C (2003) Insights into binding cooperativity of MATa1/MATalpha2 from the crystal structure of a MATa1

homeodomain-maltose binding protein chimera. Protein Sci 12:306–132

16. Monné M, Han L, Schwend T, Burendahl S, Jovine L (2008) Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats. Nature 456:653–657

17. Ullah H, Scappini EL, Moon AF, Williams LV, Armstrong DL, Pedersen LC (2008) Structure of a signal transduction regulator, RACK1, from Arabidopsis thaliana. Protein Sci 17:1771–1780

18. Pokutta S, Weis WI (2000) Structure of the dimerization and beta-catenin-binding region of alpha-catenin. Mol Cell 5:533–543

Part IV

Interaction Analysis

Chapter 10

Analytical Ultracentrifugation

Elena Krayukhina and Susumu Uchiyama

Abstract

Analytical ultracentrifugation (AUC) is a very useful technique to characterize macromolecular interactions. In AUC, a centrifugal force of up to about 250,000 g is applied to a solution of macromolecules, and the progression of sedimentation over time is monitored using an optical detection system. Significant advances in both hardware and software over the past few decades have greatly improved the applicability of AUC for the study of protein–protein interactions. The purpose of this chapter is to provide experimental strategies for the analysis of protein–protein interactions using AUC, including the determination of the association constant of self-associations, binding stoichiometry, and equilibrium binding constant of heterogeneous protein–protein associations. An overview of the method and software packages available for AUC data analysis and optimal protocols for the characterization of protein–protein interactions will be described.

Keywords Sedimentation velocity, Sedimentation equilibrium, Self-association, Hetero-associations, Isotherm analysis, SEDFIT, SEDPHAT

1 Introduction

AUC is an extremely useful technique for studying protein–protein interactions. It can be applied to broad molecular weight distributions (10^2–10^8 Da) to extract parameters such as equilibrium binding constant and binding stoichiometry. It is also a powerful method to assess protein stability and purity.

AUC experiments can be conducted in two basic modes of operation: sedimentation velocity (SV) and sedimentation equilibrium (SE). Regarding data collection, the major advantage of SV over SE is that the required run time is much shorter. Until recently, SE has been used to determine the buoyant molecular weight of the solutes and to estimate the stoichiometry and equilibrium constants of protein–protein interactions [1]. Recent advances in computational approaches for the analysis of SV data have made it possible to extract a wide variety of information from the SV runs [2–6]. Nonetheless, in cases where the number of species involved in the interaction is limited, SE remains the most accurate method to determine the equilibrium constant [7].

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks,
DOI 10.1007/978-4-431-56030-2_10, © Springer Japan 2016

Therefore, prior to SE, the sample purity and aggregation properties should be characterized with SV. The protocols for SV and SE will be described in Sect. 3. First, some general approaches for experimental design applicable to both SV and SE will be described.

2 General Experimental Setup

2.1 Optical Detection Systems

There are three different optical detection systems available for AUC. Considerations associated with each system are briefly summarized in Table 1.

Table 1
Optical detection systems for AUC

	Absorption optics	Interference optics	Fluorescence detection system
Selectivity	High (only components absorbing at the selected wavelength are detected)	Low (all components, including buffer salts, are detected)	High (only fluorescently labeled components are detected)
Loading concentrations	Concentrations producing 0.1 to ~1.5 OD at selected wavelength	Lower limit: concentrations producing a signal above the noise of acquisition (in general, ~0.1 mg/mL) Upper limit: concentrations below those causing nonideality effects (steep concentration gradients causing Wiener skew are to be avoided)	100 pM–1 μM
Scanning speed	~1 min per 1.2 cm solution column; radial scanning across solution column	Whole solution column imaged at once, ~10 s delay between scans	~90 s per 1.2 cm solution column; radial scanning across solution column
Signal-to-noise ratio	~300	>1000	Can be adjusted by changing the photomultiplier tube (PMT) voltage
Sample/reference volume matching	Not required	Exact same volumes should be loaded in sample and reference channels	Not required
Sample/reference component matching	Not required	Exhaustive dialysis, size-exclusion chromatography, or spin columns should be used to chemically equilibrate sample and reference	Not required

2.1.1 Absorption Optics

Absorption optics is the most commonly used optical detection system for AUC as it provides highly sensitive and selective protein detection. Typically the acceptable concentration range is from a few to several hundred micromolar, depending on the absorption coefficient and molecular weight of the protein of interest. The use of different wavelengths combined with 3 mm centerpieces could be employed to extend the applicable concentration range, and successful experiments have been conducted on 24 mg/mL (160 μM) samples [8]. Several important points should be considered when using absorption optics. To maximize the signal-to-noise ratio, the highest possible intensity of the xenon flash lamp is required (Fig. 1). Oil leaking from the vacuum pump can accumulate on the lamp surface and diminish the light output. To ensure the best performance of the lamp, the emission spectrum should be acquired periodically, and the lamp should be cleaned if a decrease in the emission of the peak at 230 nm is detected. Another concern associated with absorption optics is that at the selected wavelength, the total absorbance of the sample placed in the centrifugal cell should be within the dynamic range of the detector. In general, the detected signal should be linear with respect to the concentration of the solute up to 1.5 OD, but it depends on the intensity of the lamp at a particular wavelength. Thus, care must be taken to account for the relative contribution of various components of the solution, including the buffer (see Sect. 2.2) to the total signal. As such, the absorbance of the buffer should be measured against a water blank to determine its absorbance profile. An additional issue concerning absorption optics is that the wavelength accuracy of the monochromator incorporated into the AUC absorbance system is within 1 nm. When a wavelength from the steep portion of the spectrum is chosen for detection, the unpredicted shift of the wavelength during AUC experiment affects the quality of the recorded data and can result in the signal exceeding the dynamic range. The impact of wavelength imprecision

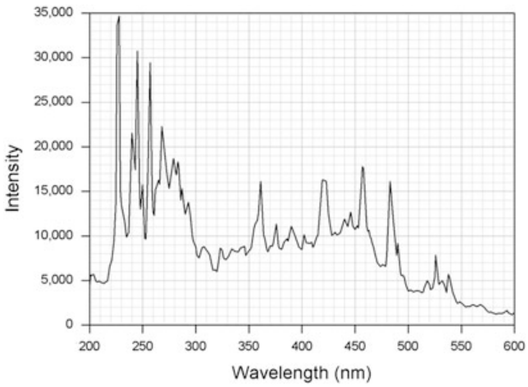


Fig. 1 The intensity profile of xenon flash lamp

can be reduced by using a relatively flat portion of the absorption spectrum such as the maximum of an absorption peak. Most proteins have an absorption peak at around 280 nm attributed to the absorption of aromatic amino acids such as tryptophan, tyrosine, and phenylalanine. Additionally, peptide bond absorption at around 220–235 nm can be used for the data acquisition of a solution with a low absorption at 280 nm. Thus, for the AUC measurements, the recommended wavelengths are 230 and 280 nm for low and medium concentrations, respectively. Highly concentrated solutions can be monitored at around 290 nm. Nevertheless, in highly concentrated solutions, the Wiener skewing effect [9] caused by the large difference in the refractive index between the solvent and solution interferes with the accurate monitoring of sedimentation profiles.

The noise of data acquisition is usually 0.005–0.01 OD, and, considering the upper limit of the dynamic range of 1.5 OD, the maximum achievable signal-to-noise ratio is approximately 300.

2.1.2 Interference Optics

In interference optics, the signal detection is based on the difference of the refractive index between the sample and reference. All components in the solution, including the buffer, contribute to the signal detected by interference optics, and different salt distributions in the sample and reference can affect the recorded signal. To obtain high-quality data, it is imperative to allow the sample and respective reference solvent to chemically equilibrate. This can be achieved through exhaustive sample dialysis against the solvent solution. Another approach is to use the sample after elution from a gel filtration column with the mobile phase being used as the reference solvent. Spin gel filtration columns have also been successfully applied for a similar purpose. Despite these technical challenges, the temporal and radial resolution of data recorded using interference optics is significantly better compared to absorbance data. An entire solution column is imaged at once with the radial step size of approximately 0.002 cm, and the time delay between consecutive scans is only 10 s. There is no specific upper concentration limit; however, significantly steep gradients should be avoided. Solvents containing strongly absorbing compounds, such as ATP, do not pose limitations on signal detection using interference optics.

2.1.3 Fluorescence Optics (Fluorescence Signal Detection System)

Recent developments of fluorescence signal detection system [10] have made it possible to use AUC for the analysis of high-affinity interactions. In addition, such system enables the detection of the sedimentation of the component of interest in complex solutions such as blood serum where other light-absorbing species are present [11]. The covalent attachment of fluorescent dyes required for fluorescence-detected AUC analysis can potentially affect the sedimentation behavior of a molecule due to modifications in its size or shape. Therefore, the impact of labeling on the structure, activity, or associations of the macromolecule should be examined. Fluorescent emission is detected in the wavelength range of 505–565 nm using laser excitation at 488 nm. Extrinsic dyes with the same excitation wavelength, such as fluorescein, Alexa Fluor 488, Oregon Green, and green fluorescent protein, can be used to label target molecules. At low concentrations, the adsorption of the protein of interest to the windows and centerpiece can potentially interfere with the analysis; thus, for low concentrations, the addition of a "carrier" protein is recommended [10, 12]. Low concentrations (0.1 mg/mL) of ovalbumin, serum albumin, and kappa casein have been used for this purpose.

2.2 Buffers

Buffers used in AUC experiments should contain sufficient salt concentrations to shield unfavorable electrostatic repulsions between molecules. If possible, no gradient forming additives, such as glycerol or sucrose, should be added to the buffer solution. For uncommon solvents, chemical resistances (http://www.uslims.uthscsa.edu/compatibility.php) should be evaluated to select a suitable centerpiece that is compatible with the solvent. In general, it is preferable to use nonabsorbing buffers. For samples with reducing agents, it should be noted that most reducing agents demonstrate significant absorption in the near-UV range that changes in a time-dependent manner. TCEP (tris(2-carboxyethyl) phosphine) is recommended to maintain the reduced state of cysteine residues during the AUC measurement.

3 Methods

3.1 Sedimentation Velocity (SV)

3.1.1 Introduction

SV is a hydrodynamic method that provides information on the size and shape of the solute. SV is applied for the determination of the solute's sedimentation coefficient distribution and to gain limited information on the hydrodynamic shape of the solute.

In SV, the solute sediments under a strong gravitational field and the sedimentation and diffusion fluxes govern the behavior of the particle. The partial differential equation describing the evolution of concentration profiles $C(r,t)$ at each radial position r and time t during the sedimentation process is the Lamm equation [13]:

$$\frac{\partial C}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}\left(rD\frac{\partial C}{\partial r} - s\omega^2 r^2 C\right), \tag{1}$$

where s and D are the sedimentation and diffusion coefficient of the solute, respectively, and ω is the angular speed. The Lamm equation is derived from equations governing sedimentation and diffusion transport processes combined with the balance equation of centrifugal, buoyant, and drag frictional forces acting on the solute molecule. For a mixture of non-interacting solutes, the total concentration of all solutes can be represented by a sum of Lamm equation solutions L for each solute in the mixture multiplied by the partial concentration c_n:

$$C(r,t) = \sum [c_n L(s_n, D_n, r, t)] \tag{2}$$

Analysis of the sedimentation data by the Lamm equation can provide information about solute sedimentation and the diffusion coefficient. Unfortunately, the Lamm partial differential equation has no general analytical solution. However, the recent availability of powerful computers has favored the development of computer programs for the numerical analysis of sedimentation experiments.

Sedimentation coefficient s (Svedberg units, $1\text{S} = 10^{-13}$ s) corresponds to speed u at which the solute molecule moves in the centrifugal field $\omega^2 r$:

$$s = \frac{u}{\omega^2 r} == \frac{M(1-\overline{v}\rho)}{Nf} = \frac{MD(1-\overline{v}\rho)}{RT}, \tag{3}$$

where M is molecular mass, $\overline{v}$ is the partial specific volume, f is the translational frictional coefficient, ρ is the buffer density, T is the absolute temperature, R is the universal gas constant, and N is Avogadro's number.

The diffusion coefficient D can be conveniently expressed through the frictional ratio f/f_0 by using the Stokes–Einstein relationship:

$$D = \frac{RT}{18\pi N(f/f_0\eta)^{3/2}\sqrt{\frac{s\overline{v}}{2(1-\overline{v}\rho)}}}, \tag{4}$$

where η is the buffer viscosity. A frictional ratio is defined as the frictional coefficient of a protein f divided by the frictional coefficient f_0 of a non-hydrated sphere of equal mass and indicates the degree of globularity of the proteins. While a non-hydrated sphere has a frictional ratio equal to 1, most globular proteins have f/f_0-values in the range 1.2–1.8. For elongated molecules, frictional ratio values can be greater than 2.

The molecular mass M can be derived from the obtained parameters (s, D) using the Svedberg equation:

$$M = \frac{sRT}{D(1-\overline{v}\rho)} \tag{5}$$

3.1.2 Experimental Design and Execution

Protocol 1

1. Choose the appropriate sample concentration. To determine if the protein of interest self-associates, the initial SV runs should be performed with at least three different protein concentrations. The initial cell-loading concentrations should cover an approximately tenfold concentration range. To study the hetero-association of two proteins (A and B), the SV

experiments should be conducted with at least one concentration of A and B alone and at least three mixtures prepared with different concentrations of A and B. In general, the mixtures are prepared in the following manner: the concentration of A is kept constant within a few folds of the expected k_d, and the concentration of B is varied approximately tenfold below and above the expected k_d.

2. Choose the appropriate optical detection system. The choice depends on the concentration range and the nature of the protein (for details, see Table 1 and Sect. 2.1)).
3. Choose the appropriate solvent: see Sect. 2.2.
4. Choose the appropriate centerpieces and load samples into cells. In most cases, standard double-sector centerpieces can be utilized. The sectors are filled with 400–450 μL of the sample. It should be noted that longer solution columns produce higher hydrodynamic resolution and better quality data can be collected for a longer amount of time. In cases where absorption optics is used, the reference sector should be filled with the buffer solution, the volume of which should exceed the sample volume by 5–10 μL to avoid complications caused by signal from the solvent meniscus. When interference optics is utilized, the volumes of the sample and reference should match. Preferably, meniscus-matching centerpieces should be used. However, if the sample and reference menisci are not precisely matched, this can be accounted for computationally during data analysis using SEDFIT software [14].
5. Choose the appropriate temperature. The sample must be stable at the experimental temperature over the course of the experiment. For most applications, 20 °C is appropriate. For the special cases, temperatures between 4 and 40 °C are available using Beckman Coulter XL-A/I ultracentrifuges. Before the run, carefully equilibrate the rotor with the samples loaded at 0 rpm for at least 30 min after the rotor reaches the target temperature. It is important to avoid convection at the beginning of the run, which is caused by the mixing of the solution layers of different temperatures.
6. Choose the appropriate rotational speed and scan interval. A speed should be chosen so that at least 40 scans can be recorded before the sedimentation of the sample is complete. Simulations available in SEDFIT [15] or UltraScan [16] software packages estimate the optimum speed and consequently an approximate time to complete sample sedimentation (http://www.analyticalultracentrifugation.com/generating_simulated_data.htm; http://www.ultrascan3.uthscsa.edu/manual/astfem_sim.html). The scan interval should be as short as possible, but the sample should completely sediment before the maximum number of the scans (999) is reached.

7. Start the method scan. Collect data until the sample sedimentation is complete, which generally requires between 2 and 12 h depending on the solute size and rotational speed.
8. Stop the run. In principle, after the AUC experiment, the samples can be recovered from the cell assembly. However, due to possible changes in the structure and aggregation state of the solutes, this is not generally recommended.
9. Clean the components of the cell assembly. It is a good practice to use the same combination of cell housing, windows, and centerpiece during cleaning and assembly. In this manner, the defective components affecting the quality of the data can be easily detected and eliminated.

3.1.3 Data Analysis

3.1.3.1 Determination of Sedimentation Coefficient Distribution

The most commonly used approach to initial data analysis is the sedimentation coefficient distribution, C(s), implemented in the SEDFIT software [3, 15]. This method requires no prior knowledge of sample properties and can be conveniently used to determine the number of sedimenting species, sedimentation coefficients, and molecular masses. C(s) is a direct least-squares method for modeling experimental data using numerical solutions of the Lamm equation. To calculate diffusion coefficients, Eq. 4 is used, where it is assumed that all sedimenting species have the same frictional ratio f/f_0. This assumption is based on the lower size dependence of diffusion relative to sedimentation and weak shape dependence of the frictional ratio. The weight-average f/f_0 value can be optimized in a nonlinear regression during C(s) analysis. For heterogeneous systems, where multiple species with different shapes are present at comparable concentrations, a single frictional ratio is not suitable to describe all the components and results in skewed molecular mass determinations. However, when a single peak is seen in the C(s) distribution, the molecular mass estimation can be expected to be within 10 % of the true value.

Protocol 2

1. Load scan files into SEDFIT. The data are color-coded according to the acquisition time: scans recorded at the beginning of the experiment are shown in black and the latest scans are indicated in red. Select the appropriate number of scans so that the transition from a green to red color is seen in the middle of the loaded data set.
2. Specify the meniscus, bottom position, and fitting limits. Set meniscus (red line) to the midpoint position of the absorbance spike corresponding to the air–sample boundary. Set the bottom position (blue line) to the maximum signal corresponding to optical artifacts at the end of the solution column. Set the left and right data analysis limits (green lines) to exclude the region of optical artifacts close to the meniscus and bottom.

3. Choose continuous C(s) distribution from the "Model" menu. In the "Parameter" box, input the minimum (s_{min}) and maximum (s_{max}) expected sedimentation coefficient values. Input the resolution. This parameter corresponds to the number of species with different s-values between s_{min} and s_{max} in which relative abundance will be determined in the C(s) analysis. Input the initial value for the frictional ratio: 1.2 for globular proteins, 1.5 for antibodies or other asymmetrically shaped proteins, and 2.0 or higher for rod-shaped and unfolded proteins, fibrils, and DNA. Input the values for partial specific volume (vbar), solvent density, and viscosity. Set the confidence level to 0.68. Check the boxes for the frictional ratio, baseline, meniscus, and time-independent noise (and radial-independent (RI) noise when interference optics is used for the data acquisition) in order to optimize these parameters.
4. Use the "Run" command to estimate the initial guesses for the parameters entered in the previous step. If the distribution significantly deviates from zero at the minimum or maximum s-value, select a higher value for s_{max} and a lower value for s_{min}, respectively. Execute the "Run" command with refined parameters. Repeat until there are no peaks at the maximum and minimum s-value in the C(s) distribution.
5. Optimize the initial parameters by executing the "Fit" command. Assess the quality of the fit by verifying that the root mean-square deviation (rmsd) does not exceed 0.1 % of the total loading signal value. The randomness of the residuals can be ensured by the absence of visible diagonal lines at the residuals bitmap. If a good quality optimization is achieved, the peaks in the resulting C(s) distribution correspond to the sedimenting species. The displayed fitted frictional ratio should be consistent with the known properties of the sample (folded/unfolded chains) and should always be >1. Values <1 indicate extra boundary broadening not originating from diffusion, but likely from rapid (k_{off} >0.01/s) chemical reactions.
6. Estimate the molecular weights of the detected species by choosing "Display Mw peaks in C(s)" from the "Display" menu or by clicking Ctrl-M. The obtained values should be interpreted with care (see Sect. 2.1).

3.1.3.2 Isotherm Analysis

The isotherm of weight-average sedimentation coefficients, s_w, as a function of protein concentration is constructed. The experiments performed at different protein concentrations are analyzed to elucidate if reversible self-association is present. Available methods for data analysis include g*(s) [17], van Holde–Weischet analysis [18], and two-dimensional spectrum analysis [4], with the C(s) analysis being the method of choice. Even though the C(s) analysis is based on the assumption that all solutes are non-interacting, the integration of the size-distribution profiles over the entire sedimentation coefficient range provides a correct weight-average sedimentation coefficient, from which the equilibrium constant for a self-association or hetero-association of the solutes can be characterized.

In addition to the determination of the presence of associations, C(s) allows for the estimation of the kinetics of those interactions. If the peaks are broad and their positions are concentration dependent, then there is a fast reaction taking place. In contrast, for a slow reversible system, the peaks would be sharper and at constant positions, and only the relative peak heights would vary with concentration.

Protocol 3

1. Analyze the collected data according to Protocol 2. Integrate the area under the corresponding peaks by selecting “Integrate distributions” under “Size-distributions options” under the “Options” menu of the SEDFIT main window or simply by clicking Ctrl-I. Note the weight-average sedimentation coefficients and write them in a second column in a tab-delimited text file, with the first column representing the loading concentrations. Alternatively, the signal-average sedimentation coefficient isotherms can be conveniently constructed using GUSSI software (http://biophysics.swmed.edu/MBR/software.html).
2. Load the isotherm file into the SEDPHAT window. In the “Experimental parameter” box, input the partial specific volume, buffer density and viscosity, extinction coefficient, and optical path length.
3. Choose the appropriate model from the “Model” menu and execute the “Fit” command. To increase the precision of the determined k_d, prior knowledge of the sedimentation coefficients of either individual components or complexes can be incorporated in the analysis. In self-associating systems, the sedimentation coefficients can be derived from available crystal structures by constructing hydrodynamic bead models using SOMO [19] or HYDROPRO [20]. For hetero-associating systems, the sedimentation coefficients of A and B can be derived from the experiments performed using the individual components.

An example of isotherm analysis conducted to study the self-association of semaphorin 6A (Sema6A) receptor-binding fragment is presented in Fig. 2a [21]. Figure 2b provides an example of the monomer–dimer–tetramer equilibrium of wild-type hemoglobin.

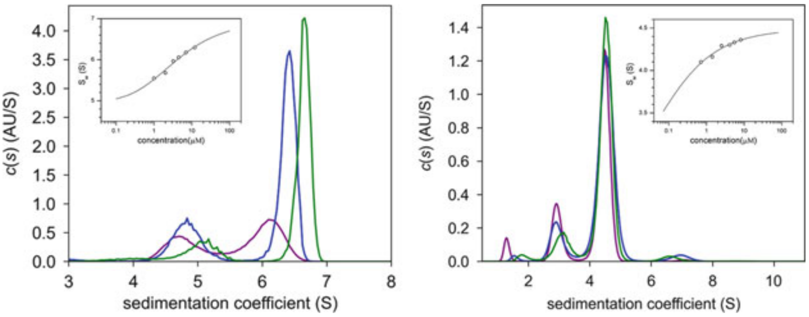


Fig. 2 Examples of isotherm analysis conducted to determine the dissociation constant of self-associating proteins. (**a**) C(s) distribution from SV experiments performed at different concentrations of the semaphorin 6A receptor-binding fragment Sema6A$_{SP}$. For the clarity of presentation, only the distributions calculated for 1 (*purple*), 3.25 (*blue*), and 12 μM (*green*) data are plotted. The concentration-dependent change observed in the sedimentation coefficient distribution indicates the presence of a monomer–dimer equilibrium. The isotherm analysis of the weight-average sedimentation coefficients yielded a k_d value of 3.5 μM (Adapted from ref. 21). (**b**) C(s) distribution from SV experiments performed at different concentrations of wild-type human hemoglobin. For the clarity of presentation, only the distributions calculated for 2.5 (*purple*), 7.5 (*blue*), and 10 μM (*green*) data are plotted. The concentration-dependent changes observed in the areas of the peaks indicate the presence of a dimer–tetramer equilibrium. The isotherm analysis of the weight-average sedimentation coefficients yielded a k_d value of 0.1 μM

3.1.3.3 Direct Boundary Modeling of SV Data: Using Prior Knowledge from Non-denaturing Mass Spectrometry

In SV, information about the molecular mass of the species is obtained using the frictional ratio parameter, which is extracted from modeling the sedimentation boundary spreading. For multicomponent solutions which contain reactive species with a broad range of sizes or shapes, the determination of molecular masses is often difficult, as in addition to diffusion, the shape of the sedimentation boundary is dependent on both conformational heterogeneity and reaction kinetics [22]. Consequently, if the model applied for the data analysis does not account for either of the factors, the estimates of the obtained parameters may be incorrect. Likewise, incorporating all factors in the fitting model can significantly complicate the analysis and potentially compromise the results.

An alternative approach to SV is mass spectrometry (MS) which is capable of providing the most accurate molecular mass determination. Nonetheless, nonspecific interactions occurring during the electrospray ionization process can affect the distribution of oligomeric species. Therefore, the combination of SV and MS may be useful for the characterization of complex protein solutions.

The study of the assembly states of the nucleosome assembly protein 1 (NAP-1) reported by Noda et al. [6] highlights the utility of proposed technique. Prior to SV, the oligomeric states of NAP-1 were characterized by MS under non-denaturing conditions. The results indicated that the primary oligomeric unit of NAP-1 was a dimer, and a portion of the dimers further assembled into higher oligomers. Then, the assembly states of NAP-1 in solution were characterized using SV. The initial data analysis performed using

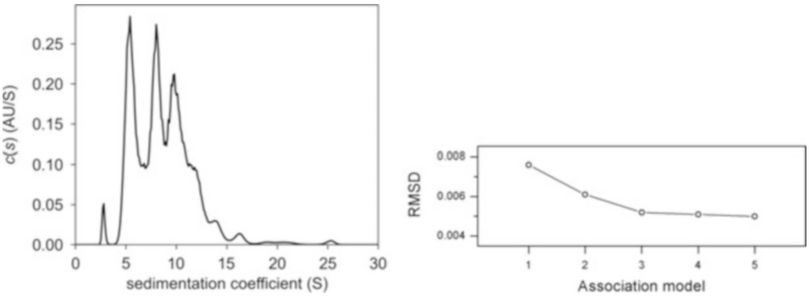


Fig. 3 Analysis of SV data from the study of the assembly states of the nucleosome assembly protein 1 (NAP-1). (**a**) C(s) distribution of human NAP-1 at 150 mM NaCl. (**b**) Plot of RMSD values from the results of "Hybrid local continuous distribution and global discrete species" analysis by the program SEDPHAT of human NAP-1. The association model number indicates *1* 1-2-4-6-8-mer model, *2* 1-2-4-6-8-10-mer model, *3* 1-2-4-6-8-10-12-mer model, *4* 1-2-4-6-8-10-12–14-mer model, and *5* 1-2-4-6-8-10-12-14-16-mer model (Adapted from Ref. 21)

the C(s) model of SEDFIT allowed accurate determination of the sedimentation coefficients and relative concentration of each oligomeric species (Fig. 3a). The assignment of molecular mass to the peaks detected in the C(s) distribution, however, was complicated by the heterogeneity of the sample and the single weight-average f/f_0 value was not suitable to describe each component individually. Thus, the findings from the non-denaturing MS measurements were incorporated as prior knowledge in the SV data analysis using SEDPHAT. "Hybrid local continuous distribution and global discrete species" analysis using a number of different models including 1-2-4-6-8-mers, 1-2-4-6-8-10-mers, 1-2-4-6-8-10-12-mers, 1-2-4-6-8-10-12-14-mers, and 1-2-4-6-8-10-12-14-16-mers was performed. With the increasing number of oligomeric species included in the model, the rmsd value decreased demonstrating a higher-quality fit (Fig. 3b). The 1-2-4-6-8-10-12-mers, 1-2-4-6-8-10-12-14-mers, and 1-2-4-6-8-10-12-14-16-mers models showed similar rmsd values, which indicated that the 1-2-4-6-8-10-12-mers model was the most appropriate for the data set analysis, according to the principle of parsimony.

3.1.3.4 Direct Boundary Modeling of SV Data: The Estimation of Kinetic Information for Systems with Reversible Associations

The sedimentation coefficients and equilibrium constants obtained from isotherm analysis can be further refined using the direct Lamm equation modeling approach.

Protocol 4

1. Load xp-files of the SV experiments into SEDPHAT. These files can be prepared while analyzing data in SEDFIT to construct the isotherm of the weight-average sedimentation coefficients. Detailed instructions on the preparation of xp-files are available elsewhere (http://analyticalultracentrifugation.com).

2. Select the model and enter the starting values for the species *s*-values and the equilibrium constant from the isotherm analysis. Estimate the chemical off-rate constant $\log_{10}(k_{off}) = -3$ for rapid interactions or -4 to -5 for slow interactions relative to sedimentation.
3. Fit this model by first optimizing only the starting concentrations. At the next step, allow the algorithm to perform the optimization of the equilibrium binding and reaction rate constants and the species *s*-values.
4. Evaluate the fit by noting the rmsd value and randomness of the residual distribution.
5. Different models can be tested. The one producing the lowest rmsd value coupled with a random distribution of residuals should be considered as the most appropriate.

3.1.3.5 Multi-signal SV (MSSV)

Multi-signal SV (MSSV) is a SV technique utilized in the study of heterogeneous protein interactions. A detailed description of this method is available in [23]. MSSV enables the investigation of binary and ternary complexes formed in mixtures of three different proteins. To resolve interacting components in MSSV, the components must show sufficiently different spectral signatures. To evaluate whether MSSV is a suitable approach for a particular mixture, the value of D_{norm} [24] is calculated based on the known extinction coefficients. Successful examples of three protein-component mixtures analyzed by MSSV are described in [25, 26].

3.2 Sedimentation Equilibrium (SE)

3.2.1 Introduction

SE experiments are conducted at lower rotational speeds than SV experiments. The sedimentation flow is opposed by counterflow diffusion that is generated according to the derivative of the concentration at a radial position. At the equilibrium state, the sedimentation force applied to the solute is balanced by the diffusion force, leading to the formation of a steady-state exponential concentration gradient. SE provides information on the total profile of detectable solute with a selected optical detection system, and therefore high purity samples containing a small number of species are preferred. Analysis of the sample by SV should be carried out prior to the SE to confirm the absence of impurities.

SE experiments provide information about solute buoyant molar mass, association constants, association stoichiometries, and second viral coefficient related to the thermodynamic nonideality of the solution. Similar to SV, the behavior of the particle in the cell is described by the Lamm equation. Unlike SV, in SE the system is studied at equilibrium, and thus the total flux, comprised of sedimentation flux and opposing diffusion flow, equals 0:

$$s\omega^2 rC - D\frac{\partial C}{\partial r} = 0 \tag{6}$$

The solution of this equation corresponds to the exponentially increasing concentration profile:

$$C(r) = C_0 e^{\frac{s\omega^2}{2D}\left(r^2 - r_0^2\right)} \tag{7}$$

where C_0 is concentration at a radial reference point r_0 in the concentration gradient. By inserting the Svedberg Eq. 5, the following expression is derived:

$$C(r) = C_0 e^{M(1-\overline{v}\rho)\frac{\omega^2}{2RT}\left(r^2 - r_0^2\right)} \tag{8}$$

Thus, the steepness of the concentration gradient at any particular rotor speed is determined by the buoyant molecular mass $M_b = M(1 - v_{\text{bar}}\,\rho)$. In contrast to SV, the molecular shape of the solute has no effect on the result of SE experiments within ideal solutions. The buoyant molecular mass thus can be obtained from SE experiments, and the weight-average molecular weight of the macromolecule of interest, M, can be estimated given an accurate partial specific volume. The partial specific volume can be determined experimentally by measuring the concentration dependence of the protein solution or by using density contrast in mixtures of light and heavy water [27, 28] or theoretically from the amino acid composition of the protein using SEDNTERP (http://sednterp.unh.edu/).

For a mixture of solutes, the total equilibrium concentration gradient is expressed by the following equation:

$$C_{\text{total}}(r) = \sum\nolimits_i C_{0,i} e^{M_i(1-\overline{v}_i\rho)\frac{\omega^2}{2RT}\left(r^2 - r_0^2\right)} + \text{baseline}, \tag{9}$$

where C_0 of the complex can be described using the C_0 values of each component and the equilibrium constant of the interaction between or among the components. In the nonlinear fitting of SE data, the C_0 values, baseline, and k_d are set as variable parameters, while M_i and v_i are typically calculated based on the amino acid composition and are set as fixed parameters.

3.2.2 Experimental Design and Execution

3.2.2.1 Self-Association by SE (Example $A + A = A_2$)

Protocol 5

1. Choose the appropriate sample concentration. In order to determine the association constant, a broad concentration range with multiple loading concentrations should be used. At low concentrations, monomers will primarily contribute to the signal, while at high concentrations the signal will be dominated by oligomeric forms. Prior to SE experiments, it is highly preferable to characterize the sample by SV according to Protocol 1 and Protocol 2. The sample should be well purified (typically more than 95 % purity) and chemically equilibrated with its reference solvent if interference optics is utilized.

2. Choose the appropriate sample volume. Usually 3 mm solution columns (100–120 μL) are sufficient for SE experiments. Longer columns require longer times to reach equilibrium; however, concentration gradients extending longer distances provide better parameter precision. For low molecular mass proteins, higher volumes may be required to produce a concentration gradient with a sufficient length of curvature. Before performing the experiment, it is recommended to simulate data using the "Estimate equilibrium rotor speeds" option under the "Calculator" menu of SEDFIT.

3. Choose the appropriate optical detection system. The choice depends on the concentration range and the nature of the protein including the amino acid composition (for details, see Table 1 and Sect. 2.1).

4. Prepare cells for sample loading. If interference optics is chosen, before loading the samples, the assembled cells should be mechanically "aged" (for details, see http://analyticalultracentrifugation.com) to minimize the impact of time-independent noise, which can change over the time course of the SE experiment due to mechanical micro-movements of the assembly parts. Similar to Protocol 1, the same (interference optics) or 5–10 μL larger solvent volumes (absorption optics) should be loaded in the reference sector.

5. Choose the appropriate temperature. The sample should be stable at the experimental temperature during the equilibrium run (depending on the settings, the run might require 1 week or longer). For most applications 20 °C is an appropriate choice. For special cases, temperatures between 4 and 40 °C are available using the XL-A/I ultracentrifuge. In contrast to the SV run, there is no need to equilibrate the rotor with the samples loaded at the target experiment temperature.

6. Choose the appropriate rotational speed. A single speed cannot distinguish interacting and non-interacting species when a sample solution with a single concentration is measured. Therefore, three rotor speeds should be chosen for the experiment. The slowest rotational speed provides a shallow gradient resulting in information about the largest species in the sample. At the highest rotational speed, meniscus depletion should be achieved and a steep concentration gradient should be observed. This data set provides information about the smallest species. Simulations available in SEDFIT or UltraScan software packages allow for the convenient estimation of the best speed (see http://www.analyticalultracentrifugation.com/generating_simulated_data.htm, http://www.ultrascan2.uthscsa.edu/manual2/finsim.html for details).

7. Collect the data. Start collecting multispeed data from the lowest speed chosen in the previous step. Data are collected every 6 h and successive scans are compared by the SEDFIT or WinMatch program. Equilibrium is attained when the subtraction of two consecutive scans produces no systematic difference. The minimum time required to reach equilibrium can be estimated using the "Calculator" menu of SEDFIT. Once equilibrium has been attained, the data can be collected at the next speed.
8. After equilibrium is attained at the highest rotor speed and all the data have been collected, stop the run and clean the components of the cell assembly.

3.2.2.2 Hetero-associations by SE (Example A + B = AB)

Protocol 6

1. Prepare a series of sample concentrations. Each component should be measured individually and a mixture of components should be prepared as a dilution or titration series. To avoid nonideality, which complicates data interpretation and analysis, the concentration range should be chosen within 0.1–10 × k_d, producing an absorbance signal within 0.1–0.75 OD or larger than 0.1 fringes. Again, if interference optics is chosen, the sample should be free from impurities and equilibrated with its reference solvent.
2. Choose the appropriate sample volume (see Protocol 5).
3. Choose the appropriate optical detection system. The detection at multiple wavelengths (230, 250, and 280 nm) combined with interference allows for a wide range of suitable loading concentrations.
4. Load cells with the sample, choose the appropriate temperature and rotational speed, and collect the data according to Protocol 5.

3.2.3 Data Analysis

Protocol 7

1. In SEDFIT, preprocess the data for further analysis using SEDPHAT. In the "Loading Options" menu of SEDFIT, choose "Sort EQ data to Disk" and convert equilibrium data to (*.xp) files suitable for the SEDPHAT analysis (for details, refer to http://www.analyticalultracentrifugation.com/se protocols.htm).
2. Analyze the data collected for the individual components. In SEDPHAT, load the xp-files associated with only one component of the interacting system. In the "Model" menu, select "A (single species of interacting system)."
3. Analyze the data acquired for the mixtures of components. In SEDPHAT, load the xp-files associated with the mixtures of

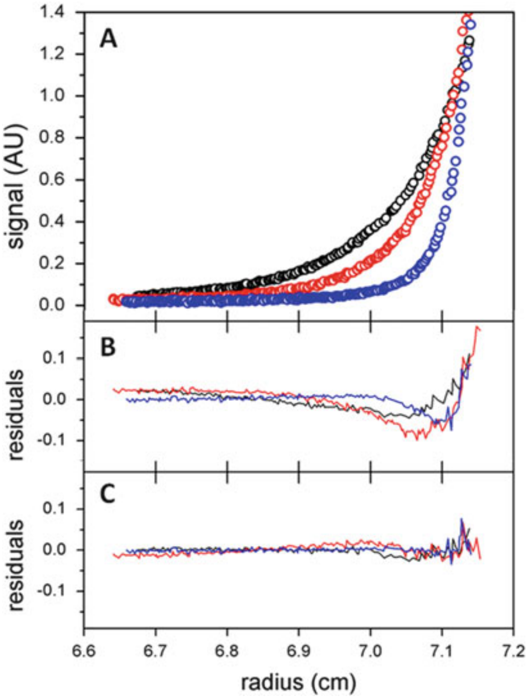


Fig. 4 SE analysis of antibody and antigen interaction. (**a**) SE concentration gradient for mAb (antibody) to NP and NP-conjugated BSA (antigen) mixed solutions each at 3.3 μM (equimolar condition). (**b**) Nonrandomly distributed residuals and high chi-squared value of 0.0231112 indicate that 1:1 interaction model is inadequate in this case. (**c**) Randomly distributed residuals and significant improvement of chi-square (0.00555707) support the 1:2 interactions

interacting components, and in the "Model" menu, choose one of the models for heterogeneous A and B interactions. To initiate the analysis, use the parameters obtained during the previous step. At this stage, time-independent noise decomposition should not be attempted as it can correlate with the model used. Different models should be tested, and the model providing the best quality fit, which is evaluated using the rmsd values for each xp-file and the randomness of the residuals, should be considered the most appropriate. Then, include "TI noise" and allow the algorithm to optimize the parameters. This should result in a decrease of the rmsd value. Ensure a relatively flat TI-noise profile with no apparent curvature.

4. Alternatively, the stoichiometry and k_d can be estimated from the nonlinear least-squares fitting of acquired data to Eq. (8) by a homemade program using software equipped with a nonlinear fitting algorithm, such as Mathematica. An example of SE analysis of antibody and antigen interaction is presented in Fig. 4.

4 Note

Recent findings suggested that time stamps recorded in the sedimentation scan files by AUC software were incorrect, leading to errors in sedimentation coefficients and molecular weight estimations [29, 30]. Even though it was discussed that the binding constants obtained from the application of isotherm analysis are unaffected by the incorrect time stamp, the absolute values of the sedimentation coefficients will be incorrect. Therefore, the use of SEDFIT (version 14.0c or later) software is recommended to compensate for possible errors.

References

1. Kato K, Sautes-Fridman C, Yamada W et al (2000) Structural basis of the interaction between IgG and Fcgamma-receptors. J Mol Biol 295:213–224
2. Philo JS (2000) A method for directly fitting the time derivative of sedimentation velocity data and an alternative algorithm for calculating sedimentation coefficient distribution functions. Anal Biochem 279:151–163
3. Schuck P (2000) Size distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and Lamm equation modeling. Biophys J 78:1606–1619
4. Brookes E, Cao W, Demeler B (2010) A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. Eur Biophys J 39:405–414
5. Oda M, Uchiyama S, Noda M et al (2009) Effects of antibody affinity and antigen valence on molecular forms of immune complexes. Mol Immunol 47:352–364
6. Noda M, Uchiyama S, McKay AR et al (2011) Assembly states of the nucleosome assembly protein 1 (NAP-1) revealed by sedimentation velocity and non-denaturing mass spectrometry. Biochem J 436:101–112
7. Oda M, Uchiyama S, Robinson CV et al (2006) Regional and segmental flexibility of antibodies in interaction with antigens of different size. FEBS J 273:1476–1487
8. Nishi H, Miyajima M, Nakagami H et al (2010) Phase separation of an IgG1 antibody solution under a low ionic strength condition. Pharm Res 27:1348–1360
9. Svensson H (1954) The second order aberrations in the interferometric measurement of concentration gradients. Optica Acta 1:25–32
10. Kingsbury JS, Laue TM (2011) Fluorescence-detected sedimentation in dilute and highly concentrated solutions. Methods Enzymol 492:283–304
11. Demeule B, Shire SJ, Liu J (2009) A therapeutic antibody and its antigen form different complexes in serum than in phosphate-buffered saline: a study by analytical ultracentrifugation. Anal Biochem 388:279–287
12. Cole JL, Lary JW, P Moody T, Laue TM (2008) Analytical ultracentrifugation: sedimentation velocity and sedimentation equilibrium. Methods Cell Biol 84:143–179
13. Lamm O (1929) Die differentialgleichung der ultrazentrifugierung. Ark Mater Astr Fys 21B: 1–4
14. Zhao H, Brown PH, Balbo A et al (2010) Accounting for solvent signal offsets in the analysis of interferometric sedimentation velocity data. Macromol Biosci 10:736–745
15. Schuck P (2005) Diffusion-deconvoluted sedimentation coefficient distributions for the analysis of interacting and non-interacting protein mixtures. In: Scott DJ, Harding SE, Rowe AJ (eds) Analytical ultracentrifugation: techniques and methods. RSC Publishing, Cambridge, pp 26–49
16. Demeler B (2005) Ultrascan: a comprehensive data analysis software package for analytical ultracentrifugation experiments. In: Scott DJ, Harding SE, Rowe AJ (eds) Analytical ultracentrifugation: techniques and methods. RSC Publishing, Cambridge, pp 210–229
17. Schuck P, Rossmanith P (2000) Determination of the sedimentation coefficient distribution g*(s) by least-squares boundary modeling. Biopolymers 54:328–341
18. Demeler B, van Holde KE (2004) Sedimentation velocity analysis of highly heterogeneous systems. Anal Biochem 335:279–288
19. Brookes E, Demeler B, Rosano C, Rocco M (2010) The implementation of SOMO

(SOlution MOdeller) in the UltraScan analytical ultracentrifugation data analysis suite: enhanced capabilities allow the reliable hydrodynamic modeling of virtually any kind of biomacromolecule. Eur Biophys J 39:423–435

20. Ortega A, Amorós D, Garcia de la Torre J (2011) Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. Biophys J 101:892–898
21. Nogi T, Yasui N, Mihara E et al (2010) Structural basis for semaphorin signalling through the plexin receptor. Nature 467:1123–1127
22. Dam J, Velikovsky CA, Mariuzza RA et al (2005) Sedimentation velocity analysis of heterogeneous protein-protein interactions: Lamm equation modeling and sedimentation coefficient distributions c(s). Biophys J 89: 619–634
23. Padrick SB, Deka RK, Chuang JL et al (2010) Determination of protein complex stoichiometry through multisignal sedimentation velocity experiments. Anal Biochem 407:89–103
24. Padrick SB, Brautigam CA (2011) Evaluating the stoichiometry of macromolecular complexes using multisignal sedimentation velocity. Methods 54:39–55
25. Houtman JC, Yamaguchi H, Barda-Saad M et al (2006) Oligomerization of signaling complexes by the multipoint binding of GRB2 to both LAT and SOS1. Nat Struct Mol Biol 13:798–805
26. Barda-Saad M, Shirasu N, Pauker MH et al (2010) Cooperative interactions at the SLP-76 complex are critical for actin polymerization. EMBO J 29:2315–2328
27. Edelstein SJ, Schachman HK (1967) The simultaneous determination of partial specific volumes and molecular weights with microgram quantities. J Biol Chem 242:306–311
28. Brown PH, Balbo A, Zhao H et al (2011) Density contrast sedimentation velocity for the determination of protein partial-specific volumes. PLoS One 6:e26221
29. Zhao H, Ghirlando R, Piszczek G et al (2013) Recorded scan times can limit the accuracy of sedimentation coefficients in analytical ultracentrifugation. Anal Biochem 437:104–108
30. Ghirlando R, Balbo A, Piszczek G et al (2013) Improving the thermal, radial, and temporal accuracy of the analytical ultracentrifuge through external references. Anal Biochem 440:81–95

Chapter 11

Mass Spectrometry

Masanori Noda, Kiichi Fukui, and Susumu Uchiyama

Abstract

The first mass spectrometry device was made in 1912 by J. J. Thomson. Until the early 1900s, the analysis of small molecules was mainly performed using electronic ionization (EI) and chemical ionization (CI) methods. However, in 1969 Beckey and others developed the electric field desorption (FD) method to analyze the molecular weight distribution of high molecular weight compounds. In subsequent years, electrospray ionization (ESI) and the matrix-assisted laser desorption/ionization (MALDI) methods have been widely used for the analysis of high molecular weight compounds such as proteins and sugars. Significant progress has been made in genomic analysis. For the proteome to be analyzed (e.g., all proteins can be included in an individual sample), mass spectrometry is needed. Recently, mass spectrometry has played an important role in the analysis of protein complexes, particularly in determining the stoichiometry of protein within complexes as well as proteomic analysis. Importantly, the mass measurement of molecular complexes composed of proteins or of proteins and low molecular weight compounds through non-covalent interactions has been enabled, accelerating the understanding of biological phenomena and drug development. In this chapter, we describe the use of mass spectrometry for the analysis of non-covalent protein–protein interactions and protein–low molecular weight compound complexes. We also discuss the validation of the molecular masses of proteins within protein complexes by using mass spectrometry.

Keywords Mass spectrometry (MS), MS measurement under non-denaturing conditions, Protein–-protein interaction

1 Introduction

In the 1990s, many functions of unknown proteins were discovered via genome and proteome analyses [1–3]. Extensive functional analysis of novel proteins was performed; however, many proteins did not appear to have a distinct in vivo function. For this reason, almost all proteins are thought to exist as complexes to be functional. Therefore, understanding the components of protein complexes and their stoichiometry is important. To study protein complexes, X-ray crystallography, nuclear magnetic resonance (NMR), and analytical ultracentrifugation methods have been used. X-ray crystallography and NMR methods can elucidate the composite structure of proteins at an atomic level [4, 5]. However,

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks,
DOI 10.1007/978-4-431-56030-2_11, © Springer Japan 2016

a large amount of proteins and protein crystals of high quality are inevitably required for X-ray crystallography. Stable isotope labeling is usually performed in NMR studies of proteins, but the limitation of molecular weights for acquiring resonance spectra is challenging to overcome. Analytical ultracentrifugation can observe protein complexes in solution and help in ascertaining equilibrium constants; however, studying complicated complexes requires extensive effort [6]. Therefore, researchers do not currently study protein complexes using a single technique; instead, they combine multiple methods to fully examine protein complexes. One of these additional techniques used to characterize protein complexes is mass spectrometry (MS).

The field of MS is relatively new. The first MS was developed in the early 1900s, and the existence of the isotope, which did not have radioactivity, was discovered by MS. This was a significant discovery in the field of chemistry, enabling most of the known isotopes to be discovered within the next 20 years. However, the measurable mass range of MS was still restricted to small molecular weight molecules. It was not until the 1990s that molecular weights of approximately 10,000 Da were measurable by MS. In the 1990s, the electron spray ionization (ESI) and the matrix-assisted laser desorption/ionization (MALDI) methods were developed, enabling the measurement of material with larger molecular weights [7, 8]. In the field of biology, MS has been used for proteomic, posttranslational, and metabolomic analyses. All proteins of various species have been identified comprehensively via proteomic analysis, and their posttranslational modifications within a cell have been elucidated using MS of the proteins involved in corresponding metabolic pathways. These techniques have been established as standard techniques that are performed for various species. In recent years, more examples using MS to analyze protein complexes have emerged [9]. Proteins need to be analyzed under non-denaturing conditions to observe the entire interacting protein complex, which helps to reveal the function of the complex. We can precisely determine the stoichiometry of the components of a complex if the protein complex can be ionized while maintaining its non-covalent bond, based on the high precision of mass measurements. In this chapter, ionization methods used for MS measurements of proteins will be introduced, and their limitations such as solvent conditions will be discussed. After that, an example of an MS measurement of a non-denatured protein will also be discussed.

2 Ionization for MS

The following ionization methods are used to ionize target molecules in MS measurements:

- EI: Electron Ionization

  A method of ionization that utilizes thermal electrons released from a heated filament to collide with sample or an atom.
- CI: Chemical Ionization

  An ionization method that is based on an electric charge exchange between sample molecules and gas molecules of a type of gas (methane) that is introduced after being ionized by the EI method.
- FD: Field Desorption

  An ionization method wherein a sample is applied to an electrode and exposed to heat via a high electric field near the electrode tip. This method uses voltage for ionization by tunneling.
- FAB: Fast Atom Bombardment

  An ionization method that involves mixing a sample with a matrix (glycerin), and then atoms belonging to a neutral element (Ar or Xe) colliding with the sample at high speeds.
- APCI: Atmospheric Pressure Chemical Ionization

  An ionization method that produces ions by using the electric discharge of a corona needle ionizer after vaporizing a sample solution by forcibly heating it to high temperatures of 400–500 °C. This method is in fact CI performed under atmospheric pressure, and the solvent that is vaporized plays a role as the reactant gas.

However, most of these techniques are not used for the ionization of proteins. Proteins are susceptible to fragmentation using high-energy ionization methods similar to those described earlier. For the analysis of proteins, two soft ionization methods are used.

2.1 MALDI: Matrix-Assisted Laser Desorption/Ionization

MALDI is an abbreviated designation of matrix-assisted laser desorption/ionization. The exact definition of the matrix is undecided, as it generally absorbs a laser beam, which promotes the ionization of a sample. Suitable matrices for nitrogen lasers that are frequently used in MALDI consist of a benzene frame. The benzene frame absorbs a laser beam, and a carboxyl group becomes the proton supplier. The hydrophilicity and hydrophobicity of the matrix depend on its position and the number of hydroxyl groups (Fig. 1).

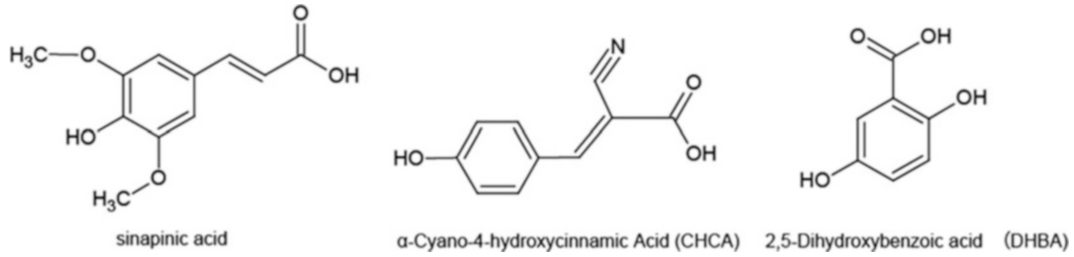

H3C—O
O
OH
HO
H3C
O
sinapinic acid
N
O
HO
OH
α-Cyano-4-hydroxycinnamic Acid (CHCA)
O
OH
OH
HO
2,5-Dihydroxybenzoic acid (DHBA)

Fig. 1 List of commonly used matrices for protein MS measurement

The sample analyzed by MALDI is uniformly mixed with a large quantity of matrix dissolved in the solvent (50 % acetonitrile aqueous with 0.1 % trifluoroacetic acid), and the solvent is vaporized naturally. After that, the minute crystal-like substances composed of the peptides or proteins and matrix are confirmed. It is desirable that the crystal-like substances are uniform as possible to get a better mass spectrum. The matrix absorbs a nitrogen laser beam (wavelength = 337 nm), and the matrix converts it into thermal energy. A small portion of the matrix is heated rapidly (within a few nanoseconds) and is vaporized with the sample. In MALDI, the delivery of the proton takes place between the matrix and samples, forming a protonation/detachment ion. Most ions generated are univalent, but polyvalent ions are produced for high molecular weight compounds, in which electric charges can easily be combined.

2.2 ESI: Electron Spray Ionization

ESI uses electrical energy to assist in the transfer of ions from solution into the gas phase. Ionic species in solution can be analyzed by ESI-MS with increased sensitivity. Neutral compounds can also be converted into an ionic form in solution or gas phase by protonation or cationization.

The transfer of ionic species from solution into the gas phase by ESI involves three steps: (1) the dispersal of a spray of charge droplets, (2) solvent evaporation, and (3) ion ejection from the highly charged droplets (Fig. 2). A mist of highly charged droplets is generated with the same polarity as that of the capillary voltage. The application of a nebulized gas, which surrounds the eluted sample solution, increases the sample flow rate. The charged droplets, generated at the exit of the electrospray tip, pass down pressure and potential gradients toward the analyzer region of the mass spectrometer. With the aid of an elevated ESI-source temperature and/or another stream of nitrogen drying gas, the charged droplets are continuously reduced in size by the evaporation of solvent, leading to an increase in the surface charge density and a decrease in the size of the droplet. Finally, the electric field strength

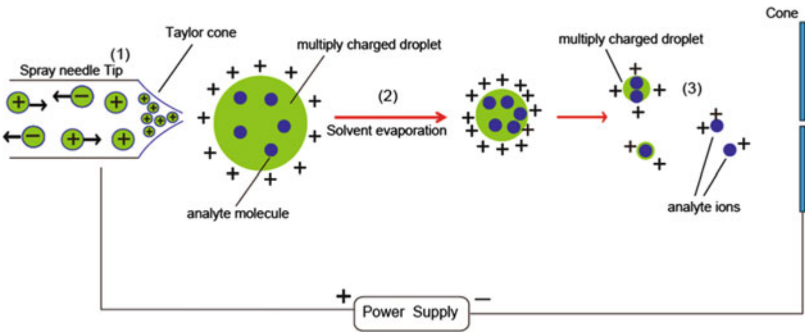


Fig. 2 The principle behind the ESI method

within the charged droplet reaches a critical point at which it is kinetically and energetically feasible for ions at the surface of the droplets to be ejected into the gas phase. The emitted ions are sampled by a sampling cone and are then accelerated into the analyzer for subsequent molecular mass and ion intensity analyses.

3 Sample Condition Suitable for MS Measurements

When MS measurements are performed, it is necessary to pay attention to the solvent (especially the solvent's purity) used. A high-purity MS-grade solvent is guaranteed, but MS should be used to verify the purity of the solvent at least once. For example, organic solvents, such as acetonitrile, consist of a mixture of high molecular weight compounds that are from the caps of the solvent container.

3.1 MALDI

MALDI is thought to have a wider molecular weight range than that of ESI. Generally, nonvolatile salts, such as surfactants and buffers, are not removed when proteins and peptides are purified and measured by MS. In this case, reversed-phase chromatography or dialysis is used to remove these salts. However, the salt removal depends on the samples. In this case, one considers whether the influence of the nonvolatile salts can be suppressed by altering the concentration of the sample.

3.2 ESI

When the ESI method is used for highly sensitive and precise MS of proteins, MS measurements should be performed under denaturation conditions, which may include increased acidity and organic solvent. Analysis of proteins under denaturation conditions can only be used to estimate the mass of molecules containing only covalent bonds. Because the proteins are denatured, many protons are added to them, and a wide range of electric charge distribution is observed.

When ESI-MS is performed on proteins under non-denaturation conditions, a sample solution should be prepared with aqueous solutions that do not contain nonvolatile salts. Because nonvolatile salts prevent the ionization of proteins, adding NaCl and glycerol must be avoided. Therefore, protein solutions have to be treated with an aqueous solution of ammonium acetate using dialysis or a gel filtration cartridge.

Because the substitution of solvent is necessary for MS measurements of proteins in a non-denatured state—and because MS measurements are performed in a gas phase—a confirmation is necessary on whether complexation occurs, which is observed by MS, being the same as that observed in the actual solution. Other methods, such as analytical ultracentrifugation (AUC) and size exclusion chromatography with a multi-angle light scattering (SEC-MALS) detector, can be used to confirm complexation. However, in recent years, complexation of proteins in their native state has been verified by ESI-MS, and the results are thought to be roughly the same in solution from a qualitative point of view.

In the next chapter, we will describe the MS measurement method used to analyze non-denatured proteins.

4 MS Measurement Under Non-denaturing Conditions

When non-denatured protein complexes are measured by ESI-MS, ions are accelerated and channelized into a collision cell containing an inert gas, such as argon, resulting in solvent removal and the release of an intact ionized protein complex. After the second ion phase passes through, packets of ions are transmitted by the pusher for separation in the time-of-flight mass analyzer (Fig. 3). Finally, mass spectra reveal that almost all proteins maintain their interactions with other proteins or with low molecular weight compounds. Higher collision energy results in the dissociation of protein complexes, and higher energy can result in structural local

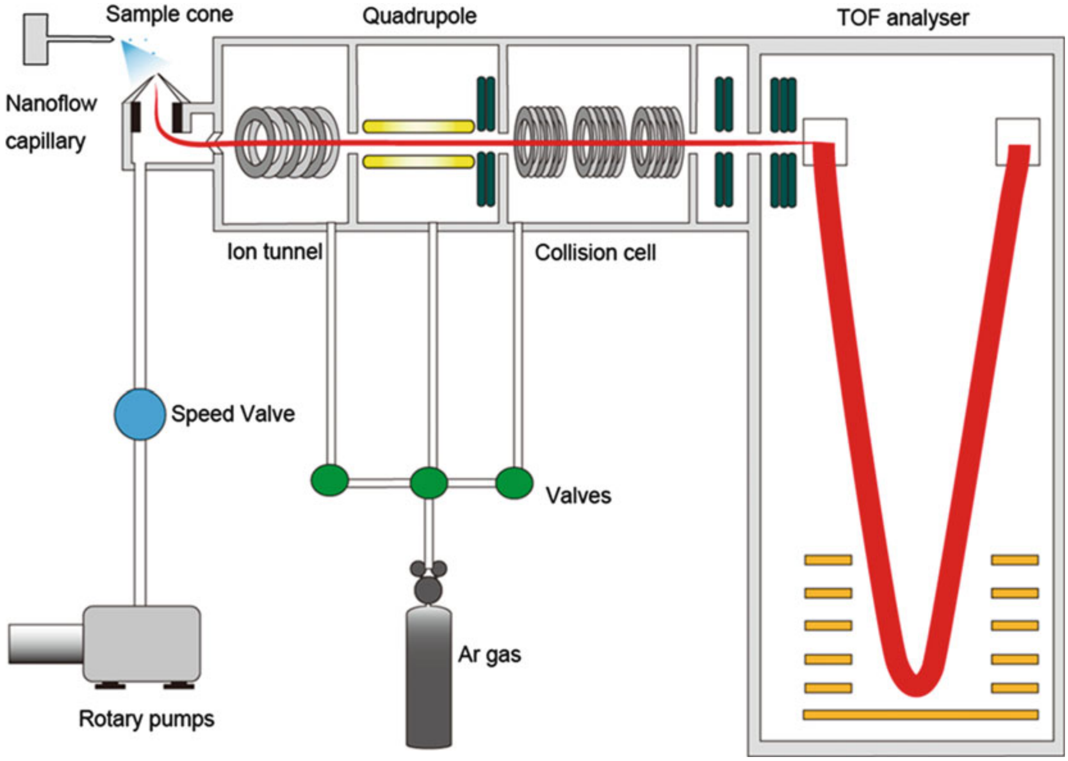


Fig. 3 Schematic representation of a Q-TOF mass spectrometer used for mass measurements of intact protein complexes. Protein complexes undergo nano-ESI through an applied capillary voltage, and then ionized species are desolvated before entering the MS. The ions enter the source, where the pressure is raised to increase the passing efficiency of large protein complexes. Next, ions pass through a quadrupole before being accelerated into a collision cell filled with inert gas. Activated ions are transferred through the TOF section

unfolding, as well as the removal of unfolded proteins from the protein complex.

4.1 Preparation for MS Measurement

4.1.1 Sample Concentration and Volume

The final concentration of complex needed for ESI measurements ranges from 1 to 20 μM. The volume requirements are 1–2 μl per nanoflow capillary. With a 1 μM concentration, a minimum concentration of approximately 5 pmol of protein complexes will be available. Higher concentrations and volumes are desirable because they provide sufficient material for optimization.

4.1.2 Buffer Exchange Method

The selection of a buffer exchange protocol is based primarily on the concentration of the complex.

Preparing a protein solution with a concentration of over 5 μM is essential. For concentrations greater than or equal to 5 μM of complex, microcentrifuge-type gel filtration columns with load volumes of 20–70 μl are commonly used. Complex-containing solutions are loaded after pre-equilibrating the column with an ammonium acetate solution at the required concentration and pH. Depending on the composition of the buffer in the original solution (containing glycerol or detergents), it may be necessary to pass the complex-containing solution two or three times through the column, although this decreases the overall recovery of the complex.

4.1.3 Preparation of ESI Nanocapillary

The internal diameter of a capillary is very important for MS measurement under non-denaturation conditions, which under ideal conditions (no back pressure) determines the flow rate. In our laboratory, an in-house capillary is used, but capillaries purchased from commercial manufacturers can also be used. Before MS measurements are taken, the tip of the capillary is reduced to an appropriate length, and 1–2 μl of sample solution is required to load the capillary. The capillary tip is positioned 1–10 mm from the cone orifice. Short distances are usually optimal at lower capillary voltages.

4.2 Parameters for MS Measurement

To achieve optimal settings, mass spectrometer parameters are tuned for maximal desolvation while attempting to minimize protein activation. The optimization of the following parameters is essential: collision voltage, cone voltage, collision gas pressure, and source pressure.

4.2.1 Source Pressure

Protein complexes generally require an increase in pressure in the transfer region between the source and analyzer. The simplest way to increase the pressure is by using a SYNAPT HDMS (Waters) to reduce the conductance of the source vacuum line to the roughing pump by partially closing the isolation valve (speed valve). Depending on the vacuum system of an instrument, it may be necessary to install or change the position of the isolation valve to allow the pressure to be varied in the source/transfer region.

4.2.2 ESI Voltage

A number of factors determine the quality of the mass spectra, based on their effects on ESI, including capillary internal diameter capillary voltage, back pressure, the position of the capillary relative to the cone, and the flow rate of the desolvation gas. Optimized ESI parameters are interdependent as well as dependent on the specific complex in solution. In general, the capillary voltage is optimal between 1000 and 1800 V, and the flow of desolvation gas between 80 and 150 h^{-1}. A back pressure (0–2 bar) can be applied to initiate the flow rate of the desolvation gas and then be reduced once the ESI is stable. However, high-quality spectra are usually obtained without any back pressure. Under such conditions, the spray may not be visible with a magnifying lens. However, in some cases, a stable spray cannot be maintained without back pressure and a high capillary voltage (1800–2000 V). In addition, it may be necessary to run the sample solution for several minutes before a stable signal is obtained from protein complexes.

4.2.3 Collision Energy Setting

Initially, complex-containing solutions are electrosprayed with intermediate voltages and the pressure in the source/transfer region is increased until charge states from the complex are detected. The charge states may not be resolved initially, and often a broad peak is distributed over a thousand or more m/z units. Further optimization depends on the configuration of the instrument. A general approach is to vary the cone and extractor voltages at several fixed back pressures. Similar spectra can be obtained using different combinations of voltages and pressures, and a trial-and-error approach is needed because optimal conditions will vary for each complex. For a quadrupole-TOF tandem mass spectrometry (Q-TOF) instrument, the collision cell pressure and voltage are additional factors to be considered. Increasing these two parameters can often improve the spectra of larger complexes (molecular weight greater than 300 kDa), in addition to increasing the extent of desolvation for complexes that are poorly resolved. For low-intensity, unresolved complexes, MS/MS with a wide isolation window can improve the transmission over a limited m/z range, increasing the collision cell voltage and pressure, which may allow the resolution of charge states. However, higher collision cell voltages can cause local unfolding. Therefore, similar to the cone and extraction voltages and back pressures, trial-and-error approaches need to be used to optimize these parameter settings.

4.3 Example 1: Protein Self-assembly

The nucleosome is a fundamental assembly of chromatin fibers in higher eukaryotes, consisting of DNA and four distinct histones, H2A, H2B, H3, and H4 [10–12]. Nucleosome assembly is not required for translation, but it is required for chromatin replication. During these processes, histones are delivered to naked DNA by proteins known as histone chaperones, which include nucleosome assembly protein 1 (NAP-1). Nucleosome formation from DNA and histones can be achieved in vitro in the presence of NAP-1 [13]. Biochemical or biophysical characterization of NAP-1 has been previously reported using yeast NAP-1 (yNAP-1) or Drosophila NAP-1 (dNAP-1). These studies showed that dNAP-1 and yNAP-1 mainly form dimers. In addition, the physiological ionic strength has been speculated to play a role in the formation of higher oligomers of yNAP-1 [14]. A subsequent study concluded that an equilibrium exists between yNAP-1 dimers, octamers, and hexadecamers based on sedimentation equilibrium analysis [15]. However, neither research on human NAP-1 (hNAP-1) nor elucidating the oligomerization mechanism of yNAP-1 could be performed. Therefore, the self-assembly of hNAP-1 and yNAP-1 was investigated by MS under non-denaturing conditions [16]. Initially, the homogeneity of hNAP-1 and yNAP-1 was assessed by MS under non-denaturing conditions to obtain precise mass information about non-covalently bound complexes coexisting in solution. The mass spectra of hNAP-1 at physiological ionic strength showed two major series of resolved peaks with different charge states in addition to broad, low-intensity unresolved peaks at higher m/z. In the case of yNAP-1, four major series of resolved peaks were detected (Fig. 4). Notably, the MS results of both yNAP-1 and hNAP-1 indicate the existence of monomers with charge states that could be assigned to folded states.

The observed charge states of higher oligomers are assigned to dimers, tetramers, hexamers, octamers, and decamers. Considering that no odd-numbered oligomers of hNAP-1 and yNAP-1 are observed, with the exception of the small peaks assigned to monomers, the results indicate that hNAP-1 and yNAP-1 exist as stable dimers. In addition, a fraction of these dimers is assembled into higher oligomers. The exact masses of unfolded hNAP-1 were obtained by increasing the collision energy. The following two series of charge states were assigned to unfolded hNAP-1 molecules with different masses: full-length (45,885 Da) and truncated hNAP-1 (45,323 Da). Because of the heterogeneity in the primary structure of hNAP-1, hNAP-1 oligomers have several molecular masses and peaks observed for hNAP-1 oligomers are broad. At a higher ionic strength (750 mM ammonium acetate; Fig. 4a), the populations of higher oligomers of hNAP-1 were significantly reduced and dimers were the dominant species.

These results indicate that the primary assembly unit of both hNAP-1 and yNAP-1 is a dimer and higher oligomers are formed at physiological ionic strength. The disruption of higher oligomers at high ionic strength indicates that the association of dimers is stabilized by electrostatic interactions.

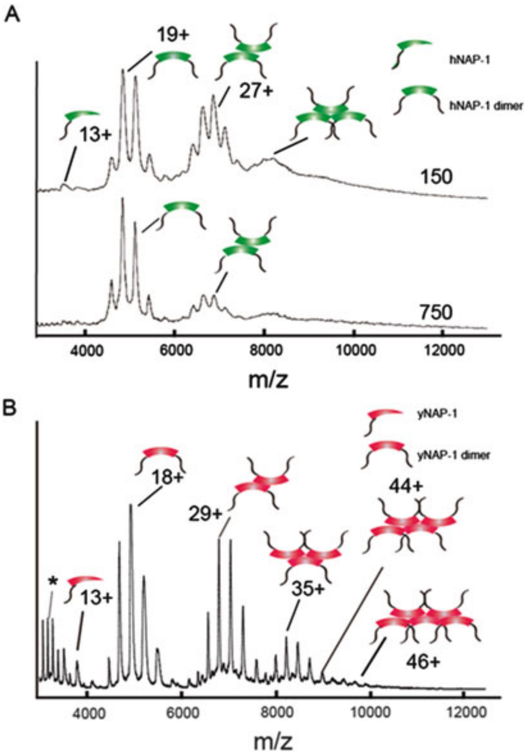


Fig. 4 MS spectra of hNAP-1 and yNAP-1 under non-denaturation conditions. (**a**) MS spectra of hNAP-1 in the presence of 150 and 750 mM ammonium acetate. (**b**) MS spectrum of yNAP-1 in the presence of 150 mM ammonium acetate. The asterisk indicates a yNAP-1 monomer formed via the dissociation of the yNAP-1 dimer

4.4 Example 2: Protein–Protein Heterointeractions, Histone Chaperone–Histone Protein Complexes

The stoichiometry between the binding of NAP-1 and histones was investigated by MS under non-denaturing conditions. Prior to these protein–protein interaction studies, the assembly states of histones in 150 mM ammonium acetate were investigated using MS.

The binding stoichiometry of the hNAP-1 dimer to both histone components was investigated using MS (Fig. 5a). At equimolar ratios of H2A-H2B dimer and hNAP-1 dimer [(hNAP-1)$_2$], a heterotetramer [(hNAP-1)$_2$(H2A-H2B)] was observed. Increasing the molar ratio of H2A-H2B dimer to hNAP-1 dimer removed free (hNAP-1)$_2$, and peaks corresponding to interactions between two H2A-H2B proteins [(hNAP-1)$_2$(H2A-H2B)$_2$] were predominantly observed.

Interactions between the hNAP-1 dimer and the (H3-H4)$_2$ tetramer were investigated in a similar manner. Increasing the amount of (H3-H4)$_2$ tetramer up to a 3:1 ratio of (H3-H4)$_2$ tetramer to hNAP-1 dimer led to formation of the complex (hNAP-1)$_2$(H3-H4)$_2$ as well as hNAP-1 dimer and free (H3-

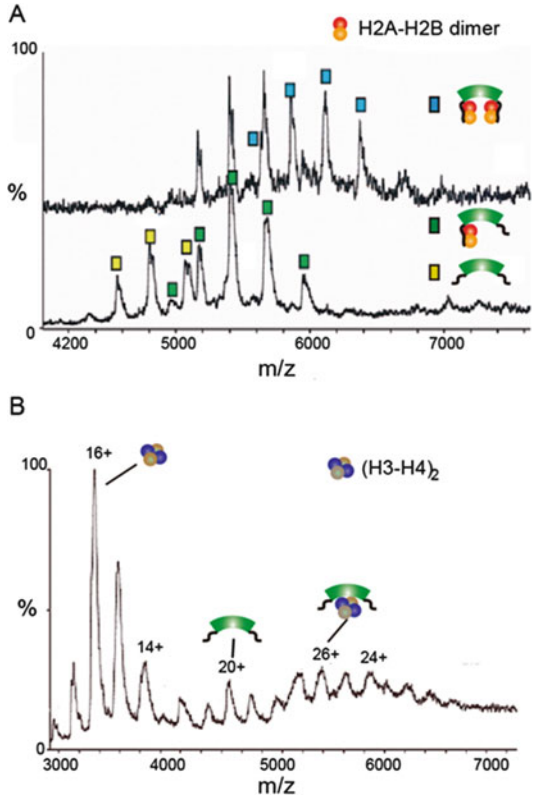


Fig. 5 MS spectra of hNAP-1 and histone complexes under non-denaturing conditions. (**a**) Complex of the hNAP-1 and H2A-H2B dimer at differing molar ratios (hNAP-1/H2A-H2B = 1:1 and 1:2, lower and upper, respectively). (**b**) Complex of hNAP-1 and the (H3-H4)$_2$ tetramer at a 3:1 molar ratio of (H3-H4)$_2$ to hNAP-1

H4)$_2$ tetramer (Fig. 5b). This stoichiometry differed from that observed with the same molar ratio of hNAP-1 dimer to H2A-H2B dimer.

4.5 Protein–Low Molecular Weight Compound Interactions

The differentiation of adipocytes or production of human chorionic gonadotropin (hCG) is mediated through the RXR-PPARγ signaling pathway [17].

To determine whether the complex formation of PPARγ with low molecular weight compounds occurs through covalent or non-covalent interactions, including ionic bonding, MS of PPARγ complexes (MS) with either compound A or compound B under non-denaturing conditions was performed (Fig. 6). The results indicated that compound A and compound B complexes form in a 1:1 ratio (Fig. 6c, f). Because no free PPARγ was detected in these spectra and, even at a highly stringent ionization conditions up to a sample cone voltage of 190 V, peaks corresponding to complexes were not disrupted, the complexes were determined to be highly stable in both cases.

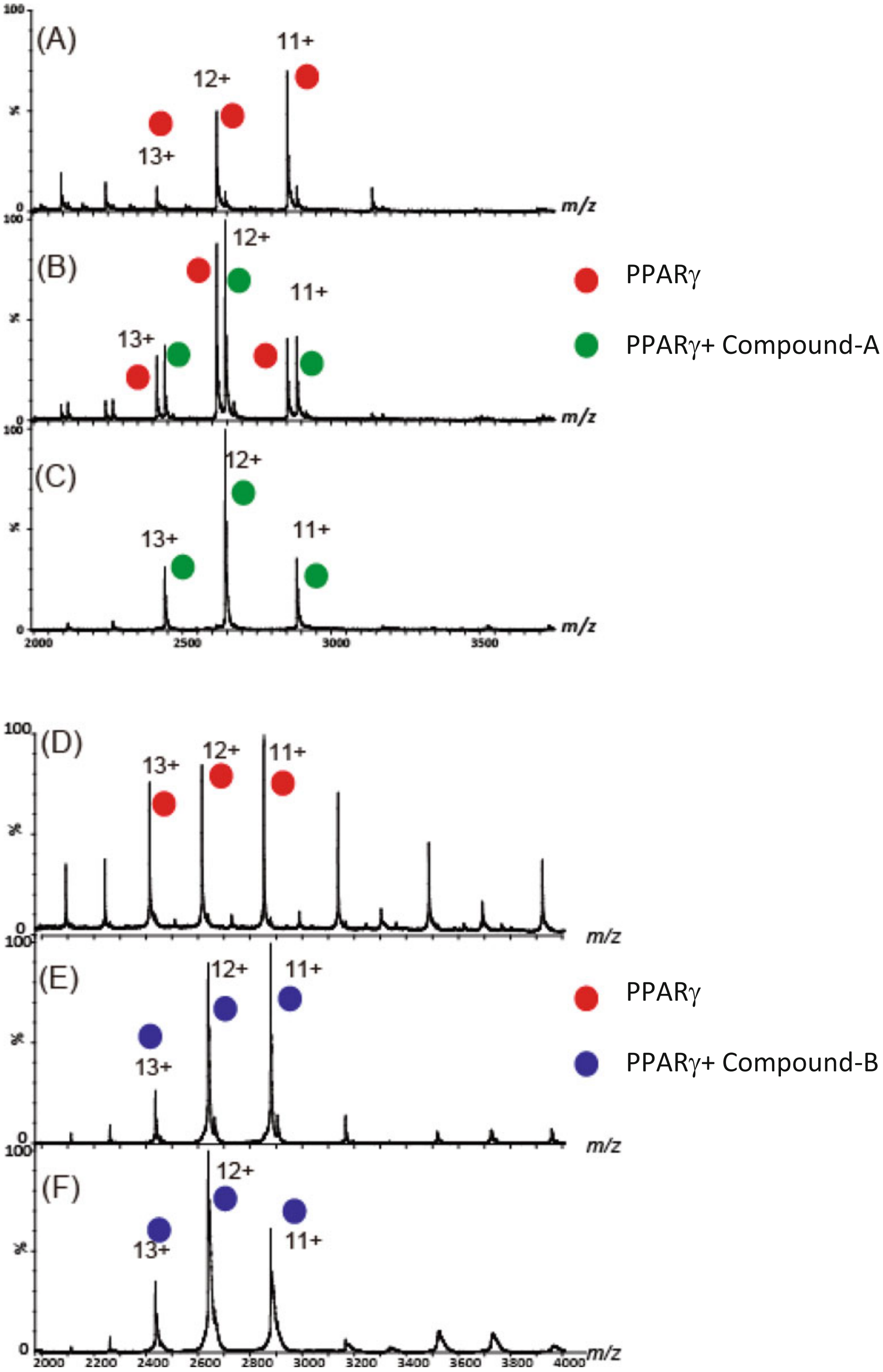


Fig. 6 Mass spectrometry analysis of PPARγ complex with compound A (**a**–**c**) or compound B (**d**–**f**) under non-denaturing conditions. The mass spectrum shows that PPARγ formed a complex with either

The addition of formic acid to mixtures, which was expected to induce the unfolding of PPARγ, resulted in different MS patterns (Figs. 6a-c) even under the same MS conditions used for the aforementioned cases. The new ion series emerged in either partially (Fig. 6b) or fully (Fig. 6a) unfolded PPARγ, which provided a molecular mass of free PPARγ (31,370.6 Da) and indicated the dissociation of compound A or compound B from PPARγ upon the acid-induced unfolding of PPARγ in solution.

References

1. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422:198–207
2. Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc Natl Acad Sci U S A 90:5011–5015
3. Uchiyama S, Kobayashi S, Takata H, Ishihara T, Hori N, Higashi T, Hayashihara K, Sone T, Higo D, Nirasawa T, Takao T, Matsunaga S, Fukui K (2005) Proteome analysis of human metaphase chromosomes. J Biol Chem 280:16994–17004
4. Kosinska Eriksson U, Fischer G, Friemann R, Enkavi G, Tajkhorshid E, Neutze R (2013) Subangstrom resolution X-ray structure details aquaporin-water interactions. Science 340: 1346–1349
5. Enokizono Y, Kumeta H, Funami K, Kumeta H, Funami K, Horiuchi M, Sarmiento J, Yamashita K, Standley DM, Matsumoto M, Seya T, Inagaki F (2013) Structures and interface mapping of the TIR domain-containing adaptor molecules involved in interferon signaling. Proc Natl Acad Sci U S A 110:19908–19913
6. Nogi T, Yasui N, Mihara E, Matsunaga Y, Noda M, Yamashita N, Toyofuku T, Uchiyama S, Goshima Y, Kumanogoh A, Takagi J (2010) Structural basis for semaphorin signalling through the plexin receptor. Nature 467: 1123–1127
7. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246:64–71
8. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yohida T (1988) Protein and polymer analyses up to m/z 100,000 by laser ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 2:151–153
9. Aquilina JA, Benesch JLP, Ding LL, Yaron O, Horwitz J, Robinson CV (2005) Subunit exchange of polydisperse proteins: mass spectrometry reveals consequences of alphaA-crystallin truncation. J Biol Chem 280: 14485–14491
10. Chakravarthy S, Park YJ, Chodaparambil J, Edayathumangalam RS, Luger K (2005) Structure and dynamic properties of nucleosome core particles. FEBS Lett 579:895–898
11. Gaillard PH, Martini EM, Kaufman PD, Stillman B, Moustacchi E, Almouzni G (1996) Chromatin assembly coupled to DNA repair: a new role for chromatin assembly factor I. Cell 86:887–896
12. Hu F, Alcasabas AA, Elledge SJ (2001) Asf1 links Rad53 to control of chromatin assembly. Genes Dev 15:1061–1066
13. Mosammaparast N, Ewart CS, Pemberton LF (2002) A role for nucleosome assembly protein 1 in the nuclear transport of histones H2A and H2B. EMBO J 21:6527–6538
14. McBryant SJ, Peersen OB (2004) Self-association of the yeast nucleosome assembly protein 1. Biochemistry 43:10592–10599
15. Fejes Toth K, Mazurkiewicz J, Rippe K (2005) Association states of nucleosome assembly protein 1 and its complexes with histones. J Biol Chem 280:15690–15699
16. Noda M, Uchiyama S, McKay AR, Morimoto A, Misawa S, Yoshida A, Shimahara H, Takinowaki H, Nakamura S, Kobayashi Y, Matsunaga S, Ohkubo T, Robinson CV, Fukui K (2011) Assembly states of the nucleosome assembly protein 1 (NAP-1) revealed by sedimentation velocity and non-denaturing mass spectrometry. Biochem J 436:101–112

Fig. 6 (continued) compound A (**c**) or compound B (**d–f**) in 1:1 molar ratio. The addition of formic acid to PPARγ–compound B complexes from 1 % (**b**) to 3 % (**a**) resulted in the gradual dissociation of their interactions

17. Kliewer SA, Umesono K, Noonan DJ, Heyman RA, Evans RM (1992) Convergence of 9-cis retinoic acid and peroxisome proliferator signalling pathways through heterodimer formation of their receptors. Nature 358: 771–774

Chapter 12

Frontal Gel Filtration

Tetsuo Ishida

Abstract

In a mixture of protein and small molecules (ligands) at equilibrium, the rates of the binding of ligands to protein molecules and the dissociation of the protein-ligand complexes are equal. Accordingly, the concentrations of free protein, free ligand, and the complex do not change with time, and all of these equilibrium concentrations can be directly determined if we can measure the concentration of free ligands.

Frontal gel filtration is a method to measure the free ligand concentration by isolating a small bit of the solution containing only free ligand molecules from the original mixture without disturbing the binding equilibrium. By using a microcolumn packed with high-resolution gel filtration medium, protein-ligand interactions can be directly examined for small amounts of sample.

Keywords Frontal analysis, Protein-ligand interaction, Gel filtration, Binding curve, Dissociation constant, Binding site, Multiple binding, Serum albumin, Microcolumn, Warfarin

1 Introduction

Proteins bind various kinds of small molecules (ligands) to perform specific functions. For example, enzymes bind substrates as the first step to catalyze relevant reactions. Serum proteins such as serum albumin bind endogenous substances such as fatty acids to transport. Therefore, examination of protein-ligand interactions is important to understand functional mechanisms of proteins. However, it is often difficult to carry out quantitative measurement of protein-ligand interaction using limited amounts of samples by means of ordinary laboratory methods such as equilibrium dialysis.

Frontal gel filtration, or frontal gel chromatography (FGC), is a method to directly measure the free ligand concentration ($[L]_f$) in a protein-ligand mixture. FGC was introduced more than 40 years ago [1–4]. In many important points including the reproducibility of the data obtained and the simplicity of the theory and technique, FGC is superior to equilibrium dialysis, which has been regarded as "gold standard" method to examine protein-ligand interactions. However, FGC requires large sample volumes (more than 10 mL)

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks,
DOI 10.1007/978-4-431-56030-2_12, © Springer Japan 2016

when columns packed with soft gel filtration media such as Sephadex G-25 are used. This disadvantage inhibited a wide range of application of FGC in protein-ligand interacting systems. In recent years, we and other research groups have tried to overcome this disadvantage by using short columns packed with high-resolution gel filtration media [5–7]. It is now possible to carry out FGC using only 100-μL volumes of samples [8].

In this section, the basics of the protein-ligand interaction at equilibrium are first explained. Second, fundamental aspects of gel filtration are illustrated to explain common technical terms used in gel filtration. Finally, using theoretical simulation of chromatograms, the theory of FGC is explained.

1.1 Protein-Ligand Interaction at Equilibrium

For simplicity, consider a protein (P) has a single binding site for a low-molecular-weight ligand (L). In a mixture of the protein and ligand molecules, the following binding (or dissociation) equilibrium is rapidly established:

$$\mathrm{PL} \rightleftarrows \mathrm{P} + \mathrm{L} \tag{1}$$

where PL denotes the protein-ligand complex. If we can determine the equilibrium concentration of free ligand, $[\mathrm{L}]_\mathrm{f}$, of this mixture without disturbing the original total concentrations of the protein and ligand, $[\mathrm{P}]_\mathrm{t}$ and $[\mathrm{L}]_\mathrm{t}$, respectively, then the equilibrium constant (dissociation constant), K_d, and the average number of bound ligand per protein, r, can be calculated using the following relationships:

$$K_\mathrm{d} = \left([\mathrm{P}]_\mathrm{t}[\mathrm{PL}]\right)[\mathrm{L}]_\mathrm{f}/[\mathrm{PL}] \tag{2}$$

$$r = [\mathrm{L}]_\mathrm{b}/[\mathrm{P}]_\mathrm{t} \tag{3}$$

where [PL] and $[\mathrm{L}]_\mathrm{b}$ are the equilibrium concentrations of the protein-ligand complex and protein-bound ligand, respectively, and in this simple case given by the following equation,

$$[\mathrm{PL}] = [\mathrm{L}]_\mathrm{b} = [\mathrm{L}]_\mathrm{t} - [\mathrm{L}]_\mathrm{f} \tag{4}$$

Thus, in principle, once we can determine $[\mathrm{L}]_\mathrm{f}$ of a given protein-ligand mixture, we can determine directly not only K_d (association constant is the inverse of K_d), but also r, the saturation level of the protein binding site.

The simple hyperbolic binding according to 1:1 binding stoichiometry (Eq. 1) is only one possible model proposed for a given protein-ligand interaction and must be verified by experiments. For this purpose, we need to prepare dozens of mixtures containing varying concentrations of the protein and ligand and measure the respective $[\mathrm{L}]_\mathrm{f}$ values. Then, the obtained binding curve, r versus $[\mathrm{L}]_\mathrm{f}$ plot, is examined to be fit for the theoretical relationship:

$$r = [\mathrm{L}]_\mathrm{f}/\left(K_\mathrm{d} + [\mathrm{L}]_\mathrm{f}\right) \tag{5}$$

As the above discussions clearly show, measuring $[\mathrm{L}]_\mathrm{f}$ is a direct method to examine protein-ligand interactions. Equilibrium dialysis and ultrafiltration are two popular methods to measure $[\mathrm{L}]_\mathrm{f}$. However, sample dilution occurs during dialysis, whereas sample concentration occurs during ultrafiltration. In addition to these drawbacks, nonspecific binding to semipermeable membranes is inevitable. Unlike these two methods, FGC does not disturb the original mixture, and in this sense FGC is the most reliable method.

1.2 Fundamental Aspects of Gel Filtration

Figure 1a illustrates a spherical gel particle with numerous pores (100–150 Å). The pores are filled with buffer solutions (the blue region in Fig. 1a), and small ligand molecules diffuse freely into, out of, and within the pores. When gel particles with an appropriate pore size are selected, both protein molecules and protein-ligand complexes cannot enter the pores due to steric hindrance. The volume of the solid material which forms the gel particle (the black region in Fig. 1a) is inaccessible to all solutes.

Figure 1b illustrates a column (typically, 1 mm in internal diameter, 3–10 cm in length) packed with the gel particles depicted in Fig. 1a. The space within the column is functionally divided into three distinct parts. The first part is the space in between the particles, and buffer in this part is referred to the moving phase (the white area in Fig. 1b, c). When the column is connected to a pump to force buffer into the column at a constant flow rate, every solute (ligand, protein, and their complex) present in the moving phase migrates from the inlet of the column to the outlet at the same speed as the flow of buffer.

The second part is the inside of the pores, and buffer in this part is referred to the stationary phase (the blue area in Fig. 1c) because molecules in the pores do not migrate along the column. The last part is the solid materials (the black area in Fig. 1b, c), into which both solvent and solutes cannot penetrate by physical hindrances.

The total volume of the moving phase is referred to the void volume (V_0 μL), and that of the stationary phase is referred to the internal volume (V_i μL). Using the column length (L mm), the cross-sectional area of the moving and stationary phase (S_0 and S_i mm^2, respectively, Fig. 1c) is defined by the following equations:

$$S_0 = V_0/L \tag{6}$$

$$S_\mathrm{i} = V_\mathrm{i}/L \tag{7}$$

If buffer flows into the column at a constant flow rate of u μL/min (volumetric flow rate), then the buffer in the moving phase migrates at a constant speed of v mm/min (linear flow rate). To obtain the relation between u and v, consider how long a protein molecule takes to pass through the column. Because the protein

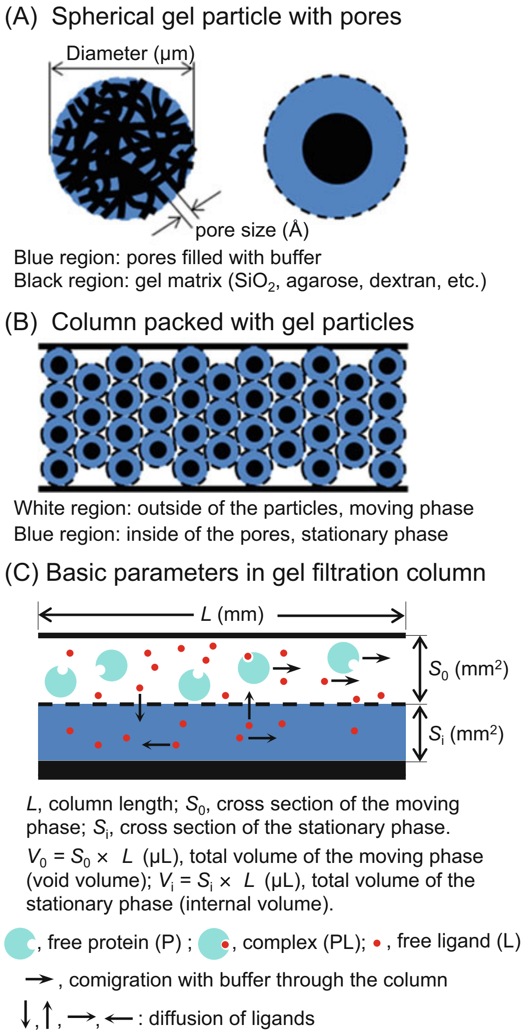

(A) Spherical gel particle with pores
Diameter (μm)
pore size (Å)
Blue region: pores filled with buffer
Black region: gel matrix (SiO_2, agarose, dextran, etc.)
(B) Column packed with gel particles
White region: outside of the particles, moving phase
Blue region: inside of the pores, stationary phase
(C) Basic parameters in gel filtration column
L (mm)
S_0 (mm^2)
S_i (mm^2)
L, column length; S_0, cross section of the moving phase; S_i, cross section of the stationary phase.
$V_0 = S_0 \times L$ (μL), total volume of the moving phase (void volume); $V_i = S_i \times L$ (μL), total volume of the stationary phase (internal volume).
, free protein (P) ; , complex (PL); ● , free ligand (L)
→ , comigration with buffer through the column
↓, ↑, →, ← : diffusion of ligands

Fig. 1 Gel filtration of a protein-ligand interacting system. Diagram of a gel particle (**a**) and a packed gel bed in a column (**b**). (**c**) Common technical terms and important parameters to understand gel filtration chromatograms

molecules stay always in the moving phase, they migrate at the same speed as the flow of buffer, v mm/min. Therefore, it takes L/v min for the protein to pass through the column. During this passage, uL/v μL of buffer elutes from the column. This elution volume is equal to the void volume:

$$V_0 = uL/v \tag{8}$$

From Eq. 8, we find that $v = uL/V_0$. After substituting Eq. 6 in this equation, v is given by the following equation:

$$v = u/S_0 \tag{9}$$

Now, consider that a ligand alone enters the column. Unlike the protein, the ligand molecules do not stay in the moving phase because they frequently leave from the moving phase into the stationary phase, and vice versa, by diffusion. If the rate of this diffusion is much higher than the migration rate of v, then the probability that the ligand is in the moving phase is given by the volume ratio of $V_0/(V_0 + V_i)$. Substituting Eqs. 6 and 7 in this ratio, we find that

$$\text{the probability of the presence of ligand in the moving phase} \\ = S_0/(S_0 + S_i) \tag{10}$$

Because the ligand migrates at the speed of v only in the moving phase, by multiplying v and this probability, the ligand migration speed, v_L mm/min, is obtained:

$$v_L = v\,S_0/(S_0 + S_i) \tag{11}$$

Substituting Eq. 9 into Eq. 11, we find that

$$v_L = u/(S_0 + S_i) \tag{12}$$

The ligand takes L/v_L min to pass through the column. During this passage, uL/v_L μL of buffer elutes from the column. This elution volume is referred to the ligand elution volume, V_L μL. Substituting Eqs. 12, 6, and 7 in the equation $V_L = uL/v_L$, we find that

$$V_L = V_0 + V_i \tag{13}$$

In usual gel filtration, the purpose of the gel chromatography is to separate the protein and ligand, as in the cases of protein desaltation and buffer exchange. Therefore, to attain the complete separation between the protein and ligand, the sample volume injected into the column is maximally about 15 % of the column volume. Figure 2 is a theoretical simulation of such chromatograms expected for the experiments where a small volume (1.3 μL) of the protein solution (Fig. 2a), the ligand solution (Fig. 2b), or the protein-ligand mixture (Fig. 2c) is injected into a column (1 mm in internal diameter, 75 mm in length) which has the V_0 of 22.4 μL and V_i of 21.3 μL (for information on the simulation, refer to Note 1). In the simulations, the total concentration of each solute is supposed to be 10 μmol/L.

As shown in Fig. 2a, protein molecules elute in the void volume, V_0, as a single peak as expected (Eq. 8), and the peak protein concentration is significantly reduced from the original concentration. In the absence of the protein, ligand molecules elute as a single broad peak at the volume of V_0 + V_i as expected (Eq. 13),

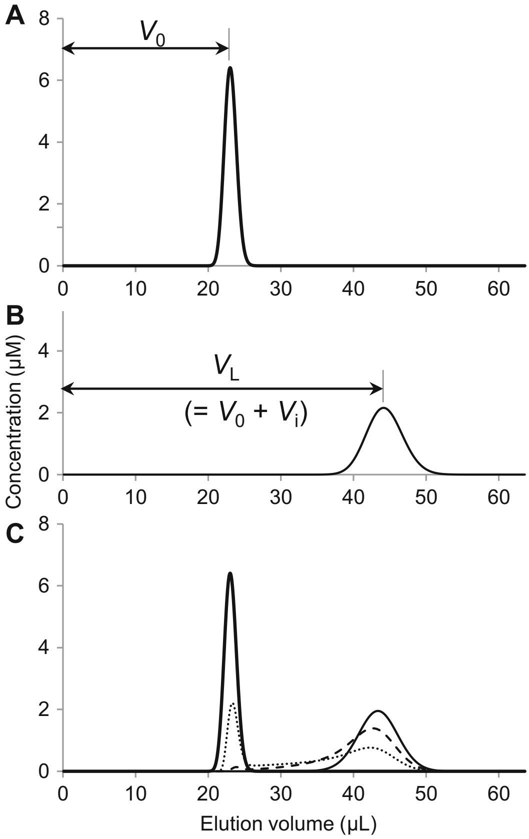


Fig. 2 Theoretical chromatogram of a small volume sample consisting of protein alone (**a**), ligand alone (**b**), and protein-ligand mixture (**c**). Chromatogram simulation was performed using the in-house program listed in Note 1 and the following parameters: the internal diameter and length of the column are 1.0 and 75 mm, respectively. The void and internal volumes, V_0 and V_i, are 22.4 and 21.3 μL, respectively. The sample volume applied to the column is 1.35 μL. The total concentrations of the protein and ligand are both 10 μmol/L. (**c**) The values of the dissociation constant (K_d) used are 5 (*thin line*), 1 (*broken line*), and 0.2 (*dotted line*) μmol/L, respectively

and the peak ligand concentration is about one fifth of the original concentration (Fig. 2b).

Figure 2c shows theoretical chromatograms of the protein-ligand mixture. Coexistence of ligand shows no effect on the elution pattern of the protein: it is identical to the elution pattern depicted in Fig. 2a. In contrast, the elution pattern of the ligand is dependent on the strength of the interaction between the protein and ligand. When the protein-ligand interaction is relatively weak (K_d is larger than about 5 μmol/L), the ligand elutes in a single peak as in Fig. 2b. However, when the protein-ligand interaction is strong (K_d is smaller than about 1 μmol/L), the ligand elution shows two peaks. The first elution peak overlaps the protein elution peak, and the second elution peak is protein-free, appearing slightly before the volume of V_0 + V_i. It should be noted that ligand molecules elute continuously until the second peak appears. These results clearly show that relatively weak protein-ligand interaction is overlooked if you apply small volumes of samples to gel filtration columns.

1.3 Theory of FGC

In FGC, it is essential that elution of the original sample forms a plateau region in the chromatogram. This fundamental requirement is usually satisfied by the injection of a bed volume of the sample (bed volume is the total column volume). In the following discussions, for simplicity, we consider a protein-ligand mixture in which both $[P]_t$ and $[L]_t$ are 10 μmol/L and the dissociation constant is 5 μmol/L. Then the equilibrium concentration of the free ligand, $[L]_f$, is 5 μmol/L.

Figure 3a is a theoretical FGC chromatogram when a protein solution (10 μmol/L, 44.8 μL) is injected into the same gel filtration column (1 mm in diameter, 75 mm in length, bed volume of 58.9 μL) as used in Fig. 2. The elution of the protein starts at the void volume (V_0, 22.4 μL), followed by the elution of the original protein solution as a plateau, and ends at the volume of V_0 + V_S, where V_S is the sample volume injected into the column (44.8 μL). The protein concentration of the plateau region is 10 μmol/L, identical to that of the original sample.

Figure 3b is a theoretical FGC chromatogram when a ligand solution (10 μmol/L, 44.8 μL) is injected into the column. The ligand elution starts at the volume of $V_L(= V_0 + V_i)$, and then the original ligand solution elutes to form a plateau. Finally, the ligand elution ends at the volume of V_L + V_S.

Figure 3c shows a typical FGC chromatogram of the protein-ligand mixture. Because both the protein and protein-ligand complex stay in the moving phase, the protein elutes in the same elution pattern as the sample containing only the protein (Fig. 3a, blue line), namely, the protein elution starts at the void volume, V_0, and ends at the volume of V_0 + V_S, forming a single plateau where the total protein concentration is identical to that of the original mixture. In contrast, the ligand elution shows two plateau regions (thick red line in Fig. 3c). The ligand elution starts at smaller elution volume than the V_L (the elution volume of free ligand, thin broken red line in Fig. 3c), and the first plateau (the β phase in Fig. 3c) appears. Then, the first plateau ends at the elution volume of V_0 + V_S and is immediately followed by the second plateau (the γ phase in Fig. 3c). The second plateau ends at the volume of V_L + V_S, which is identical to the corresponding volume of ligand in the absence of the protein (Fig. 3b). The end of the β phase coincides with the end of the protein elution. In the first plateau region, the original mixture itself elutes.

FGC theory demonstrates that the ligand concentration of the second plateau region ($[L]_\gamma$) is equal to the equilibrium free ligand concentration of the original sample ($[L]_f$).To understand that $[L]_\gamma = [L]_f$, first, consider the ligand migration rate in the β

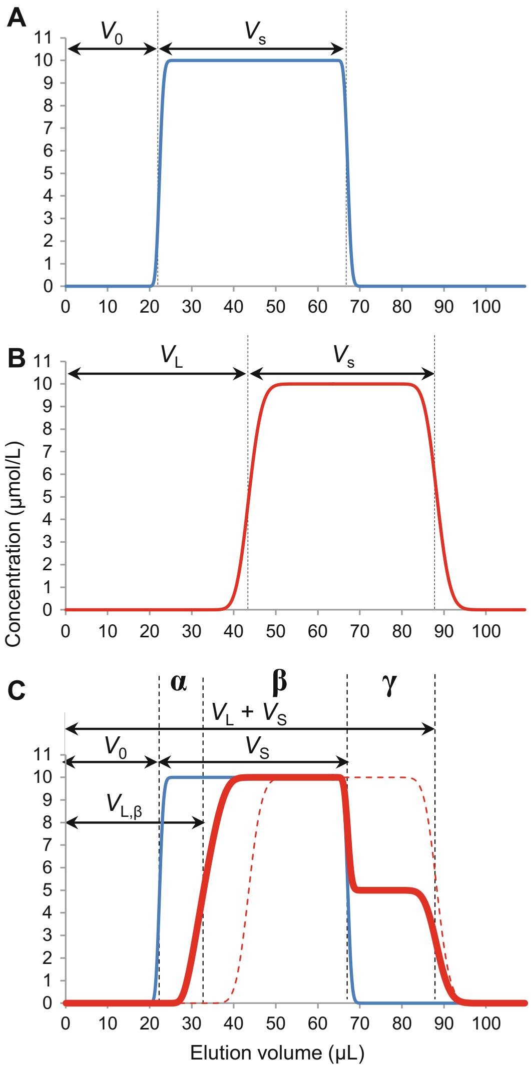


Fig. 3 Theoretical chromatogram of a large volume sample consisting of protein alone (**a**), ligand alone (**b**), and protein-ligand mixture (**c**). Simulation conditions are the same as used in Fig. 2 only except for the sample volume. In the present simulation, the applied sample volume (V_S) was 44.9 μL, about 76 % of the bed volume of the column. (**c**) The K_d value of 5 μmol/L was used. The *blue line* shows the elution of the protein. It is of note that the protein elution pattern is identical to that in (**a**), indicating that the protein-ligand complex migrates at the same speed as free protein. The *red thick line* is the ligand elution, which clearly forms two plateaus. The ligand concentration of the first plateau (β phase) is 10 μmol/L, matched with the total ligand concentration of the original mixture. On the other hand, the ligand concentration of the second plateau (γ phase) is 5.0 μmol/L, matched with that of the free ligand in the original mixture. The *red broken line* is the same chromatogram depicted in (**b**) and is included in this figure as a reference

phase. Ligand molecules eluting in the β phase coexist always with protein molecules in the column. Therefore, ligand molecules take three possible states: free form in the moving phase, protein-bound form in the moving phase, and free form in the stationary phase. The total amount of the ligand in the moving phase is $[\mathrm{L}]_{\mathrm{t}}S_0$ per unit column length, whereas that in the stationary phase is $[\mathrm{L}]_{\mathrm{f}}S_{\mathrm{i}}$. Therefore, the probability that ligand molecules are found in the moving phase is $[\mathrm{L}]_{\mathrm{t}}S_0/([\mathrm{L}]_{\mathrm{t}}S_0+[\mathrm{L}]_{\mathrm{f}}S_{\mathrm{i}})$. By multiplying the migration rate, v $(=u/S_0)$, and this probability, the ligand migration rate in the β phase ($v_{\mathrm{L},\beta}$) is given by the following equation:

$$v_{\mathrm{L},\beta} = u[\mathrm{L}]_{\mathrm{t}}/\left([\mathrm{L}]_{\mathrm{t}}S_0 + [\mathrm{L}]_{\mathrm{f}}S_{\mathrm{i}}\right) \tag{14}$$

Because $[\mathrm{L}]_{\mathrm{f}}$ is smaller than $[\mathrm{L}]_{\mathrm{t}}$, the migration rate, $v_{\mathrm{L},\beta}$, in the β phase is larger than that in the γ phase, v_{L}, ($v_{\mathrm{L}} = u/(S_0 + S_{\mathrm{i}})$, Eq. 2). In the presence of protein, the ligand takes $L/v_{\mathrm{L},\beta}$ min to pass through the column. During this passage, $uL/v_{\mathrm{L},\beta}$ μL of buffer elutes from the column. This corresponds to the volume at which the first ligand plateau (the β phase) starts, $V_{\mathrm{L},\beta}$ (Fig. 3c). Substituting Eq. 14 into $uL/v_{\mathrm{L},\beta}$, we find that

$$\begin{aligned} V_{\mathrm{L},\beta} &= L\left([\mathrm{L}]_{\mathrm{t}}S_0 + [\mathrm{L}]_{\mathrm{f}}S_{\mathrm{i}}\right)/[\mathrm{L}]_{\mathrm{t}} = V_0 + V_{\mathrm{i}}[\mathrm{L}]_{\mathrm{f}}/[\mathrm{L}]_{\mathrm{t}} \\ &= V_0 + (V_{\mathrm{L}} - V_0)[\mathrm{L}]_{\mathrm{f}}/[\mathrm{L}]_{\mathrm{t}} \end{aligned} \tag{15}$$

where Eqs. 6, 7, and 13 are used to derive the final expression. Solving this equation for $[\mathrm{L}]_{\mathrm{f}}$, we find that

$$[\mathrm{L}]_{\mathrm{f}} = \left(V_{\mathrm{L},\beta} - V_0\right)[\mathrm{L}]_{\mathrm{t}}/(V_{\mathrm{L}} - V_0) \tag{16}$$

Now consider conservation of mass. Because all the ligand molecules injected into the column ($[\mathrm{L}]_{\mathrm{t}}V_{\mathrm{S}}$ mol) pass through the column within the two plateaus regions, the area surrounded with the thick red line and the x-axis in Fig. 3c is equal to $[\mathrm{L}]_{\mathrm{t}}V_{\mathrm{S}}$. Considering that the original sample elutes in the first plateau, we obtain the following relationship:

$$\begin{aligned} [\mathrm{L}]_{\mathrm{t}}V_{\mathrm{S}} = {} & [\mathrm{L}]_{\mathrm{t}}\left(V_0 + V_{\mathrm{S}} - V_{\mathrm{L},\beta}\right) \\ & + [\mathrm{L}]_{\gamma}(V_{\mathrm{L}} + V_{\mathrm{S}} - V_0 - V_{\mathrm{S}}) \end{aligned} \tag{17}$$

Solving this equation for $[\mathrm{L}]_{\gamma}$, we find that

$$[\mathrm{L}]_{\gamma} = \left(V_{\mathrm{L},\beta} - V_0\right)[\mathrm{L}]_{\mathrm{t}}/(V_{\mathrm{L}} - V_0) \tag{18}$$

Comparing Eqs. 16 and 18, we can conclude that

$$[\mathrm{L}]_{\gamma} = [\mathrm{L}]_{\mathrm{f}} \tag{19}$$

Equation 19 means that the ligand concentration in the second plateau region is identical to the equilibrium free ligand concentration of the original sample.

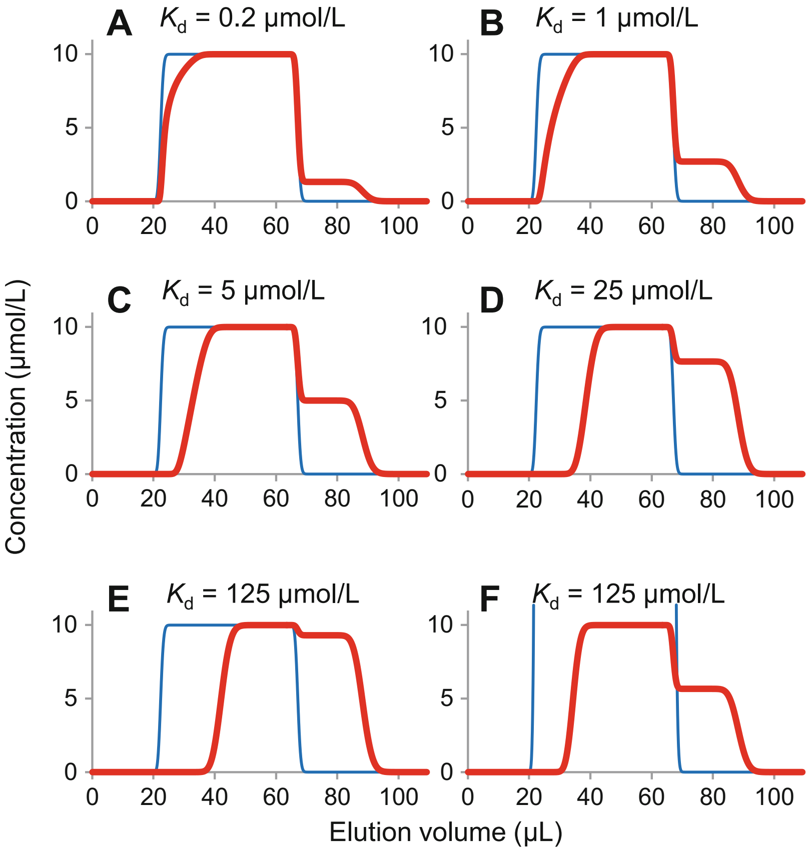


Fig. 4 Dependence of the ligand elution pattern on the strength of the protein-ligand interaction. Simulation conditions are the same as used in Fig. 3c, and only the K_d value increased from 0.2 to 125 μmol/L. The total ligand concentration is 10 μmol/L. The total protein concentration is 10 μmol/L except for (F), in which the protein concentration was 100 μmol/L

Figure 4 shows FGC chromatograms expected for samples containing 10 μmol/L protein and 10 μmol/L ligand as a function of the binding strength (the dissociation constant, K_d) (simulation conditions are the same as in Fig. 3). When the K_d value is in the range of 0.2–25 μmol/L, the second plateau (the γ phase) is clearly formed (Fig. 4a–d). As shown in Fig. 4e, it becomes difficult to discriminate the second plateau from the first plateau as the protein-ligand interaction is very weak (the K_d value is larger than about 100 μmol/L). However, if we increase the protein concentration up to 100 μmol/L, the second plateau (the γ phase) is clearly formed (Fig. 4f). These results indicate that FGC is especially suited for the direct measurement of relatively weak protein-ligand interaction (K_d = 0.1–1,000 μmol/L).

In the above discussions, to explain the essence of FGC, it is supposed that protein-ligand interaction follows 1:1 stoichiometry and that protein and protein-ligand complex are completely excluded from the pores of gel particles. Actually, FGC is effective to examine multiple binding systems only if the following conditions are satisfied [1]. The free protein and all kinds of the protein-ligand complexes (PL_i, the protein molecule that binds i ligand molecules, $i = 1, 2,$) migrate at the same speed [2]. The migration rate of free ligand molecules is sufficiently slow compared to the protein molecules to form clearly the second plateau in the ligand chromatogram (the γ phase in Fig. 3c).

For further detailed information on FGC theory, refer to the *Migration of Interacting Systems* [9]. Refer to the *Biothermodynamics* for detailed information on multiple binding of protein and ligands [10].

2 Materials

2.1 Gel Filtration Media Selection

Commercially available media suitable for FGC are listed in Table 1. When choosing an appropriate medium, consider the following factors:

2.2 Possible Interaction Between the Gel Medium and Protein (or Ligand)

Weak reversible interactions of gel medium and solutes are acceptable, but strong irreversible interactions are unacceptable. Highly acidic or basic substances and aromatic materials may interact with the gel matrix. For example, basic proteins such as histones have a tendency to strongly interact with silica-based media.

Table 1
Commercially available high-resolution gel filtration materials

Name[a]	Particle size (μm)	Pore structure (Å)	Exclusion limit (kDa)	Max. back pressure (Mpa)
TSKgel Super SW2000	4	125	150	12
Superdex peptide	13		20	1.8
Superdex 75	13		100	1.8
Agilent Bio SEC-3	3	100	100	12
Agilent Bio SEC-3	3	150	150	12
COSMOSIL 5Diol-120-II	5	120	100	20

[a] TSKgel gel is a product of TOSOH; Superdex gels are products of GE Healthcare; Agilent Bio SEC gels are products of Agilent; COSMOSIL gel is a product of Nacalai Tesque

2.3 High Resolution Between the Protein and Ligand of Interest

In FGC, it is essential to obtain ligand elution profile containing two plateau regions: the elution of the original sample (the first plateau) and that of free ligand (the second plateau) (Fig. 3). The duration of the first plateau is determined by the sample volume applied, whereas that of the second plateau is determined by the difference in the elution volume between the protein and ligand (V_0 and V_L in Fig. 3). The second plateau must be sufficiently long to determine accurately the equilibrium free ligand concentration, $[L]_f$.

2.4 Operating Pressures Required

As the gel particle size becomes smaller, the back pressure put on to a pump increases. Check the maximum operating pressure of the pump which you want to use. If the maximum pressure is higher than 4 MPa, then the pump can be used to control buffer flow into microcolumns packed with any of the gel filtration media listed in Table 1. Usually, a syringe pump which is used for a liquid handling system has the maximum operating pressure of about 0.5 MPa and can be used only for a column packed with 13 μm Superdex particles.

2.5 Column Size

Commercial prepacked gel filtration columns are usually too large to be used for FGC experiments. In FGC, application of a sufficient volume of sample (a column volume of sample) is essential to ensure the elution of the original sample to form the first plateau region of the chromatogram (Fig. 3). Therefore, columns with 1.0 mm in internal diameter and 35–100 mm in length (bed volume of 27–80 μL) are suited for FGC. It may be possible to order prepacked microcolumns. Alternatively, you can pack yourself gel filtration medium into a microcolumn (stainless steel column, GL Sciences, Tokyo, Japan; quartz column, Kohoku Kougyo, Shiga, Japan). In-house column packing is relatively easy to carry out, and in Note 2 a method is described in detail. In the examples of FGC described below, we used a homemade TSKgel Super SW2000 column with 1.0 mm in internal diameter and 75 mm in length.

2.6 Buffers

Buffer composition should be determined primarily on the basis of the biological properties of the protein molecules of interest. Then, select gel filtration medium compatible with the selected buffer. For example, silica-based media are unusable at extremely acidic or basic buffer conditions. The presence of 0.1–0.15 mol/L NaCl (or equivalent ionic strength) is effective to prevent nonspecific ionic interactions between gel media and protein (or ligands). To examine the interactions of human serum albumin and warfarin, we used HEPES buffer: 50 mmol/L, pH 7.5, $I = 0.15$ (adjusted with NaCl).

2.7 Preparation of Warfarin and Human Serum Albumin

Racemic warfarin and R-(+) warfarin were purchased from Sigma-Aldrich, and S-(−) warfarin was separated from the racemic warfarin using established protocols [11]. Human serum albumin was purified from the Fraction V albumin purchased from Sigma-Aldrich according to the published method [6].

2.8 Instrumentation for Manual FGC

Figure 5 shows a setup of instruments for manual FGC. The pump 1 is used to control buffer flow into a column (1.0 mm in internal diameter and 50–75 mm in length). It is important to use flow rates that allow sufficient time for small molecules to diffuse in and out of the pores of gel particles in order to achieve chromatographic and binding equilibrium of the protein and ligand molecules in the column. For most FGC experiments, the flow rate of 5–10 μL/min is adequate.

It should be noted that the pump 1 must control an accurate flow rate at a constant pressure. Otherwise, even slight variation in the pressure makes a significant periodic noise on FGC chromatograms monitored by a UV-Vis detector. This phenomenon occurs probably because the flow rate is low compared to the volume of the sample loop and the column. In the experiments shown in Figs. 6 and 7, we used an intelligent pump 301M (OmniSeparo-TJ, Hyogo, Japan) equipped with a degassing unit DG661 (GL Sciences, Tokyo, Japan). The degassing unit is essential to avoid the

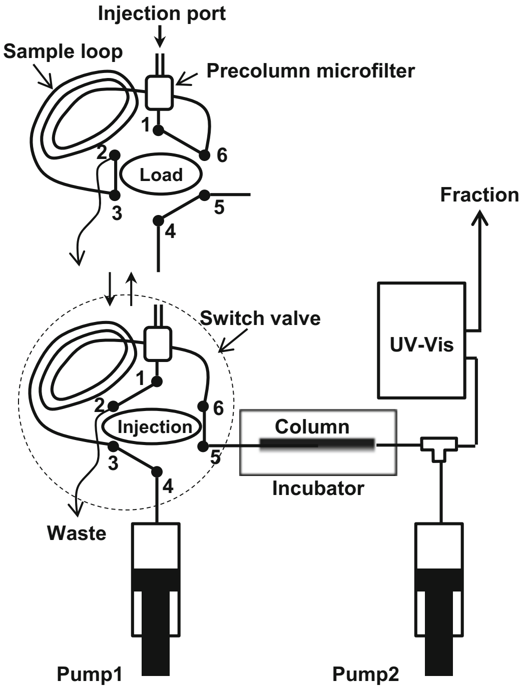


Fig. 5 Schematic diagram of a frontal gel chromatography system

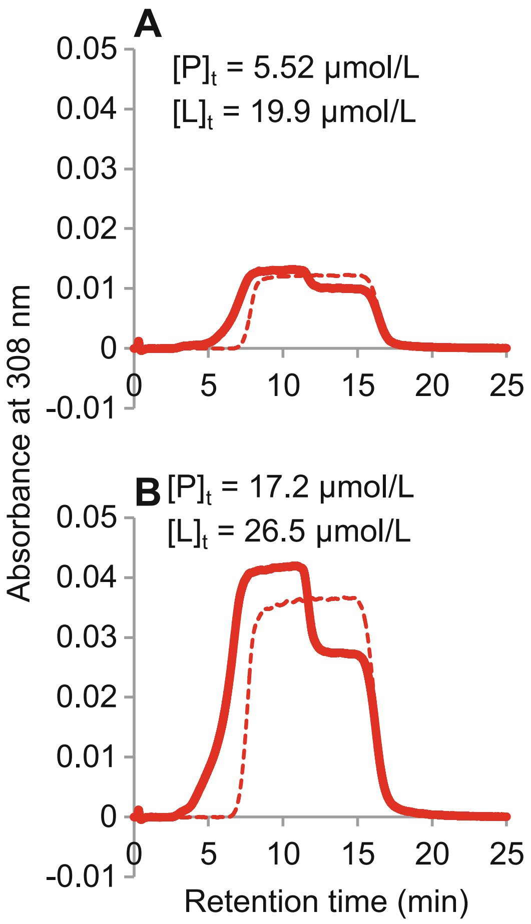


Fig. 6 Typical chromatograms obtained from frontal gel chromatography on a microcolumn (1.0 × 75 mm) packed with TSKgel Super SW2000. A total volume of 300 μL solution containing 5.52 μmol/L human serum albumin and 19.9 *S*-warfarin (**a**) or 17.2 μmol/L and 26.5 μmol/L, respectively (**b**), was prepared in 50 mmol/L HEPES buffer, $I = 0.15$ mol/L, pH 7.5. First, a gas-tight syringe was washed with a 50-μL aliquot of this sample. Second, using this sample-washed syringe, a 200-μL aliquot of the remaining sample was loaded into a 157-μL sample loop. Finally, only a 90 μL of the sample in the loop was applied to the column by returning the valve position from injection to load (see Fig. 5) 9 min after the start of the injection. Column flow rate, 10 μL/min; dilution flow rate, 90 μL/min; temperature, 25 °C; column mobile phase, 50 mmol/L HEPES buffer, $I = 0.15$ mol/L, pH 7.5; dilution solution, Milli-Q water. The *red broken lines* are the chromatograms obtained for sample solutions containing only *S*-warfarin at the corresponding concentrations. The elution of warfarin was monitored by measuring the absorbance at 308 nm. It is noted that human serum albumin shows a significant level of absorption at this wavelength

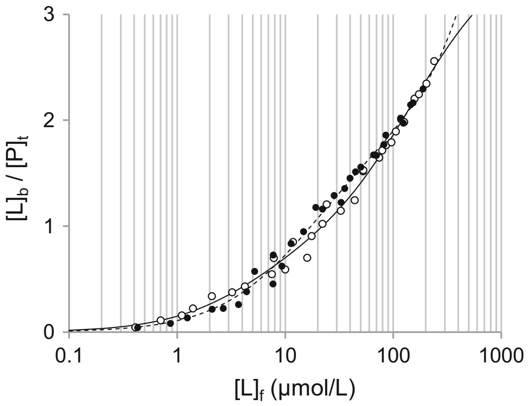


Fig. 7 *S*- and *R*-warfarin binding to human serum albumin. The binding of *S*- and *R*-warfarin to human serum albumin was examined at 25 °C in 50 mmol/L HEPES buffer, $I = 0.15$ mol/L, pH 7.5 by frontal gel chromatography according the same experimental conditions described in Fig. 7. *Open circle*, *S*-warfarin; *filled circle*, *R*-warfarin. The *S*-warfarin binding data suggest that the albumin has one primary binding site ($K_d = 5.0$ μmol/L) and about three secondary binding sites ($K_d = 159$ μmol/L). On the other hand, the R-warfarin binding data indicate that the albumin has two primary binding sites for this enantiomer ($K_d = 16.4$ μmol/L). The *black* and *broken lines* are the best-fit theoretical binding curves

formation of air bubbles in the pump head, the packed column, and the flow cell of the UV-Vis detector.

The pump 2 is used to increase the flow rate of the eluent from the column to 100 μL/min before it enters the UV-Vis detector. Because the flow cell attached to ordinary detectors has a volume of 2–10 μL, without this pre-dilution, the detector signals are complicated due to the eluent dilution in the flow cell. As in the case of the pump 1, the pump 2 must operate at a constant pressure. To satisfy this requirement, the 301M pump was also used for this pre-detector dilution to examine the binding of warfarin to serum albumin (Figs. 6 and 7).

In FGC experiments, application of a column volume of sample is needed. As shown in Fig. 5, a PEEK tube (1/16 in. in outer diameter, 0.50 mm in internal diameter, 80 cm in length) is connected to the position 3 and 6 of a six-position switch valve (e.g., 401, OmniSeparo-TJ). This PEEK tube is used as a 157 μL sample loop.

Injection port is made at the position 1 of the switch valve according the following method. A precolumn microfilter (M-560, Upchurch Scientific) is set to the position 1, and a short Teflon tube (2 cm, 1/16 in. in outer diameter, internal diameter fitted to the outer diameter of the needle of a gas-tight syringe that is used for sample injection, GL Sciences) is connected to this microfilter.

The pump 1 is connected to the position 4 of the switch valve, and the gel filtration column is connected to the position 5. The position 2 is used as an outlet through which washing solutions and extra samples flow out. It is essential that the column temperature is kept strictly at a desired temperature. In the binding experiments shown in Figs. 6 and 7, we put the gel filtration column into an IGLOO-CIL column incubator (OmniSeparo-TJ).

The outlet of the column is connected to the pump 2 and a UV-Vis detector (e.g., 1,200 series, Agilent Technologies) by using a PEEK tee (P-727, Upchurch Scientific). PEEK tubing with internal diameter of 0.1 mm (Upchurch Scientific) is used for these connections.

The raw signals from the UV-Vis detector are collected by a chromatographic integrator (e.g., Smart Chrom, KYA Technologies, Tokyo, Japan). The collected data are converted into text files, and then they are transferred to a personal computer. Data analysis can be carried out on the personal computer using Microsoft Excel and in-house programs written in Visual Basic (Microsoft).

3 Methods

As shown in Fig. 4, in the case of the dissociation constant, K_d, smaller than 100 μmol/L, frontal gel chromatographic analysis can clearly reveal the protein-ligand interaction using samples containing 10 μmol/L of a protein and ligand. Therefore, if nothing is known about the strength of the interaction of the protein of interest and ligands, it is recommended to prepare samples in which the concentration of the protein and ligand is about 10 μmol/L. In the case that the protein-ligand interaction is expected to be very weak (K_d larger than 100 μmol/L), the protein concentration of samples should be higher than 100 μmol/L to unambiguously detect the interaction (Fig. 4f).

3.1 Buffer Preparation

First, it is needed to determine buffer conditions in which the interaction between the protein of interest and ligands is examined. Protein stock solutions, ligand stock solutions, and the protein-ligand mixtures are all prepared using this buffer. This buffer is also used as the running buffer for frontal gel filtration chromatography.

1. Select a buffer and its conditions (pH, ionic strength, etc.) primarily on the basis of the biological activity of the protein of interest.

   If nothing is known about the target protein, start with HEPES buffer: 50 mmol/L, $I = 0.15$ (ionic strength adjusted with NaCl), pH 7.5 (25 °C).
2. Prepare at least 500 mL of the chosen buffer.

Because the same buffer prepared is used for the sample preparation and as the running buffer for FGC, it is better to prepare a sufficient amount of buffer to perform a series of binding experiments. The HEPES buffer (1 L) is prepared as follows: Dissolve 11.9 g HEPES and 7.39 g NaCl in about 800 mL Milli-Q water in a 1-L volumetric flask. Then, add 23.6 g of 1 M NaOH (aqueous solution) into the flask and adjust volume to 1 L with additional Milli-Q water. Confirm the pH of the buffer solution to be 7.5 using a pH meter. If the pH is smaller than 7.5, adjust the pH to 7.5 with 1 M NaOH (less than 3 mL is enough for the adjustment).

3. Filter the buffer through 0.45 μm filters.

3.2 Preparation of Protein Stock Solution

Generally, at least 10 nmol of the purified protein of interest (about 1 mg) is needed for FGC. The purified protein is usually preserved in buffer conditions different from those for the binding assay. If there is no possibility of the partial denaturation of the protein during the storage, a column packed with Sephadex G-25 (about 5 mL bed volume, GE Healthcare) is convenient to exchange buffer. Otherwise, gel filtration on a high-resolution column (about 14 mL bed volume) is recommended to remove the denatured proteins together with buffer exchange. If the protein concentration of the stored sample is less than 0.5 mg/mL, it is recommended to concentrate the sample up to 1 mg/mL or higher concentration using a centrifugal filter device (Ultracel YM-30 or YM-10, Millipore).

1. Equilibrate the column with at least two column volumes of the buffer used for binding assay.
2. Filter 500 μL of the stored protein solution using a centrifugal filter device (0.45 μm pore size)
3. Apply the filtered protein solution to the column, in a volume of 500 μL in the case of Sephadex column and 100 μL in the case of high-resolution column, respectively.
4. Elute the protein with the equilibration buffer by gravity (Sephadex column) or at 0.8–1.0 mL/min (high-resolution column).
5. Collect the desalted intact protein contained in the main part of the elution peak.
6. In the case of high-resolution column, repeat several times the steps 3–5.
7. Concentrate the collected protein to a volume of about 100 μL using a centrifugal filter device (membrane exclusion limit of 10–30 kDa).
8. Determine the protein concentration of the concentrated solution (protein stock solution).

3.3 Preparation of Ligand Stock Solution

It is desirable that a stock solution of the ligand of interest contains at least 100 μmol/L of the ligand in the binding assay buffer. However, it is not uncommon that the ligand shows limited solubility to aqueous buffers. If the solubility of the ligand is unknown, first, dissolve 1 mg of the ligand of interest in 10 mL buffer. If the ligand does not completely dissolve in the buffer, increase the buffer to 100 mL. It should be noted that some ligands take long time to dissolve. For example, 1 mg of crystalline warfarin completely dissolves in 1 ml of HEPES buffer, pH 7.5, but it takes several hours of continuous vigorous mixing to obtain the solution.

3.4 Sample Preparation

Although the sample volume injected into a gel filtration column is 50–100 μL, it is necessary to prepare 300 μL of sample. A 50 μL of the sample is used to wash a gas-tight syringe, which is used to load the sample into a sample loop (157 μL). To fill the sample loop with the original sample as completely as possible, all reaming sample (250 μL) is injected into the loop by using the sample-washed syringe. These careful steps are essential to obtain the elution of the original sample as the first plateau of FGC chromatogram.

In the following explanation, the concentrations of the protein and ligand stock solutions are postulated to be 100 and 150 μmol/L, respectively.

First, prepare three pipets (1,000, 100, and 20 μL) and calibrate the respective pipets to the sampling volumes of 300, 30, and 20 μL by weight, respectively, on the basis of the assumption that specific gravity of water is 1.0. For example, pipette Milli-Q water using the 300-μL adjusted pipet and transfer the water onto a dish on an electronic balance. If the weight of the water is out of the range of 300.0 ± 1.0 mg, then readjust the pipet until the measured weight fits in the range. In the following method, use these calibrated pipets for liquid handling.

1. Add 300 μL of the assay buffer into a 1.5-mL polypropylene tube.
2. Pipette 20 μL of the buffer from the tube and discard it.
3. Pipette 30 μL of the buffer from the tube and discard it.
4. Add 20 μL of the ligand stock solution into the tube.
5. Add 30 μL of the protein stock solution into the tube.
6. Vortex the tube gently to mix the solution.

By performing the above steps, a sample containing 10 μmol/L of the protein and 10 μmol/L of the ligand is obtained. Without performing steps 2 and 4, a sample containing only 10 μmol/L of the protein is prepared. Similarly, without performing steps 3 and 5, a sample containing only 10 μmol/L of the ligand is prepared.

By changing the concentrations of stock solutions and the handling volumes, you can prepare samples containing various concentrations of the protein and ligand according to the above-explained method.

3.5 Frontal Gel Chromatography

First, equilibrate the gel filtration column with the buffer for the binding assay at the desired temperature for about 1–2 h. During the equilibration, confirm that your frontal gel chromatography system is working correctly.

1. Aspirate 50 μL of the sample from the 1.5-mL tube into a gas-tight syringe (250 μL capacity) by pulling up the syringe plunger.
2. Discard all taken sample by pushing down the plunger.
3. Aspirate all remaining sample in the tube into the gas-tight syringe.
4. Confirm that the switch valve position is Load.
5. Inject all aspirated sample into the sample loop through the precolumn microfilter by slowly pushing down the plunger.
6. Change the switch valve position to Inject, and start the recording of the chromatogram.
7. Change the switch valve position to Load immediately after the desired volume of sample injected into the column.
8. Wash the injection port and the sample loop by injecting 200 μL Milli-Q water several times.
9. Before next sample analysis, re-equilibrate the column with one column volume of the buffer after the free ligand elution is completed.

It should be noted that the elution of the ligand is unable to be monitored continuously in the case that the ligand shows no measureable levels of UV-Vis absorption. In that case it is needed to collect the eluent corresponding to the first and second plateau regions into fractions by taking the elution pattern of the protein into consideration. The ligand concentration of the collected fractions can be determined by an appropriate method such as mass spectrometry.

3.6 Data Analysis

First of all, you must examine whether the conditions essential for FGC are all satisfied. As shown in Fig. 3, compare the chromatograms obtained for a set of three samples: sample 1 (protein alone), $[P]_t = 10$ μmol/L, $[L]_t = 0$ μmol/L; sample 2 (ligand alone), $[P]_t = 0$ μmol/L, $[L]_t = 10$ μmol/L; and sample 3, $[P]_t = 10$ μmol/L, $[L]_t = 10$ μmol/L.

1. If the ligand chromatogram of the sample 3 lacks the first plateau, then it is suspected that the ligand interacts strongly

with the gel matrix. In this case, it is recommended to increase the sample volume or shorten the gel filtration column length.

2. In the case that the ligand chromatogram of the sample 3 shows a typical two plateau regions, then confirm that the start of the second plateau coincides with the end of the protein elution and that the protein elution profile of the sample 3 is the same as that of the sample 1. If these conditions are satisfied, then confirm that in sample 2 all ligands entered into the column are eluted in the plateau region, namely, the absence of irreversible interaction between the ligand and the gel matrix. This can be done as follows: collect the eluting ligand and determine the volume of the collected ligand solution, and then measure the absorption spectrum of the solution.
3. If all conditions examined above are satisfied, then the equilibrium free ligand concentration of the sample 3 is obtained by multiplying the ligand concentration of the sample 2 (10 μmol/L) by the ratio of the second plateau height of the sample 3 and that of sample 2.

3.7 Examples of FGC

FGC is the most powerful method to examine multiple binding systems in which protein has more than two binding sites with greatly different affinity to ligands. The interaction between human serum albumin (HSA) and warfarin (anticoagulant) is one of the most investigated examples. Figure 8 shows a crystal structure of HSA-warfarin complex [12, 13] and the chemical structure of warfarin. We examined in detail the interaction of HSA and *S*- and *R*-warfarin by FGC to reveal the stereospecific binding mechanism.

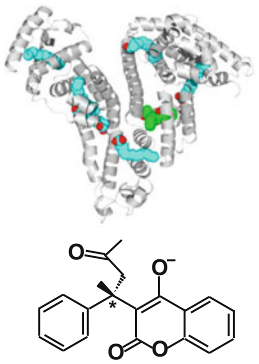


Fig. 8 Structure of human serum albumin-myristate-*S*-warfarin complex (PDB 1HA2). Myristate and *S*-warfarin molecules are shown in a space-filling representation. The carbon atoms of myristate are *colored blue*, and those of *S*-warfarin are *colored green*. The oxygen atoms of the ligands are *colored red*. The figure was prepared using MolFeat (FiatLux, Tokyo, Japan). The chemical structure shows *S*-warfarin, and the asymmetric center is indicated by *asterisk*

We prepared 60 samples containing various concentrations of HSA and *S*- and *R*-warfarin and analyzed them by FGC on a microcolumn (1 × 75 mm) packed with TSKgel Super SW2000. Figure 6 shows typical FGC chromatograms, and the obtained binding curves are shown in Fig. 7. These results strongly suggest that a binding site of HSA, which is distinct from the primary binding site, has strong stereoselectivity for *R*-warfarin, whereas the primary binding site shows only slightly stronger binding to *S*-warfarin.

4 Notes

1. Theoretical Simulation of Frontal Gel Filtration

In FGC, the shape of the ascending and descending parts of the plateau region is not utilized to determine the free ligand concentration. In the theoretical simulation of FGC, neglect of axial diffusion (diffusion of solutes in the moving phase along the longitudinal axis of the column) does not affect the plateau regions, because in the plateau regions there is no axial concentration gradient of solutes. Therefore, strict simulation is possible only on the basis of mass conservation.

A program to simulate FGC of 1:1 binding of protein and ligand is given in Appendix. You can run the program using Excel. In the program, it is postulated that binding equilibrium and the partition of free ligands between the moving and stationary phases are rapid compared to the migration rate of buffer. This program can treat the cases in which protein enters partially into the pores of gel particles, and nonspecific interactions between ligand and the gel matrix occur.

2. Column Packing

You can obtain sufficient amounts of gel particles by removing them from the new or used prepacked columns (bed volume larger than 1 mL). First, detach the end fitting of the column outlet, and then connect the column inlet to a HPLC pump. Wash out the gel particles into a 50-mL polypropylene tube containing about 20 mL of Milli-Q water at the flow rate of 1 mL/min. In the case of used columns, only collect the gel particles contained in the outlet side half of the column.

Column packing can be done by the following method:

1. Connect the one end of a stainless steel column (1.0 × 50–75 mm, outer diameter of 1/8 in., GL Science) with a slurry reservoir (25 mL, CP-25, GL Science) using an attachment (CPA-3, GL Science).
2. Attach an end fitting with frit to the other end of the column.

3. Fill the slurry reservoir with Milli-Q water and cap the reservoir.
4. Connect a HPLC pump to the reservoir and flow Milli-Q water at the rate of 1 mL/min for about 30 min to remove air from the column, the attachment, and the reservoir.
5. Cap the end fitting of the column with a plug, and disconnect the reservoir from the pump.
6. Remove the water in the reservoir using a pipet.
7. Disperse about 150 μL of gel particles into about 25 mL of Milli-Q water.
8. Fill the reservoir with this gel slurry, and cap the reservoir.
9. Connect a HPLC pump to the reservoir and set the maximum pressure limit to 4 MPa. If the maximum operating pressure of the gel particles is lower than 4 MPa, then set the pressure limit to the appropriate pressure.
10. Remove the plug and flow Milli-Q water at the rate of 0.4 mL/ min until the pressure increases to the pressure limit. Then reduce the flow rate to 0.1 mL/min until the pressure increases to the pressure limit. Finally, flow Milli-Q water at the rate 20 μL/min for 1 h.
11. Cap the end fitting of the column with a plug, and disconnect the reservoir from the pump.
12. Remove the remaining solution in the reservoir using a pipet.
13. Remove the column from the reservoir and the attachment.
14. Attach an end fitting with frit to the open end of the column packed with gel particles.

Appendix

FGC simulation program

```
    Sub SimulationFGC1()

    rg = 0.004    'mm gel particle diameter
    lu = 100 * rg  'mm length of the functional unit
    ru = 200 * rg  'mm diameter of the functional unit

    ru = 1

    vu = 3.14 * ru * ru * lu / 4  'uL volume of the functional
unit
    v0u = 0.38 * vu      'uL void volume of the functional unit
    viu = 0.36 * vu      'uL internal volume of the functional
unit
```

```
    vip = 0 * viu  ’uL the volume of the internal space accessible for protein
    ngs = 0   ’uM concentration of the binding site of the gel matrix in MFU
    ng = ngs * vu  ’pmol
    kg = 100    ’uM

    m = 188    ’total number of the functional unit

    B = m * lu   ’the length of the column
    V0 = m * v0u  ’uL void volume of the column
    Vi = m * viu  ’uL internal volume of the column
    Vp = m * vip

    dV = 0.8 * v0u  ’elution volume per unit calculation step

    Vmax = (V0 + Vi) * 2.5   ’total elution volume (including sample volume)
    nv = Int(Vmax / dV)    ’total number of calculation step

    Vs = 0.06 * V0  ’sample volume applied

    ns = Int(Vs / dV)
    C0 = 10    ’uM the total concentration of ligand (L)
    P0 = 10   ’uM the total concentration of acceptor protein (P)
    kd = 0.2    ’uM dissociation constant (one to one stoichiometry)

    d00 = kd + P0 - C0
    d01 = d00 ^ 2 + 4 * kd * C0
    d02 = d01 ^ 0.5
    lf = 2 * kd * C0 / (d00 + d02)

    Cells(2, 15) = nv
    Cells(1, 17) = B
    Cells(2, 17) = ru
    Cells(3, 17) = V0
    Cells(4, 17) = Vi
    Cells(5, 17) = V0 + Vp
    Cells(6, 17) = dV
    Cells(7, 17) = Vmax
    Cells(8, 17) = Vs
    Cells(9, 17) = C0
    Cells(10, 17) = P0
    Cells(11, 17) = kd
    Cells(12, 17) = lf
    Cells(13, 17) = ng * m
    Cells(14, 17) = kg
```

```
v00 = v0u + vip
b0 = v0u + viu

QP = P0 * dV
QL = C0 * dV

    pt = QP / v00

    b1 = b0 * (kd + kg) + pt * v00 + ng - QL
    b2 = b0 * kg * kd + pt * v00 * kg + ng * kd - QL * (kd + kg)
    b3 = -QL * kd * kg

    x01 = C0
    For j1 = 1 To 10
    bunsi = b0 * (x01 ^ 3) + b1 * (x01 ^ 2) + b2 * x01 + b3
    bunbo = 3 * b0 * (x01 ^ 2) + 2 * b1 * x01 + b2
    x02 = x01 - bunsi / bunbo
    x01 = x02
    Next j1
    CL = x01

    CP = pt / (1 + CL / kd)
    CPL = pt * CL / (kd + CL)
    Lt = CL + CPL

Cells(1, 1) = pt
Cells(1, 2) = QP
Cells(1, 3) = Lt
Cells(1, 4) = QL

For j = 2 To m
Cells(j, 1) = 0
Cells(j, 2) = 0
Cells(j, 3) = 0
Cells(j, 4) = 0
Next j

  Cells(1, 10) = dV
  Cells(1, 11) = Cells(m, 1)  'Pt
  Cells(1, 12) = Cells(m, 3)  'Lt

  For j = 1 To m
  Cells(j, 5) = Cells(j, 1)
  Cells(j, 6) = Cells(j, 2)
  Cells(j, 7) = Cells(j, 3)
  Cells(j, 8) = Cells(j, 4)
  Next j
```

```
For i = 2 To ns
   QP = Cells(1, 6) + P0 * dV - Cells(1, 5) * dV
   QL = Cells(1, 8) + C0 * dV - Cells(1, 7) * dV

    pt = QP / v00

      b1 = b0 * (kd + kg) + pt * v00 + ng - QL
      b2 = b0 * kg * kd + pt * v00 * kg + ng * kd - QL * (kd + kg)
      b3 = -QL * kd * kg

      x01 = C0
      For j1 = 1 To 10
      bunsi = b0 * (x01 ^ 3) + b1 * (x01 ^ 2) + b2 * x01 + b3
      bunbo = 3 * b0 * (x01 ^ 2) + 2 * b1 * x01 + b2
      x02 = x01 - bunsi / bunbo
      x01 = x02
      Next j1
      CL = x01

      CP = pt / (1 + CL / kd)
      CPL = pt * CL / (kd + CL)
      Lt = CL + CPL

Cells(1, 1) = pt
Cells(1, 2) = QP
Cells(1, 3) = Lt
Cells(1, 4) = QL

   For j = 2 To m
    QP = Cells(j, 6) + Cells(j - 1, 5) * dV - Cells(j, 5) * dV
    QL = Cells(j, 8) + Cells(j - 1, 7) * dV - Cells(j, 7) * dV

     pt = QP / v00

      b1 = b0 * (kd + kg) + pt * v00 + ng - QL
      b2 = b0 * kg * kd + pt * v00 * kg + ng * kd - QL * (kd + kg)
      b3 = -QL * kd * kg

      x01 = C0
      For j1 = 1 To 10
      bunsi = b0 * (x01 ^ 3) + b1 * (x01 ^ 2) + b2 * x01 + b3
      bunbo = 3 * b0 * (x01 ^ 2) + 2 * b1 * x01 + b2
      x02 = x01 - bunsi / bunbo
      x01 = x02
      Next j1
      CL = x01

      CP = pt / (1 + CL / kd)
      CPL = pt * CL / (kd + CL)
      Lt = CL + CPL
```

```
Cells(j, 1) = pt
Cells(j, 2) = QP
Cells(j, 3) = Lt
Cells(j, 4) = QL

   Next j

  Cells(i, 10) = i * dV
  Cells(i, 11) = Cells(m, 1)  ’Pt
  Cells(i, 12) = Cells(m, 3)  ’Lt

  For j = 1 To m
  Cells(j, 5) = Cells(j, 1)
  Cells(j, 6) = Cells(j, 2)
  Cells(j, 7) = Cells(j, 3)
  Cells(j, 8) = Cells(j, 4)
  Next j

Cells(1, 15) = i

Next i

For i = ns + 1 To nv

  QP = Cells(1, 6) - Cells(1, 5) * dV
  QL = Cells(1, 8) - Cells(1, 7) * dV

   pt = QP / v00

    b1 = b0 * (kd + kg) + pt * v00 + ng - QL
    b2 = b0 * kg * kd + pt * v00 * kg + ng * kd - QL * (kd + kg)
    b3 = -QL * kd * kg

    x01 = C0
    For j1 = 1 To 10
    bunsi = b0 * (x01 ^ 3) + b1 * (x01 ^ 2) + b2 * x01 + b3
    bunbo = 3 * b0 * (x01 ^ 2) + 2 * b1 * x01 + b2
    x02 = x01 - bunsi / bunbo
    x01 = x02
    Next j1
    CL = x01

    CP = pt / (1 + CL / kd)
    CPL = pt * CL / (kd + CL)
    Lt = CL + CPL

Cells(1, 1) = pt
Cells(1, 2) = QP
Cells(1, 3) = Lt
Cells(1, 4) = QL
```

```
  For j = 2 To m

   QP = Cells(j, 6) + Cells(j - 1, 5) * dV - Cells(j, 5) * dV
   QL = Cells(j, 8) + Cells(j - 1, 7) * dV - Cells(j, 7) * dV

    pt = QP / v00

     b1 = b0 * (kd + kg) + pt * v00 + ng - QL
     b2 = b0 * kg * kd + pt * v00 * kg + ng * kd - QL * (kd + kg)
     b3 = -QL * kd * kg

     x01 = C0
     For j1 = 1 To 10
     bunsi = b0 * (x01 ^ 3) + b1 * (x01 ^ 2) + b2 * x01 + b3
     bunbo = 3 * b0 * (x01 ^ 2) + 2 * b1 * x01 + b2
     x02 = x01 - bunsi / bunbo
     x01 = x02
     Next j1
     CL = x01

     CP = pt / (1 + CL / kd)
     CPL = pt * CL / (kd + CL)
     Lt = CL + CPL

Cells(j, 1) = pt
Cells(j, 2) = QP
Cells(j, 3) = Lt
Cells(j, 4) = QL

 Next j

  Cells(i, 10) = i * dV
  Cells(i, 11) = Cells(m, 1)  ’Pt
  Cells(i, 12) = Cells(m, 3)  ’Lt

  For j = 1 To m
  Cells(j, 5) = Cells(j, 1)
  Cells(j, 6) = Cells(j, 2)
  Cells(j, 7) = Cells(j, 3)
  Cells(j, 8) = Cells(j, 4)
  Next j

Cells(1, 15) = i

Next i

End Sub
```

References

1. Nichol LW, Winzor DJ (1964) The determination of equilibrium constants from transport data on rapidly reacting systems of the type A + B = C. J Phys Chem 68:2455–2463
2. Cooper PF, Wood GC (1968) Protein-binding of small molecules: new gel filtration method. J Pharm Pharmacol 20:150S–156S
3. Nichol LW, Jackson WJH et al (1971) The binding of methyl orange to bovine serum albumin studied by frontal analysis in Sephadex chromatography. Arch Biochem Biophys 144: 438–439
4. Keresztes-Nagy S, Mais RF et al (1972) Protein binding methodology: comparison of equilibrium dialysis and frontal analysis chromatography in the study of salicylate binding. Anal Biochem 48:80–89
5. Pinkerton TC, Koeplinger KA (1990) Determination of warfarin-human serum albumin protein binding parameters by an improved Hummel-Dreyer high-performance liquid chromatographic method using internal surface reversed-phase columns. Anal Chem 62: 2114–2122
6. Honjo M, Ishida T et al (1997) Semi-microscale frontal gel chromatography of interacting systems of a protein and small molecules: binding of warfarin, tryptophan, or FMN to albumin, and *o*-nitrophenol to catechol 2,3-dioxygenase. J Biochem 122:258–263
7. Sawada O, Ishida T et al (2001) Frontal gel chromatographic analysis of the interaction of a protein with self-associating ligands: aberrant saturation in the binding of flavins to bovine serum albumin. J Biochem 129:899–907
8. Senda M, Kishigami S et al (2007) Molecular mechanism of the redox-dependent interaction between NADH-dependent ferredoxin reductase and Rieske-type [2Fe-2S] ferredoxin. J Mol Biol 373:382–400
9. Nichol LW, Winzor DJ (1972) Migration of interacting systems. Oxford University Press, London
10. Edsall JT, Gutfreund H (1982) Biothermodynamics: the study of biochemical processes at equilibrium. John Wiley & Sons Ltd, Chichester
11. West BD, Preis S et al (1959) Studies on the 4-hydroxycoumarins XVII The resolution and absolute configuration of warfarin. J Am Chem Soc 83:2676–2679
12. Petitpas I, Bhattacharya AA et al (2001) Crystal structure analysis of warfarin binding to human serum albumin. J Biol Chem 276: 22804–22809
13. Ghuman J, Zunszain PA et al (2005) Structural basis of the drug-binding specificity of human serum albumin. J Mol Biol 353:38–52

Chapter 13

Surface Plasmon Resonance

Yoshihiro Kobashigawa, Natsuki Fukuda, Yusuke Nakahara, and Hiroshi Morioka

Abstract

Almost two decades had passed since the first biosensor based on surface plasmon resonance (SPR) had become commercially available. Among them, the Biacore is the most widely used SPR-based system. More than 10,000 papers, which reported the results obtained using the Biacore (GE Healthcare), had been published until 2015. The most notable progress in the Biacore in this decade is marked reduction of the noise level, which enabled acquisition of the thermodynamic parameters, application to the low molecular weight analytes, observation of the thermodynamic parameters for the activated state, and analysis using further complicate binding model. This chapter aims to provide guidance to users of SPR, with an emphasis on acquiring the thermodynamic parameters for the molecular interaction of two-state binding mechanism, the system exhibiting the interconversion between the transient and the stable complex. No attempt will be made to describe the routine operation and maintenance of the Biacore, as this is comprehensively described elsewhere (Nagata K, Handa H (eds), Real-time analysis of biomolecular interactions: application of BIACORE. Springer, Tokyo, 2000).

Keywords Langmuir, Two state, Induced fit, Thermodynamics, van’t Hoff equation

Abbreviations

AGE	Advanced glycation end product
CDR	Complementarity-determining region
GA	Glycolaldehyde
IgG	Immunoglobulin G
ITC	Isothermal titration calorimeter
RU	Resonance unit
scFv	Single-chain variable fragment
SPR	Surface plasmon resonance
V_H	Variable region of immunoglobulin heavy chain
V_L	Variable region of immunoglobulin light chain

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks, DOI 10.1007/978-4-431-56030-2_13, © Springer Japan 2016

1 Introduction

Surface plasmon resonance (denoted as SPR) is a key principle for the measurement of molecular interaction between two molecules in Biacore system (GE Healthcare). The sensor of SPR-based instruments, Biacore, is composed of a micro flow cell, through which an aqueous solution (denoted as the running buffer) passes under continuous flow (1–100 μL/min). For detection of the intermolecular interaction, one molecule (the ligand) is required to be immobilized onto the sensor surface, to which its binding partner (the analyte) is injected in aqueous solution (sample solution) through the flow cell under continuous flow. The analyte binds to the ligand and accumulated on the surface, which increases mass of the sensor surface within 100 nm approximately. After injective flow, the sample solution is changed to the running buffer, dissociation of the analyte proceeds, and the surface mass of the sensor decreases. This surface mass change is optically measured in real time, and the result plotted as response or resonance units (RUs) versus time (a sensorgram), which enables evaluation of the association and the dissociation rate constant as well as the equilibrium binding constant of the molecular interaction.

In recent years, Biacore system is markedly upgraded. The first notable progress is reduction of the noise level (Table 1), by which ligand concentration fixed on the sensor surface can be markedly reduced without sacrificing the quality of the sensorgram. Low ligand density exhibits two important merits for reducing the following artifact. Firstly, the association rate of the analyte to the surface exceeds the rate of transportation of the analyte to the sensor surface, in case of the high ligand density. In this situation, transportation of the analyte to the sensor surface becomes the rate-limiting step, resulting that the measured association rate constant (k_a) becomes slower than the true k_a. Secondly, dissociated analyte can rebind to the neighboring unoccupied ligand, resulting that the measured dissociation rate constant (k_d) becomes slower than the true k_d. Hence, the advanced Biacore with the lower noise level

Table 1
Noise level of the Biacore systems

System	Noise level (RU)
Biacore 1000	±2
Biacore 2000	±1
Biacore 3000	±0.3
Biacore T100	±0.1
Biacore T200	±0.03

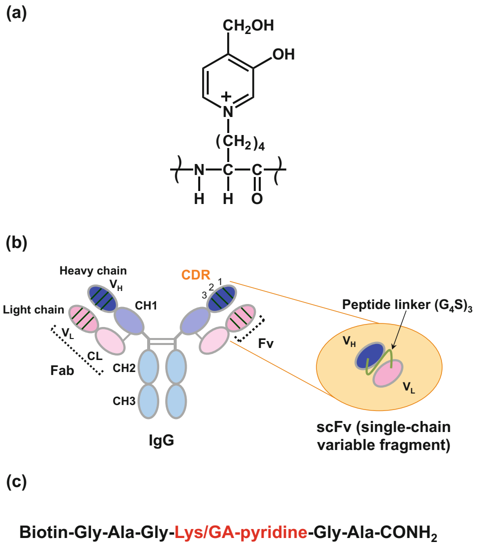


Fig. 1 **a** Chemical structure of GA-pyridine. This compound can be formed by glycation of ε-amino group of lysine residues in the proteins. **b** Schematic diagram of IgG and single-chain variable fragment (scFv). **c** Amino acid sequence of the GA-pyridine peptide as the ligand for Biacore analysis. Biotin was attached to immobilize the ligand peptide on the sensor chip SA by tight streptavidin-biotin interaction

allows for obtaining precise kinetic parameters as compared with the former system.

The second notable progress of the Biacore system is the improvement of the temperature controller, which allows for estimation of thermodynamic parameters by least-square fitting of the several sets of the SPR data acquired at the different temperature to the van't Hoff equation [2–4]. Moreover, the least-square fitting of the kinetic parameter at different temperatures to the Eyring equation allows for estimation of the thermodynamic parameters of the transient state [5, 6]. In this chapter, the authors will mention about the thermodynamic analysis using the Biacore system. In addition, advanced Biacore system clarified that some of the intermolecular interactions obey not to the 1:1 Langmuir binding mechanism, but to the two-state binding mechanism, in which structural rearrangement is induced after initial complex formation [6–8]. The authors will describe also analysis using the two-state binding model in this chapter. These topics are mentioned in the chapter as an example case of the molecular interaction system between GA-pyridine (Fig. 1a), a kind of advanced glycation end product (AGE) [9, 10], and the single-chain variable fragment (scFv) of the antibody (Fig. 1b) against GA-pyridine (anti-GA scFv).

2 Materials

2.1 Chemicals

(1) Anti-GA scFv (analyte)

(2) HBS-EP: 10 mM HEPES pH 7.4, 150 mM NaCl, 3.0 mM EDTA, and 0.005 % (v/v) Tween-20

(3) Regeneration Sol A: 10 mM NaOH

(4) Regeneration Sol B: 100 mM HCl

(5) Biotinylated peptide containing GA-pyridine (ligand) (Fig. 1c)

(6) Biotinylated unmodified peptide (negative control)

(7) Series S Sensor Chip SA (GE Healthcare)

2.2 Equipment

(1) Biacore T100 (or T200) (GE Healthcare)

(2) Biacore Control Software Version 2.0.2 (GE Healthcare)

(3) Biacore Evaluation Software Version 2.0.2 (GE Healthcare)

3 Methods

3.1 Immobilization of the Ligand to the Biacore Sensor Surface

The initial step of the Biacore experiment is immobilization of one of the molecules (the ligand) to the sensor surface without disrupting its activity. Several types of the Biacore sensor chip for various immobilization techniques can be commercially available, which is documented comprehensively in the Biacore handbook. In the present molecular system, biotinylated peptide containing GA-pyridine (GA-pyridine peptide; Fig. 1c) was immobilized on the Series S Sensor Chip SA, on which streptavidin had already been covalently immobilized. It has two important advantages. First, quantity of the immobilized GA-pyridine molecule can be easily controlled by duration of injection. Second, the GA-pyridine is seldom inactivated by indirect coupling, since all the molecules are immobilized in a known and consistent orientation on the surface. Resonance unit of approximate 30 RU of the GA-pyridine peptide was immobilized on the sensor surface, which exhibited the maximum resonance unit (RU_{max}) value of approximate 60 by the association of the anti-GA scFv, while the low-density sensor surface did not spoiled the quality of the data.

Procedure for Immobilization of the Ligand

(1) Set flow rate to 5 μL/min with the running buffer of HBS-EP.

(2) Inject 2.5 μL/min of biotinylated GA-pyridine peptide in HBS-EP (100 nM).

(3) Flow the running buffer.

(4) Repeat (2) and (3) where the amount of immobilized ligand reaches 30 RU approximately.

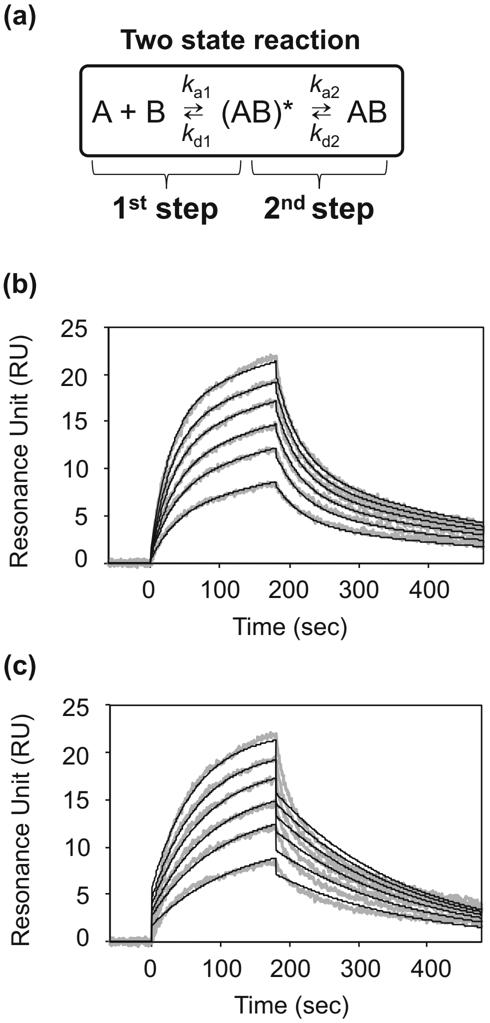


Fig. 2 (**a**) Schematic representation of the two-state binding model. Overlay of a series of the sensorgrams where the concentration of the analyte of 20, 30, 40, 50, 60, and 70 (nM) from *bottom* to *top*, respectively. Experimental curves (*gray line*) and the curves generated by fitted data to the two-state binding model (**b**) and the 1:1 Langmuir binding model (**c**), respectively, are presented

3.2 Acquisition and the Kinetic Analysis of the Sensorgram by the Two-State Binding Model

A two-state binding model is the most simplified scheme for quantitatively describing a two-step association process (Fig. 2a), where A and B represent antibody and antigen, respectively, and (AB)* and AB represent encounter complex/transition state and final stable (rearranged) complex, respectively. This model is an "induced fit" model, with a conformational change occurring after initial complex formation. Simulated sensorgrams for the two-state binding model are shown in Fig. 3. The equilibrium constants of the individual steps can be expressed as equations (Eqs. 1 and 2):

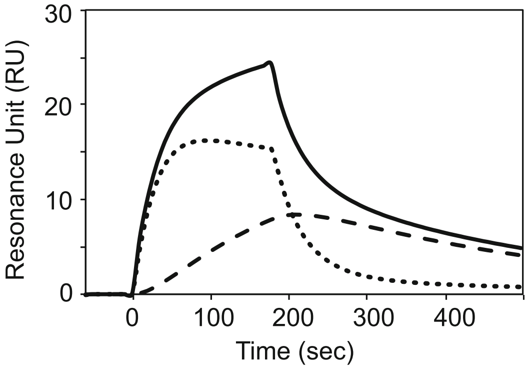


Fig. 3 Simulated component curves of the binding between the anti-GA scFv and the immobilized GA-pyridine peptide. Sensorgram (*solid line*) was fitted to the two-state binding model, and the simulated component curves were shown as transient complex $(AB)^*$ (*dotted line*) and stable complex AB (*dashed line*), respectively

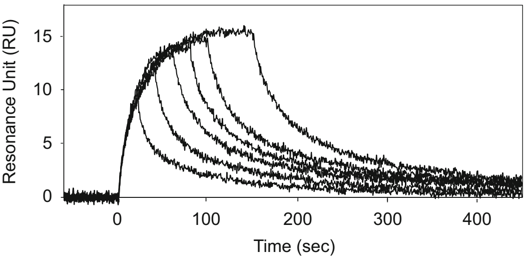


Fig. 4 Overlay of a series of the sensorgrams where the duration of the injection of 20, 40, 60, 80, 100, and 150 (sec) from *left* to *right*, respectively. Biotinylated GA-pyridine peptide (Fig. 1c) was immobilized on the sensor chip SA as ligand, and 70 nM of anti-GA scFv was injected as an analyte

$$K_{\mathrm{d1}} = \frac{k_{\mathrm{d1}}}{k_{\mathrm{a1}}} \tag{1}$$

$$K_{\mathrm{d2}} = \frac{k_{\mathrm{d2}}}{k_{\mathrm{a2}}} \tag{2}$$

and the overall equilibrium binding constant can be calculated as

$$K_{\mathrm{D}} = \frac{(k_{\mathrm{d1}} \times k_{\mathrm{d2}})}{(k_{\mathrm{a1}}(k_{\mathrm{a2}} + k_{\mathrm{d2}}))} \tag{3}$$

Statistically, the two-state binding model is complicated and fitting is better than the 1:1 binding model (Fig. 2b, c), while this is not always an experimental proof for selection of the proper model. To apply the two-state binding model, a series of the sensorgrams with various durations of the injection (Fig. 4) should be obtained and confirm that dissociation becomes slower as the duration of the injection

Table 2
Kinetic parameters for association between GA-pyridine peptide and anti-GA scFv at 25 °C

	$k_{a1} \times 10^{5}$ (1/Ms)	$k_{d1} \times 10^{-2}$ (1/s)	$k_{a2} \times 10^{-3}$ (1/s)	$k_{d2} \times 10^{-3}$ (1/s)	$K_D \times 10^{-8}$ (M)
anti-GA scFv	4.1 ± 0.5	3.7 ± 0.3	4.1 ± 0.2	3.6 ± 0.2	4.2 ± 0.6

increases, which is caused by accumulation of the stable complex (AB) for longer injection period, and experimentally validating application of the two-state binding model. Once application of the two-state binding model is validated, a series of the SPR data can be analyzed using Biacore Evaluation Software included in the instrument (Table 2).

Procedure for SPR Measurement

(1) Set flow rate to 50 μL/min with the running buffer of HBS-EP.

(2) Inject 150 μL/min of anti-GA scFv (analyte) in HBS-EP.

(3) Flow the running buffer for more than 0.50 mL.

(4) Inject 15 μL of Regeneration Sol A once, and 15 μL of Regeneration Sol B twice, to regenerate the sensor chip.

(5) Flow the running buffer for more than 0.25 mL.

3.3 Thermodynamic Analysis Based on van't Hoff and Eyring Equations

For further description of the molecular interaction, thermodynamic parameters including the binding energy change (ΔG), the enthalpy change (ΔH), and the entropy change (ΔS) could be used. The binding energy change can be expressed as (Eq. 4):

$$\Delta G = \Delta H - T\Delta S \tag{4}$$

Of these parameters, ΔG can be obtained by SPR data, and ΔH can be measured indirectly by van't Hoff analysis. Assuming that ΔH and ΔS are temperature independent, the linear form of the van't Hoff equation (Eq. 5) can be used:

$$\Delta G = RT\ln K_{\mathrm{D}} = \Delta H - T\Delta S \tag{5}$$

$$\ln K_{\mathrm{D}} = \Delta H/RT - \Delta S/R \tag{6}$$

The dissociation constant (K_{D}) was measured over a range of temperatures and ln (K_{D}) were plotted against $1/T$. The slope of this plot is equal to $\Delta H/R$ and the intercept $-\Delta S/R$ (Eq. 6). In real cases, ΔH varies with temperature for protein/ligand interactions and the plot is not linear. Consequently, K_{D} needs to be measured over a small range around the temperature of interest, where the plot can be assumed to be linear. Obtained thermodynamic parameters were plotted in Fig. 5b. Another approach is to use a nonlinear form of the van't Hoff equation (Eq. 7):

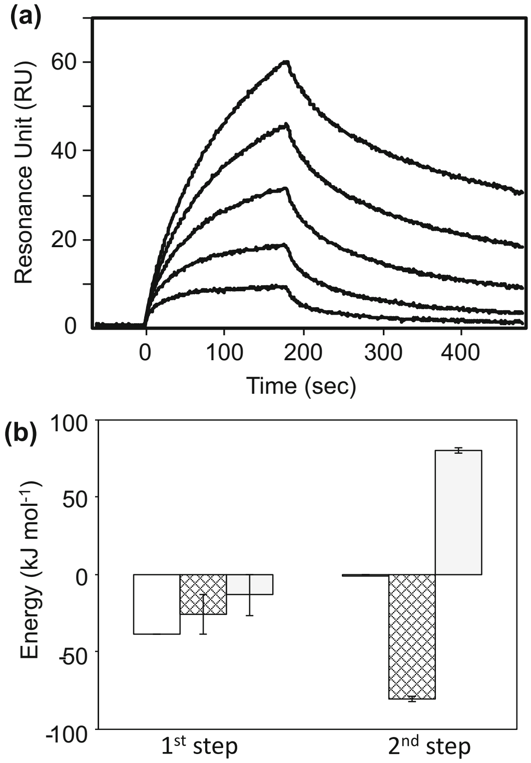


Fig. 5 (**a**) Overlay of a series of the sensorgrams where the temperature of 13, 17, 21, 25, and 29 (°C) from *bottom* to *top*, respectively. (**b**) Thermodynamic parameters estimated by van't Hoff equation (Eq. 7). Gibbs free energy change (Δ*G*) is shown in white bar, enthalpy change (Δ*H*) in *meshed bar*, and −*T*Δ*S* in *gray bar*, respectively. Thermodynamic parameters for 1st and 2nd steps are separately plotted

$$\begin{aligned}\Delta G_{T_0} &= RT\ln K_{\mathrm{D}}\\ &= \Delta H_{T_0} - T\Delta S_{T_0} + \Delta C_p(T - T_0) - T\Delta C_p\ln(T/T_0) \quad (7)\end{aligned}$$

where T is the Kelvin temperature (K), T_0 is a reference temperature (e.g., 298.15 K), ΔH_{T_0} is the enthalpy change upon binding at T_0 (kcal/mol), ΔS_{T_0} is the entropy change upon binding at T_0 (kcal/mol), and ΔC_p is the specific heat capacity change (kcal/mol·K), a measure of the dependence of ΔH (and ΔS) on temperature. In the molecular interaction system between anti-GA scFv and GA-pyridine peptide, temperature range of 13–29 °C was used for the analysis (Fig. 5a). A set of the sensorgrams at the analyte concentration from 20 to 70 nM was obtained at the different experimental temperatures. Thermodynamic parameters can be obtained by fitting a series of the sensorgrams using Biacore Evaluation Software. Obtained thermodynamic parameters were plotted in Fig. 5b. The first step of the binding appeared to be both entropy- and enthalpy-driven reaction, while the second step enthalpy driven.

Microcalorimeter (e.g., MicroCal-ITC) is a superior way for obtaining thermodynamic parameters, since it allows for direct observation of ΔH and ΔG upon ligand association. There is, however, a drawback that microcalorimeter requires about 100-fold more protein than the Biacore. Thus, the Biacore may be the only means of acquiring thermodynamic parameters under the limited sample amount.

Besides static thermodynamic parameters, Biacore allows for evaluation of the thermodynamic parameters for the transition state. The k_a and k_d generally increase with temperature. By fitting temperature dependency of the k_a and k_d to the Eyring equation (Eq. 8),

$$\begin{aligned}\ln(k/T) &= -\Delta G^{0\ddagger} + \ln(k_\mathrm{B}/h)\\ &= \Delta S^{0\ddagger}/R\ - \Delta H^{0\ddagger}/RT + \ln(k_\mathrm{B}/h) \qquad (8)\end{aligned}$$

where k is the relevant rate constant (e.g., k_a and k_d), R is the gas constant, k_B is a Boltzmann constant, and h is the Planck constant, thermodynamic parameters for the transition state could be obtained. This analysis is also implemented in the Biacore Evaluation Software. Obtained thermodynamic parameters for the molecular interaction system between anti-GA scFv and GA-pyridine peptide were plotted in Fig. 6a–c. The higher entropic energy barrier was occurred for the second step in the association phase, while there was no enthalpic barrier.

4 Conclusion

SPR is one of the powerful tool for the analysis of intermolecular interactions. In this chapter, the authors mentioned analysis of the intermolecular interaction based on two-state binding model. Calorimeter allows for measuring the affinity and the thermodynamic parameters for protein/ligand interaction while does not allow for separately determining the parameters for the first and the second step. Moreover, SPR has advantage in obtaining the thermodynamic parameters for the transition state. Thus, SPR-based thermodynamic parameter is effective for describing dynamic feature of the molecular interaction. For static thermodynamic analysis, SPR exhibits advantage that small amounts of protein are required as compared to calorimetry. Thus, the advanced Biacore system will contribute for accelerating molecular interaction study.

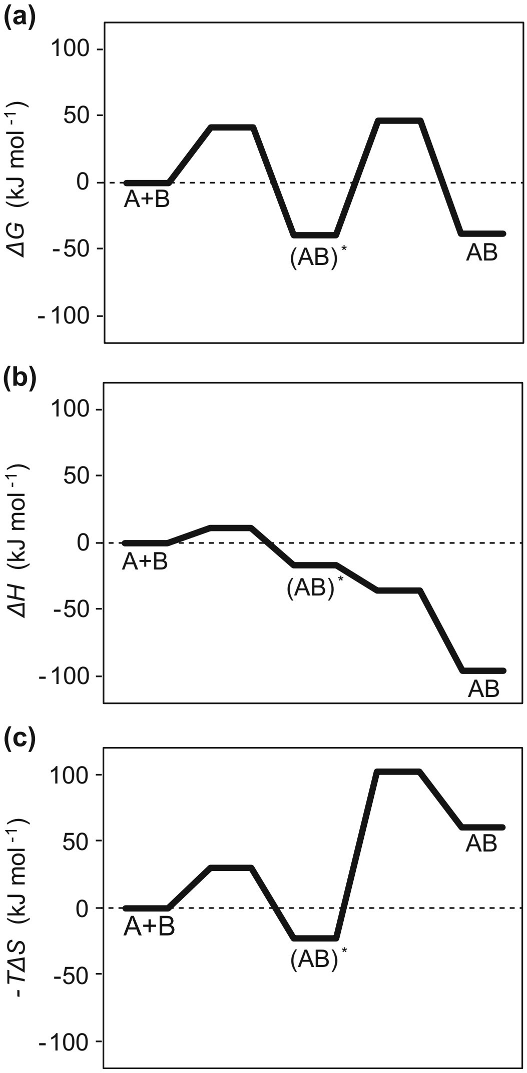


Fig. 6 Thermodynamic parameters for the binding pathway between GA-pyridine and anti-GA scFv estimated by Eyring equation (Eq. 8). The Gibbs free energy change (ΔG) is shown in (**a**), the enthalpy change (ΔH) in (**b**), and $-T\Delta S$ in (**c**), respectively

Acknowledgments

This work was partly supported by the Grants-in-Aid for Scientific Research (YK and HM) and the Grant-in-Aid for Scientific Research on Innovative Areas (HM) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Nagata K, Handa H (eds) (2000) Real-time analysis of biomolecular interactions: application of BIACORE. Springer, Tokyo
2. Tsumoto K, Yokota A, Tanaka Y, Ui M, Tsumuraya T, Fujii I, Kumagai I, Nagumo Y, Oguri H, Inoue M, Hirama M (2008) Critical contribution of aromatic rings to specific recognition of polyether rings: the case of ciguatoxin CTX3C-ABC and its specific antibody 1C49. J Biol Chem 283:12259–12266
3. Edink E, Rucktooa P, Retra K, Akdemir A, Nahar T, Zuiderveld O, van Elk R, Janssen E, van Nierop P, van Muijlwijk-Koezen J, Smit AB, Sixma TK, Leurs R, de Esch IJ (2011) Fragment growing induces conformational changes in acetylcholine-binding protein: a structural and thermodynamic analysis. J Am Chem Soc 133:5363–5371
4. Madura F, Rizkallah PJ, Miles KM, Holland CJ, Bulek AM, Fuller A, Schauenburg AJ, Miles JJ, Liddy N, Sami M, Li Y, Hossain M, Baker BM, Jakobsen BK, Sewell AK, Cole DK (2013) T-cell receptor specificity maintained by altered thermodynamics. J Biol Chem 288:18766–18775
5. Suzuki N, Tsumoto K, Hajicek N, Daigo K, Tokita R, Minami S, Kodama T, Hamakubo T, Kozasa T (2009) Activation of leukemia-associated RhoGEF by Galpha13 with significant conformational rearrangements in the interface. J Biol Chem 284:5000–5009
6. Walsh ST (2010) A biosensor study indicating that entropy, electrostatics, and receptor glycosylation drive the binding interaction between interleukin-7 and its receptor. Biochemistry 49:8766–8778
7. Futamura M, Dhanasekaran P, Handa T, Phillips MC, Lund-Katz S, Saito H (2005) Two-step mechanism of binding of apolipoprotein E to heparin: implications for the kinetics of apolipoprotein E-heparan sulfate proteoglycan complex formation on cell surfaces. J Biol Chem 280:5414–5422
8. Lipschultz CA, Li Y, Smith-Gill S (2000) Experimental design for analysis of complex kinetics using surface plasmon resonance. Methods 20:310–318
9. Nagai R, Hayashi CM, Xia L, Takeya M, Horiuchi S (2002) Identification in human atherosclerotic lesions of GA-pyridine, a novel structure derived from glycolaldehyde-modified proteins. J Biol Chem 277:48905–48912
10. Greven WL, Waanders F, Nagai R, van den Heuvel NC, Navis G, van Goor H (2005) Mesangial accumulation of GA-pyridine, a novel glycolaldehyde-derived AGE, in human renal disease. Kidney Int 68:595–602

Part V

Advanced Methods for Structural Analyses

Chapter 14

Structural Biology with Microfocus Beamlines

Kunio Hirata, James Foadi, Gwyndaf Evans, Kazuya Hasegawa, and Oliver B. Zeldin

Abstract

Protein microcrystallography, which analyzes crystals smaller than a few tens of microns, is becoming one of the most attractive fields in structural biology. To realize the complete potential of this technique, it is inevitable that microcrystallography has to be combined with novel data collection instruments and strategies. Recently, a highly brilliant X-ray beam with micron size has enabled the measurement of diffraction data from such microcrystals (Smith JL, Fischetti RF, Yamamoto M, Micro-crystallography comes of age. Curr Opin Struct Biol 22:602–612, 2012). Here, we describe important instrumentation at synchrotron facilities and experimental strategies.

Keywords Protein crystallography, Microcrystals

1 Introduction

X-ray crystallography is one of the most effective techniques for elucidating the detailed structural features of proteins. The first X-ray structural analysis of myoglobin was achieved 50 years ago. Since then, the considerable evolution of various techniques has resulted in the current focus on macromolecular crystallography (referred to as MX). One of these techniques involves the focusing of X-rays achieved at synchrotron facilities.

Reflection intensity, which is required for structural analysis by MX, weakens as crystal size decreases. The primary solution for enhancing intensity is to expose the crystal to intense X-rays. Matching the beam size to the crystal size is also effective in reducing background from the non-crystal volume. When only micron-sized crystals are available, microfocused beams with high brilliance are required.

There have been numerous improvements in the application technologies of synchrotron facilities (Sect. 2). For example, recent designs for supporting mechanics stabilize optical elements, eventually improving both beam position and intensity.

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks,
DOI 10.1007/978-4-431-56030-2_14, © Springer Japan 2016

Novel techniques for fabricating the surface of focusing mirrors have increased photon density at a sample position. Enhancements in pixel resolution, readout time, and the quantum efficiency of the X-ray detector have achieved remarkable productivity improvements in the data collection process. There have also been technological innovations in preparing protein crystals and computation. For example, the so-called lipidic cubic and sponge phases have drastically accelerated crystal structure analysis of membrane proteins. The introduction of robotics for crystallization as well as visualization has increased throughput in the crystallization step. Further, recently developed software has reduced difficulties related to both phasing and refining of structures.

Although technologies have improved, data collection, which is mainly conducted by humans, is still the most critical and important process in MX. There are difficulties involved in determining suitable experimental conditions such as the oscillation step, total number of images, camera distance, beam size, and exposure time. In particular, it is quite difficult to prevent radiation damage of the crystals. From one perspective, the history of MX can be regarded as a history of the battle against radiation damage. Radiation damage of crystals increases as the beam intensity increases. If damage was preventable, high-resolution structural analysis could be easily achieved by exposing the crystals to as many X-ray photons as possible. However, as described in Sect. 3, radiation damage exists in the real world and often hinders high-precision data collection. Thus, appropriate data collection strategies should be devised before starting diffraction experiments.

Let us consider the following simulation. We have a lysozyme crystal 100 μm in size and an X-ray with a wavelength which corresponds to 1 Å with a photon flux of 10^{10} photons/s in a 1-μm square area. The absorbed dose of lysozyme crystal, estimated with the RADDOSE program, reaches 20 MGy after a 7-s exposure. If we utilize the 100-μm square beam with the same photon flux, 10^{10} photons/s, the exposure for 20 MGy roughly corresponds to 70,000 s. It is worth noting two points. The first is that the crystal loses roughly half its diffracting power when it absorbs an energy of 20 MGy from incident X-rays. The second is that the quantitative difference in absorbed dose can be explained by the "photon flux density." The photon flux density in the first example is 10^{10} photons/s/μm^2, and in the second example, it is 10^6 photons/s/μ m^2. The difference in the order of absorbed dose is comparable to the difference in photon flux density. Once the absorbed dose for the sample crystal for one beamline is calculated, the dose for other beamlines can be roughly determined by calculating the photon flux density of the two beamlines. Ideally, we would obtain reflections with roughly the same signal-to-noise ratio, or "resolution limit," by exposure to the same number of X-ray photons. If we choose 1-s exposure time for both beam sizes, the 1-μm focused beam gives only 7 frames and the 100-μm beam can give 70,000 frames with the same resolution. This example roughly demonstrates the challenges inherent in "protein microcrystallography."

Although these difficulties exist, the use of microbeams is inevitable for crystals that are <20–30 μm. Especially in protein microcrystallography, data collection strategies should be carefully considered. The primary branch point is the selection of the study objective. Is it for phasing? Is it for native structure at higher resolution? Can you permit the use of multiple crystals? The lifetime of a crystal is independent of your objective and cannot be changed. Thus, the crystal lifetime should be distributed either "resolution" or "redundancy." In extreme cases, you can choose to obtain only one diffraction image by delivering X-rays for the crystal lifetime. This approach would give you the highest resolution from the crystal. Therefore, multiple crystals would be required to complete the dataset. This "multiple crystal strategy" is useful for collecting native data at higher resolution. The associated technical issues and methods used to merge multiple datasets are described in Sect. 4. Although the multiple-crystal strategy is also effective for initial phasing, as reported in the Sulfur SAD papers [2], there also are difficulties involved in preparing isomorphous crystals suitable for this purpose. In these cases, we have no choice but to collect a dataset from one crystal to reduce systematic errors in merging datasets and enhance anomalous signals. Hence, in this strategy, the lifetime of the crystal is utilized to increase redundancy at the expense of resolution. For this purpose of the microbeam, the so-called helical data collection is effective, both for enhancing the S/N ratio and increasing the redundancy for one crystal (Sect. 5).

Radiation damage of protein crystals exhibits different characteristics at different energies. In particular, using higher energy for microcrystallography may allow us to extract maximum information with minimum radiation damage (Sect. 6).

Recently, the novel light source, X-ray free electron laser (XFEL), has had a tremendous impact on biology. This source delivers extremely intense femtosecond X-ray pulses and allows the structural determination of proteins without radiation damage. In Sect. 7, we will describe the current state of MX using XFELs and present some experimental results.

2 Concept and Design of Microfocus Beamline

To enhance the signal-to-noise ratio of diffraction, background scattering from solvent and the sample support should be reduced. For this primary purpose, a microfocus beam is required to collect good diffraction data from microcrystals. Beamlines dedicated to microcrystallography target samples smaller than 20 μm; the beams range in the size of 1–20 μm. Thus, microbeams can also be utilized to illuminate a well-diffracting region within an imperfect crystal or a single-crystal region in a multi-crystal sample. For example, biological supramolecular complexes often yield clustered or inhomogeneous crystals. In such cases, a diffraction volume within a crystal can be selected with a microfocused beam.

Beamline ID13 at the European Synchrotron Radiation Facility (ESRF) was the first example of a microfocus beamline for macromolecular crystallography (MX). The microbeam of beamline ID13 was critical for several structure determinations before the turn of the century (reviewed in [3]) and led other synchrotron facilities to develop microbeam beamlines dedicated to MX. They achieved 1×1 μm focusing with a K-B mirror and 10^{10} photons/s. After this pioneering work, beamline ID23 was constructed at ESRF to focus the beam to 5 μm [4]. Let us consider the types of advanced technologies required for achieving this type of micron-sized X-ray beam.

At third-generation synchrotron radiation facilities, low-emittance electron beams (<10 nm rad) and coupled insertion devices are key technologies for microfocusing. In general, for diffraction data collection, a smaller beam divergence is preferable for separating neighboring diffraction spots on an X-ray detector. A low-emittance electron beam is the most important property for both smaller beam size and smaller beam divergence. Combining better electron beams and advanced insertion device produces X-rays with brilliance on the order of 10^{20} photon/mm^2/mrad^2/0.1 % BW in the energy range of 5–35 keV.

Remarkable improvements to the beamline optical elements help the system maintain a stable microbeam. One of them is a new cooling technique for monochromatization crystals. The brilliant source from the insertion device generates white X-ray beams that have energy spectra. In general, the monochromator crystal, which must provide a stable output beam under varying thermal load conditions, is exposed to the white beam and receives an enormous heat load. This heat load often changes the local shape of the crystal and the direction of the monochromatized X-ray output. Because the water-cooling technique lacks the ability to remove the heat load enough, almost all of the recent high-flux beamlines have adopted liquid nitrogen flow for cooling monochromator crystals. This technique eliminates the positional drift of the output X-rays and improves the stability of the beam position and intensity. In addition, the "top-up" operation mode [5] of the synchrotron radiation facility significantly reduces thermal drift because the heat load is kept constant.

Another improved technique is concerned with the "focusing element," such as focusing mirrors. Surface error on the mirror element generates residual scattering and makes it difficult to achieve good focusing. The fabrication technique of mirrors is one of the most important challenges for microfocus beamlines. For example, the elastic emission machining (EEM) technique, developed in Osaka University [6], can process a mirror surface with "atomic level." The first success of EEM mirrors was reported in 2005 [7]. Although EEM mirrors have a fixed curvature, it is often preferable to utilize various beam divergences and different focal points for data collection. For example, when the focal point is set to the X-ray detector surface, diffraction spots can be separated, under ideal conditions, during the integration of the diffraction intensities. This purpose demands tunability of the mirror curvature. Although the mirror curvature can be changed by bending its shape with motorized mechanics, it is often difficult to make the best shape for microfocus using this technique. The reason is that the technique generates unintended mirror shapes using one axis to push/pull the plate of the mirror. To avoid this problem, the so-called bimorph mirror was developed and utilized in several microfocus beamlines. The bimorph mirror has a series of piezo-actuators that are used to shape its surface. These actuators enable more precise mirror shaping within a local area of the mirror. The bimorph mirror enables changing the divergence and focal point, allowing a greater variety of diffraction experiments.

Temperature change and vibration are obvious difficulties in making a microfocus beam. Compared with conventional X-ray beam for protein crystallography, positional instability of the microbeam is critical for data quality in protein microcrystallography. For example, when the 1-μm focused beam is exposed to 1-μm crystal, a positional change of approximately 0.6 μm reduces the diffraction intensity to half. For this reason, positional errors of the light source, optical elements, and diffractometer should be as small as possible. Substantial resistance to vibrations and temperature fluctuations can be achieved simply by granite support tables and heavy, stiff mechanisms for all of the optical components. Intensity fluctuation should also be removed to ensure precise data collection [8].

At the beamline BL32XU at SPring-8, one of the most important developments to eliminate the fluctuation of both position and intensity is the double-crystal monochromator (DCM). Cooling agents, such as liquid nitrogen and water, vibrate monochromatizing crystals and cause both these fluctuations. Much effort has been focused on developing rigid DCMs for stable beamline operation. For example, a smaller number of motorized axes enhances rigidity and largely eliminates the vibration. We should also take steps to eliminate vibrations from outside the monochromator chamber. Vacuum pumps, which lower the electron density in the X-ray path, often generate substantial vibration at the beamline. Rubber seats are a good solution that can stop the transition of vibration from the pump to the ground [9].

For more precise detection of the diffraction intensity from microcrystals, the time scale of the X-ray beam fluctuations is also important. The exposure times necessary to obtain a diffraction image range from a few seconds, under conditions typically available at a modern beamline, to a few tens of milliseconds when the beamline is equipped with a fast-framing detector. Given these detection technologies, fluctuations in beam intensity on the time scales of 0.01–200 Hz can result in more precise data collection.

The optical design of the beamline absolutely defines beamline ability. The first step in designing a microfocus beamline is to select the intended target proteins. Using the molecular weight of the target proteins, a distribution of the cell parameters that can be expected can be obtained from the RCSB-PDB (http://www.rcsb.org/). After considering the selected cell parameters and expected mosaic spread, diffraction patterns should be simulated to evaluate the overlap of diffraction patterns on the X-ray detector. This simulation lets the designer determine the allowable divergence of the microbeam. In general, beam divergence becomes larger when a smaller size beam is required. The balance between size and divergence can be tuned by setting the proper optical configuration of the beamline. Modern microfocus beamlines and their concepts are summarized in greater detail in a study published by Janet Smith [1].

2.1 Diffractometer for Microcrystallography

In this section, modern techniques for precise data collection from tiny protein crystals are reviewed. The required specifications for the diffractometer are very simple. They are to precisely irradiate the microcrystal with the microbeam and to detect weak diffraction intensity quickly and efficiently. This section is divided into five sections entitled "To Know the Position of Sample," "Move It and Fix It," "Detecting Weak Diffraction," "Efforts to Enhance the Signal-to-Noise Ratio," and "Manipulation of Microcrystals" (Sects. 2.1.1, 2.1.2, 2.1.3, 2.1.4, and 2.1.5).

2.1.1 To Know the Position of Sample

A high-magnification optical microscope is required to position the microsized sample in the center of the beam. Constructed beamlines recently have often adopted a coaxial microscope, which gives the researcher a viewpoint along the direction of the X-rays. The merit of using the coaxial microscope is that it is easy to hit the sample compared with a microscope aligned with the camera, whose viewpoint is inclined from the beam. Moreover, this camera can visualize the precise beam position from the X-ray scintillator, such as a crystal made of YAG, to the sample position. This beam position monitor helps to fix problems that arise during beamline tuning.

Visible light is not often able to detect the crystal position when some reagents that work as light shields exist in the sample loop, such as ice and lipids. Several interesting techniques to enable the visualization of sample position in the beamline have been developed. One of these techniques is to utilize UV light to detect crystal position. Proteins that contain tryptophan can be excited with UV light and produce fluorescence, which can be detected as visible light [10, 11]. Another development is a technique referred to as second-order nonlinear imaging of chiral crystals (SONICC) [12].

In this technique, a femtosecond pulsed laser exploits the frequency-doubling properties of most protein crystals to locate them in the presence of noncrystalline substances. When these robust techniques are used to visualize sample crystals independent of the sample environment, such as cryoprotectants and lipids, the efficiency of protein microcrystallography will be dramatically improved.

2.1.2 Move It and Fix It

The specifications for the goniometer must include high positional precision, particularly for protein microcrystallography. This includes three-dimensional translation axes and a spindle axis for the rotation method. Translation axes with submicron precision are generally employed in microfocus beamlines. A linear encoder monitoring the absolute position of the motorized translational axes with submicron precision helps to ensure that the sample is fixed at the same position or correctly positioned. The most important axis for precise data collection is the rotation axis. If the position along this axis lacks precision, the diffraction signals will be reduced because the crystal will be off-center with respect to the beam during its data collection. This is extremely difficult to distinguish from radiation damage particularly for microcrystals. The air-bearing goniometer substantially reduces the sphere of confusion and has been adopted in many microfocus beamlines. For example, at BL32XU at SPring-8 (Harima, Japan), the sphere of confusion of the air-bearing goniometer corresponds to 0.5 μm [13].

Vibration in the experimental hutch causes deterioration of the data quality because it causes fluctuation in the beam intensity at the detector during X-ray exposure. At microfocus beamline, the base plate of diffractometer often comprises granite, which is suitable for eliminating the vibration because it is heavy and comprises multiple rock products. The heavy base plate of the diffractometer should be fixed tightly to the ground to ensure high rigidity. Vibration of the sample crystal can be minimized by utilizing this type of highly rigid diffractometer.

In addition, it is important to control the temperature inside the hutch. All materials show thermal expansion; thus, their lengths are easily affected by temperature change. For example, the coefficient of linear expansion of the iron corresponds to $12.1 \times 10^{-6}/$ K. Thus, a bar composed of iron whose length is 100 mm expands 1.2 μm when the temperature is raised by 1 °C. This change is critical for data collection from few micron-sized crystals. Materials with smaller coefficient of linear expansion should be adopted for instrumentation, although it is expensive. For this reason, a precise air-conditioning system that can control room temperature ≤ 0.1 °C is required for stable experiments.

2.1.3 Detecting Weak Diffraction

Diffraction signal becomes weaker when the protein crystal is smaller, according to Darwin's formula. The quantum efficiency of X-ray detectors has shown dramatic improvement recently. Combining the advanced technologies of CCD chips, tapered fibers, phosphor screens, and several CCD detectors allows recent instrumentation to detect one X-ray photon at 12.4 keV. PILATUS [14], a pixel array detector, which can directly count the incident X-ray photons, is one of the most powerful detectors. It has been adopted in many synchrotron radiation facilities. In general, a direct detector is incapable of counting the large intensity region with linearity, although readout noise is negligible. This deficiency can be overcome by altering the frame rate. The frame rate of the PILATUS detector corresponds to a few milliseconds. The higher intensities on the direct device should be sliced along the rotation axis and detected as weaker signals suitable for photon counting with good linearity. For this reason, a fine ϕ slicing data collection technique is particularly suitable for a direct detector. Improvements that enhance the speed of the data collection can be observed in CCD/CMOS technologies [15]. In addition to the speed of data collection, the pixel size and detector size are also important to discriminate diffraction spots on the detector surface and should be considered with the beam size and the beam divergence at the beamline. There is a report that radiation damage can be mitigated using a highly efficient detector [16]. The technological advancements in these X-ray detectors are allowing high speed and highly efficient data collection in protein microcrystallography.

2.1.4 Efforts to Enhance the Signal-to-Noise Ratio

In addition to improving detector efficiency, background noise should be reduced to enhance the signal-to-noise ratio of diffracted intensities. General microfocus beamlines adopt pinhole or metal pipe techniques to protect against air scattering generated by the intense X-rays [17]. The beamline BL32XU at SPring-8 is equipped with a pinhole of 30-μm diameter 7-mm upstream of the sample. This pinhole can eliminate both parasite scattering, which comes primarily from the focusing mirrors, and air scatterings. A metal tube with a diameter of 300 μm is attached to the pinhole, and this mitigates the additional weak-intensity parasite scattering from the pinhole. It is extremely important to take care of the sample environment because the amount of air scattering increases proportionally with the incident beam intensity. At BL32XU, the background scattering detected with a Rayonix MX225HE detector corresponds to a maximum of approximately 80 ADU/pixel at the lowest resolution range when collecting a total of 3×10^{11} photons. The diameter of the beam stopper should be small to detect diffractions at the lower resolution range.

The helium chamber surrounding the diffractometer is undergoing development at Photon Factory BL17 and SPring-8 BL32XU. Background scattering is reduced by purging the chamber with helium gas. For this purpose, the diffractometer is surrounded by an acrylic box, and helium gas flow is utilized to cool the sample crystal. After sealing, the purged helium gas shows 10 % lower background scattering at 12.4 keV compared with an environment of air [13].

2.1.5 Manipulation of Microcrystals

In general, sample crystals are selected and manipulated manually using a cryo-loop or similar devices. Protein microcrystals are considerably sensitive to physical damage caused by manipulation, temperature change, osmotic pressure change, etc. In addition to these, it is difficult to visualize the sample crystal when its size is <10 μm. Manipulating the microcrystals becomes an exceedingly difficult process compared with the conventional crystallography.

To overcome these problems, techniques for manipulating microcrystals are being developed at several synchrotron facilities. At SPring-8 and the Diamond Light Source, a robot with *laser* tweezers is being developed for this purpose [18, 19]. The laser optical tweezers generate motive force by the laser and can manipulate the protein crystals with optical force. The manipulation robot developed at SPring-8 can flash-cool the crystal after collecting it using cryo-loops. Crystals with a maximum size of 30 μm can be manipulated. The "acoustic mount" developed at NSLS in the USA makes protein microcrystals jump to the cryo-loop from the harvest solution using sound waves [20]. This is one of the techniques that can automatically mount many tiny crystals onto loops.

It is tedious to search for the best conditions for cooling protein crystals. Microcrystals, in particular, require more rapid screening of crystallization conditions to investigate different conditions. One of the solutions for this requirement is called plate screening. In this method, crystallization drops on the crystallization plate are directly exposed to an X-ray beam [21, 22]. An advantage of this method is that it circumvents the process of cooling the crystals. This helps the crystallographer examine the diffraction quality of the crystals without any deterioration occurring during the crystal manipulations that are required for cooling. This method is quite powerful, particularly for microcrystals. In some cases, a full dataset can be collected using multiple crystals at room temperature [23].

3 Data Collection

In this section, detailed methods and difficulties in data collection using a microfocused beam with high flux are described.

The BL32XU beamline at SPring-8 is dedicated to protein microcrystallography. Diffraction data with atomic (1.7 Å) resolution has been acquired from a 5-μm-sized protein crystal at this beamline. This crystal was a polyhedral virus crystal with space group *I*432. From the unit cell volume, this crystal includes 10^6–10^7 unit cells. The number of unit cells included in illuminated volume by the X-rays is a useful benchmark for evaluating the limitation of diffracting power from small crystals [3]. The size limitation is being reduced as beamline techniques improve.

The microfocused beam is a powerful tool for structure determination using microcrystals, but it is difficult to align the sample with the X-rays. One of the reasons is the size of the crystal. Optical resolution is insufficient to visualize a sample with a size <10 μm. In addition, the lens effect caused by the surrounding cryoprotectant disturbs direct observation of the actual crystal position with an optical microscope. The location of the crystals is skewed by this effect, and this phenomenon is critical for aligning tiny micron samples with the X-rays. Moreover, membrane protein crystals grown in a lipidic cubic phase [24] cannot be visualized with an optical lens primarily because of frozen lipids. The layers are almost opaque to visible light.

To avoid these problems, a "coaxial microscope" described in the previous Sect. 2.1.1 is normally utilized on a microfocus beamline. This camera can more readily remove the lens effect compared with a non-coaxial camera, although problems of invisible crystals still remain. To find and align the crystal, one of the solutions is a "raster scan" with the X-ray beam [25, 26]. This method visualizes the crystal position using diffraction images that are collected from different irradiation points. After this sequence, the invisible crystal can be detected where significant diffraction spots are observed. In this case, matching the beam size to the crystal size is also very important. When the crystal size is 10 μm and a 50-μm beam is utilized for raster scanning, the background scattering from the lipids causes the signal to disappear. The X-ray dose for a raster scan should be very small because this process is normally conducted for finding the crystal before data collection. Residual background scattering should also be reduced as much as possible. When the loop size is wide, many images are required for this method. Thus, the readout speed of the X-ray detector is critical for finding the crystal. In addition, the quantum efficiency of the detector is important to minimize the absorbed dose during the raster scan. During/after the raster scan, acquired images are analyzed to determine if diffraction is observed. Because many frames should be processed for detecting the crystal position, this should be automated [27]. This type of programs often misinterprets the crystal position because quantifying the diffraction quality depends on many parameters such as intrinsic diffraction power of crystals and background noise. Besides, diffraction images to be analyzed sometimes include strange patterns like ring from ice. Likewise, lipidic cubic phase crystals normally generate ring-shaped diffraction patterns from the lipid layer structure, which causes problems during auto-detection of the diffraction spots using the raster scan process with low-dose exposure. More suitable and efficient software with rationally defensible judging parameters should be developed at the beamlines. Moreover, it is considerably important to enhance the ability of the beamline control interface for an easy raster scan [28].

The structure of the human β_2 adrenergic G-protein-coupled receptor was revealed and reported in 2007 [29]. In this paper, the authors stressed that utilizing the microfocus beam was required for the data collection. The data collection was conducted at ESRF and the Advanced Photon Source (APS). The microbeam, whose size was 4 × 6 μm, was used to irradiate a crystal with a size of 300 × 30 × 10 μm. Full diffraction data collection, covering an oscillation range of 182°, was completed by changing the irradiation points on the crystal during data collection. The desired resolution, from 3.4 to 3.7 Å, limited the rotation range for data collection from each irradiation point to 5–10° because of radiation damage.

These results provided us with three important hints to solve crystal structures from microcrystals. The first hint was, of course, the importance of using a microfocus beam that matches the crystal size. Second, radiation damage to the protein crystal was more severe when the crystal size was small. Third, severe radiation damage could be avoided by translating the irradiation points using a microfocused beam. In addition, we should be able to recognize when diffraction datasets from multiple crystals should be merged.

The next section (Sect. 3.1) demonstrates the simple physics of radiation damage and its method of estimation. Subsequently, one of the data-merging techniques named the clustering technique will be described in detail (Sect. 4.1). Finally, the data collection strategy employed with one crystal will be shown, based on the experimental results at SPring-8 BL32XU (Sect. 5).

3.1 Radiation Damage with Micron-Sized Beams

An unavoidable phenomenon associated with the use of high-intensity X-rays to obtain atomic structures from protein crystals is radiation-induced damage. This has presented a significant challenge to macromolecular crystallographers ever since the earliest days of the field. Although cryo-cooling increases the radiation lifetime of crystals by around a factor of 70 [30], many data collections are still limited by radiation damage, particularly at the high brilliances associated with microbeams.

For a typical 100-μm-thick crystal, only about 2 % of a 12.4 keV incident beam will interact with the crystal [31]. Out of this 2 %, only 8–0.16 % of the incident beam will be elastically scattered, contributing to diffraction. The remaining 92 % of the interacting photons will contribute to radiation damage, principally in the form of photoelectric absorption of the incident photon, which leads to electron cascades as the highly energetic photoelectron propagates through the crystal. A smaller fraction of damage is caused by the inelastic (Compton) scattering of incident photons, where a recoil electron is emitted, leading to damage. The absorbed energy is measured in grays (Gy) (J/kg), and this is widely accepted as the appropriate metric against which to evaluate radiation damage.

The probability of photoelectric absorption depends on the atomic number and is energy dependent, meaning that choice of buffer can have a significant impact on the radiation sensitivity of protein crystals. Trying to avoid heavy atoms in crystallization conditions wherever possible can increase the lifetime of a crystal; see [32] for a list of heavy atom concentrations required to double the probability that an X-ray will be absorbed in a typical protein crystal, calculated with RADDOSE v2 [33].

Because cross sections for damage processes are energy dependent, it is desirable to consider if there is an optimum energy at which to collect data in MX. Although the elastic cross section increases gradually with energy, the inelastic, damage-causing, Compton cross section also increases. Phasing considerations and the elemental composition of a crystal will affect the theoretical optimal energy for data collection, and there is no clear consensus within the field for a "best practice" with respect to collecting data at very high or low energies [34]. Furthermore, despite the promise of potential advantages with high-energy data collection, most MX beamlines are optimized for the ~10 keV energy range (principally for phasing reasons), and so significant practical hurdles remain.

The damage caused by these initial absorption events of the incident beam in the crystal is referred to as primary damage. However, after these have taken place, we are left with highly energetic electrons still in the crystal: photoelectrons, Auger electrons, and recoil electrons. These particles will progress through the crystal, losing energy as they interact with atoms, knocking off more electrons, and forming energetic species. This secondary damage is the dominant driver of damage in MX experiments.

The specific mechanisms that cause damage are highly temperature dependent, with a change in behavior associated with the glass transition at around 180 K [35]. Above this transition, there is a great deal of complex radiation chemistry that can occur, and we observe an extremely wide range of sensitivities to damage [36]. This variability is principally attributed to the diffusion of reactive species created in the buffer, mother liquor, and protein. Below ~180 K, the diffusion of many larger reactive species appears to be frozen out, as suggested based on experimental observations of OH radicals using UV-vis microspectrophotometry by Owen et al. [37]. Because of this, cryo-cooled crystals are typically ~70 times more radiation resistant than room temperature crystals [30]. Due to the simplified radiation chemistry, cryo-cooled protein crystals are often considered as an amorphous glass, with a more consistent decay behavior as they are exposed to an X-ray beam compared with crystals at room temperature.

In the MX experiment, we observe a diffraction image on the detector and then process a set of these to produce a spatially and temporally averaged model of the macromolecule of interest. Radiation damage manifests itself in the diffraction data, and we classify damage by how it is observed. Global damage is observed in reciprocal space: in the diffraction image and in the statistics of a dataset. Specific damage refers to structural changes observed in the final electron density maps as a consequence of damage.

3.1.1 Global Damage

The main consequences of global damage on a MX dataset are:

1. A loss of intensity throughout the diffraction pattern, in particular at high resolution. This can be seen by comparing the three images in Fig. 1.
2. An increase in mosaicity, which is so significant across these three images that we can also directly see the spots become larger and less well defined in the extreme dose image in Fig. 1c.
3. An increase in unit cell volume. This is not visible by the eye from Fig. 1.
4. Increase in R factors. As the structure becomes more damaged, it is intuitive that metrics measuring its fidelity to an undamaged model will get worse, so R factors (R_{free}, R_{cryst}, R_{pim}, R_{sim}) increase.
5. An increase in B factor. The Wilson B factor increases with damage, as the average atomic positions defining the structure become less well defined.
6. Loss of isomorphism. Individual monomers in the unit cell can rotate or move slightly as damage progresses. This is particularly nefarious, since, along with the increase in unit cell volume, it can overwhelm the small changes in intensity that must be measured for anomalous phasing techniques.

3.1.2 Specific Damage

Unlike global damage, specific damage is highly dependent on the local environment, and a quantitative theory has not yet been developed. At 100 K, there is a broad agreement on what types of damage are most commonly observed and on the sequential order in which these take place:

1. *Reduction of metallocenters* occurs at doses as low as 45 kGy at 100 K [38]. There is some evidence [39] that cooling to 40 K can have a significant effect on the sensitivity of metallocenters, and helium cooling may thus be advisable for the study of proteins with metals in the active site, if their oxidation state is of interest.
2. *Elongation and breaking of disulfide bridges* [40–42].
3. *Decarboxylation of glutamates and aspartates* [40–42].

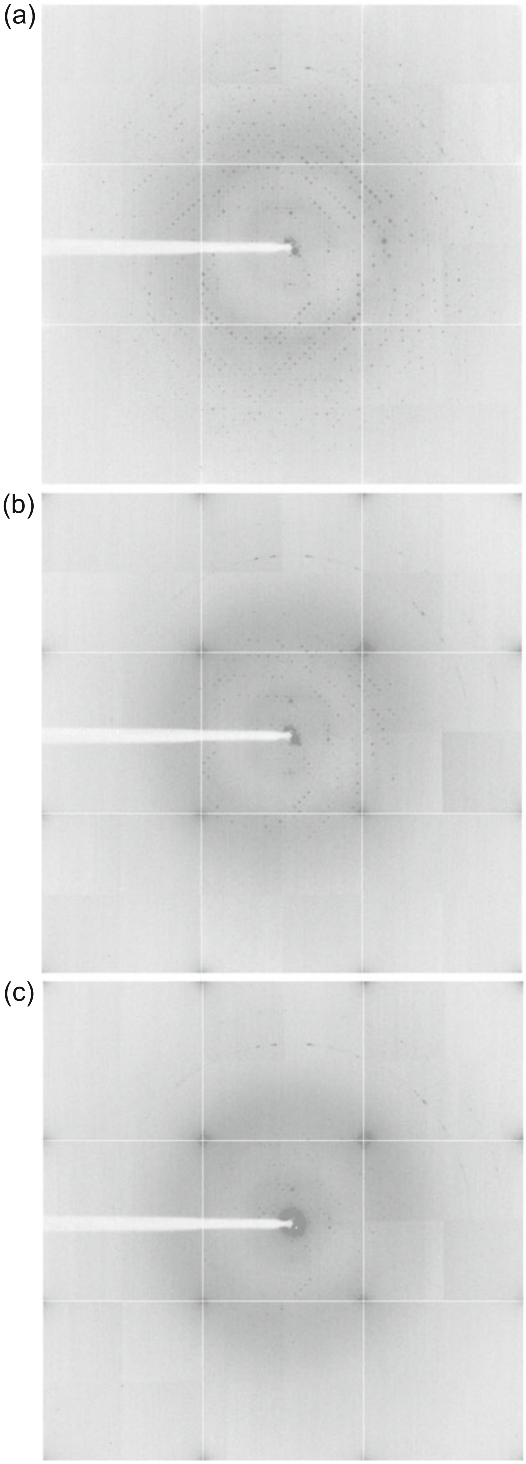


Fig. 1 Examples of images collected with the same exposure parameters from a crystal of cubic insulin at 100 K after various levels of radiation damage. Note how the intensity drops in the high-resolution bins and how the remaining spots increase in size due to mosaicity increases. (**a**) Low dose (∼kGy) dataset. (**b**) High dose (30 MGy) dataset. (**c**) Extreme dose (60 MGy) dataset

4. *Dehydroxylation of tyrosine* [40, 41].
5. *Cleavage of covalent bonds to heavy atoms* is common, as was observed for the cleavage of a mercury derivative by Ramagopal [43].

3.2 Predicting Crystal Lifetime

Several studies have shown that radiation damage is a function of energy absorbed per unit mass (dose = J/kg). Owen et al. [44] found that overall diffraction intensity decays linearly with dose at 100 K and that a dose of 30 MGy was a useful upper limit for the dose tolerance of macromolecular crystals at 100 K. Around the same time, Kmetko et al. [45] independently performed a similar analysis and proposed the use of the relative B factor, B_{rel}, which is the difference in scaling B factor when scaling together successively damaged datasets as a linearly dose-dependent damage metric. The gradient of this line, the coefficient of sensitivity, was found to be approximately 0.014 $Å^2$/MGy.

These two metrics are the currently most widely used measures of radiation damage. It is worth noting that both of these important studies went to significant lengths to ensure that the distribution of dose within the crystal volume was even and so that a one-dimensional treatment of dose could be applied. A one-dimensional treatment is appropriate for a crystal that is completely immersed in a beam with a flat, top-hat profile, as was implemented in the software program RADDOSE v1-3 [33, 34, 46].

For routine application, this approximation has obvious limitations in an era when many beams are smaller than the crystals being irradiated and when the beams can have highly featured profiles [47]. A small beam will, under rotation, create a dose hot-spot where the beam and rotation axes intersect, leading to a very large range of dose values where the peak dose can be of one or more orders of magnitude higher than the average dose within the crystal [48]. In order to solve this problem, a weighting scheme – diffraction-weighted dose – has been proposed which aims to give a consistent measure of dose in cases where microbeams lead to uneven dose profiles. Combined with a measure of total elastic scattering, this provides a powerful tool to quantify the relative radiation damage effectiveness of different proposed strategies. 3D models of dose are implemented in the program RADDOSE-3D [49], available at http://raddo.se.

4 Techniques for Merging Data from Multiple Crystals

Microcrystals can diffract X-rays effectively only for a limited amount of time, due to the ensuing radiation damage that, inevitably, degrades their lattices. This is especially true at third-generation synchrotrons, where X-ray brilliance is extremely high. The chances to collect datasets suitable to produce proportionate electron density maps, i.e., datasets with at least 90 % completeness, are very slim. The best option with microcrystals is for a complete dataset to be assembled out of separate short sweeps from individual crystals. The result will be appropriate for electron density calculations only if individual crystals have a reasonable level of isomorphism. For this reason it is essential to carry out data assemblage in a systematic fashion, using computing algorithms and procedures somewhat different from traditional data processing.

A number of independent researchers [50–54] have applied techniques from the field of cluster analysis to the aggregation or separation of data from multiple crystals. Their aim is the production of one or more complete datasets with acceptable merging statistics. Their approaches to the analysis and processing of multiple datasets show similarities and differences. In this section the procedures and algorithms utilized in the highlighted research programs will be reviewed and illustrated with applications to a test case. The emphasis will be put on one of these procedures, coded in *BLEND* [54], a computer program for the analysis and processing of data from multiple crystals. A short summary of the methods and insights brought about by other researchers will follow.

4.1 The Grouping of Multiple Sweeps with Cluster Analysis

Cluster analysis [55] is a well-established technique of multivariate statistics, often applied to the field of data exploration. It is also a statistical research topic in continuous evolution; a good review of this subject can be found in the article by Jain et al. [56]. As the name suggests, the purpose of this technique is to create groups (clusters) out of a given number of individual objects, the grouping being based on similarity criteria. In the so-called hierarchical cluster analysis, individual elements are joined together into progressively larger groups, according to how close to each other the elements are. Proximity between all couples is measured with some kind of generalized geometric distance. In order to measure all distances, it is thus necessary to assign generalized coordinates to the individual objects. These are called *statistical descriptors* and can be any set of variables characterizing each object's features in a unique way. A typical succession of steps in cluster analysis includes estimation of the statistical descriptors, calculation of generalized distance between all couples of elements, accorpation of the closest objects into larger and larger groups, and, finally, the drawing of a so-called dendrogram, a kind of inverted tree where individual elements are like leaves merging into small branches, merging into larger branches, and eventually converging into an all-encompassing trunk. Any node in the dendrogram is associated to a group of elements whose degree of similarity is a function of the generalized distance. There exist many types of distance functions (average distance, minimum distance, maximum distance, etc.), and in cluster analysis these bear on *linkage methods*, because they also control the way in which smaller clusters are merged into larger clusters. The main steps involved in a clustering procedure are sketched in Fig. 2.

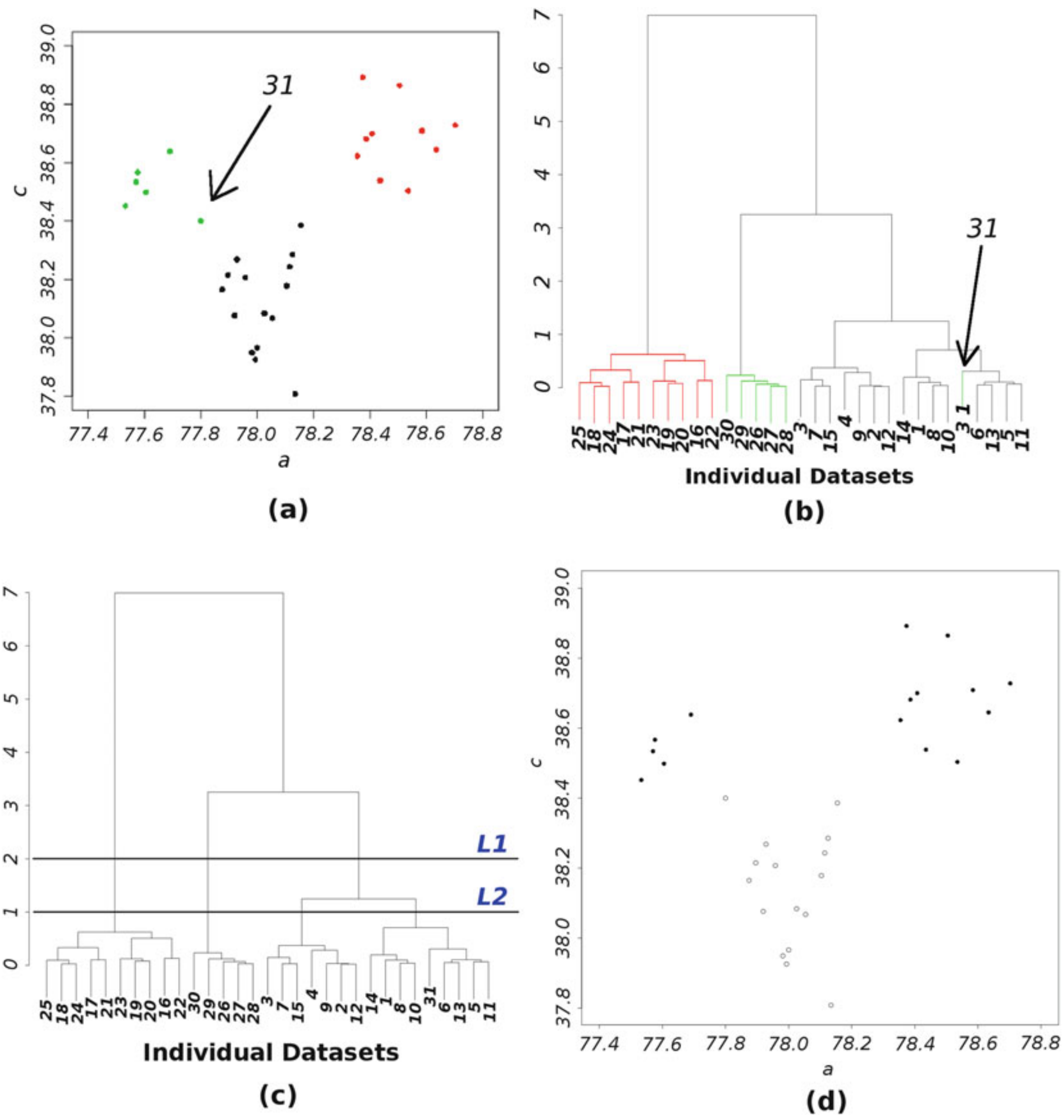


Fig. 2 Main steps in hierarchical clustering. (**a**) A number of objects, in this case 31 sweeps from a multiple-crystal data collection, are assigned statistical descriptors, i.e., parameters to identify them in a multidimensional Cartesian space. In the present example, only two descriptors are used, the *a* and *c* unit cell side lengths for a tetragonal system. From the *a*-*c* plot, it is clear that all objects group in three separate clusters, colored in *black* (sweeps 1–15), *red* (sweeps 16–25), and *green* (sweeps 26–31). Although sweep 31 belongs to the *green* group, its position is equally close to objects in the *black* group; this can have consequences for its inclusion in specific cluster during later analysis. (**b**) Based on the distance between points in the plot and on the specific type of linkage method adopted (ward linkage in this example), individual objects are merged in clusters of increasing size until everything is joined into a single large cluster. The process can be followed in the depicted diagram, known as dendrogram. Individual objects at the *bottom* are joined up initially in groups of two elements and later in groups of three, four, and more elements. The clustering process represented by the dendrogram provides an uncharacterized group of objects with a hierarchical structure; this can be used to reduce greatly the huge set represented by the combination of objects. In the dendrogram shown above, it is clear that the three groups have been correctly identified by the clustering process, with the exception of

As in this section, the focus is on applications of cluster analysis to the analysis of diffraction data; the individual objects to be grouped are single diffraction sweeps collected in a continuous fashion from multiple crystals, from multiple parts of a crystal, or from both. Such sweeps can be by themselves not sufficient to form complete datasets or, as stressed by Liu et al. [50–52], can provide a not strong-enough anomalous signal. It is, therefore, desirable to group them together into larger datasets. Unfortunately, the number of possible groups increases exponentially with the number of individual sweeps. With ten sweeps, for instance, it is possible to form 10 groups of one sweep each, 45 groups of two sweeps each, 120 groups of three sweeps each, 210 groups of four sweeps each, etc. A total of 1023 groups can be obtained with ten sweeps. In general, $2^n - 1$ groups can be formed out of n sweeps. A possible way out of this *combinatorial explosion* is through the formation of clusters based on similarity criteria, using cluster analysis. With hierarchical clustering, for instance, n-1 nodes are formed out of n individual sweeps. This means that only $2n - 1$ datasets will be obtained, a remarkable reduction from the potential $2^n - 1$ datasets.

The statistical descriptors used in BLEND are the crystal cell parameters as measured by integration programs. Their number goes from one for the cubic crystal system (the cell's side length) to six for the triclinic system. Each sweep is characterized by one set of cell parameters. Accordingly, sweeps will be geometrically represented as points in a space with a specific dimensionality, from a minimum of one dimension to a maximum of six dimensions. Next, the Cartesian distance between all couples of points is calculated, and their merging into larger and larger clusters is carried out using the *ward linkage* [55]. Several other types of linkage methods have been tried [54], but the ward linkage has shown an overall better performance, ultimately allowing the emergence of a higher number of datasets with good merging statistics. Other descriptors can be used to measure data similarity. Giordano et al. [53] have implemented a promising procedure using statistical descriptors based on the correlation of scaled intensities. It should be pointed out, though, that a robust scaling is only feasible with sweeps for structures of high-symmetry space groups. Thus this type of descriptors is more effective for complete datasets. In such an instance, the focus is on the increase of data redundancy while

Fig. 2 (continued) sweep 31, wrongly assigned to the black group. This type of error is quite often present in cluster analysis and should be taken care of in the follow-up analysis. (**c**, **d**) Selection of one or more merging groups can be carried out using one or two numerical values for the height of the dendrogram. In the specific example the selected group corresponds to all nodes included between L1 and L2; only one node happens to be included between these levels, the one essentially corresponding to the *black* group (depicted with *white open circle* in the figure) maintaining a good degree of isomorphism, a situation normally met when the need is to increase the anomalous signal. Data handled by BLEND typically span small wedges of reciprocal space and are, thus, unfeasible for direct scaling. As a consequence, only quantities directly derived from the integration process or some simple adaptation of such quantities can be used as statistical descriptors for sweep clustering. So far only cell parameters have shown a clear and consistent propensity to highlight isomorphism among groups of different crystals.

4.1.1 *BLEND*

A diagram of the main components and procedures of the program is shown in Fig. 3. Inputs are integrated (but unscaled) single-sweep data from either MOSFLM [57] or XDS [58]. The correct output to be used from XDS is "INTEGRATE.HKL," rather than "XDS_ASCII.HKL." BLEND can be executed in three modes: *synthesis*, *analysis*, and *combination*. The first run is always in analysis mode; subsequent runs can be either repetitions of the first run in analysis mode, with one or more input parameters modified, or runs in either synthesis or combination modes. During the analysis

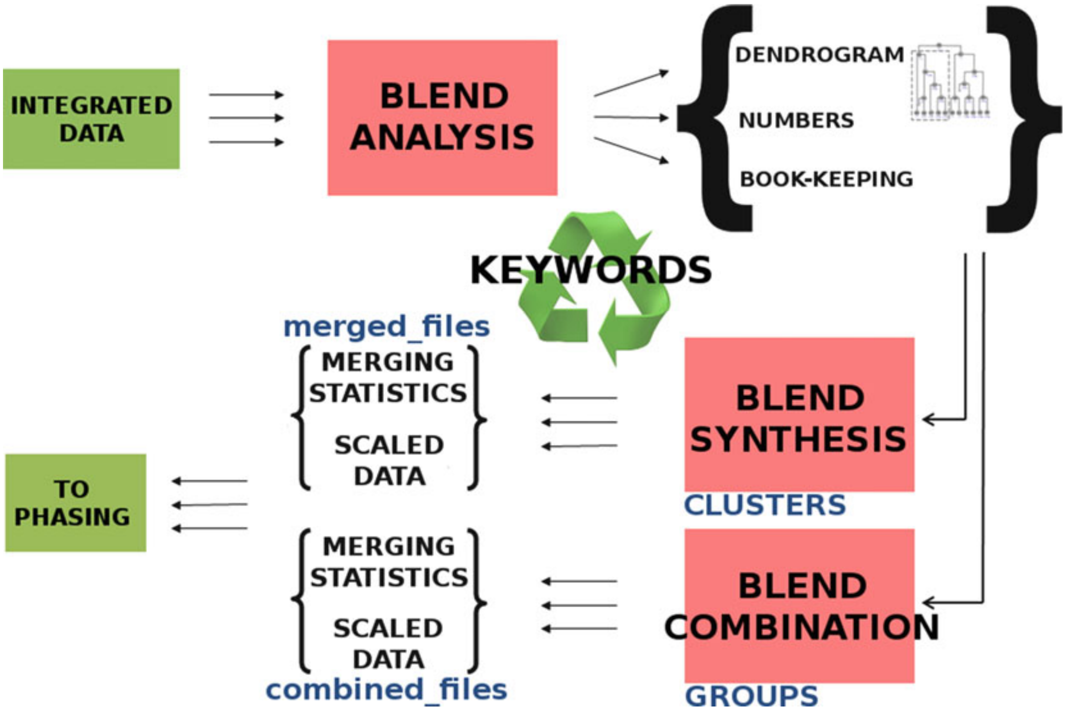


Fig. 3 Schematic diagram of the process flow in *BLEND*. Input to the program consists of integrated but unscaled sweeps of data from either MOSFLM or XDS. BLEND can be executed in three modes: analysis, synthesis, and combination. Output from the analysis is the dendrogram, a few numeric tables and bookkeeping files for subsequent runs. Synthesis and combination modes can be executed as many times as desired after BLEND has been executed at least once in analysis mode. The final output consists of scaled files in mtz format, log files, and tables of merging statistics

mode, BLEND examines all individual input files and discards those containing either multiple sweeps or formatting errors or both. Next, BLEND extracts statistical descriptors and carries out cluster analysis. The output is here formed by a series of ASCII files with tabulated data from all accepted sweeps, a dendrogram in both graphical and text forms and a binary file with information needed for all runs in synthesis or combination modes. During the analysis pass, the program also calculates a single parameter supposed to provide an intuitive measure of non-isomorphism in the group of crystals investigated. This is called linear cell variation (LCV) [54] and essentially measures the largest variation of the diagonals on the three crystal cell faces, across all crystals under study. It has been observed empirically that values of LCV around 1.5 % or less correspond to non-noticeable structural changes.

After having executed BLEND in analysis mode, the user can decide whether to carry out scaling of specific clusters by providing one or two numeric values for the height in the dendrogram. This is then executed by BLEND in synthesis mode. More in-depth analysis of results from the synthesis run might point to specific sweeps or groups of sweeps that do not perform well but, rather, deteriorate merging statistics. In such cases it is possible to execute BLEND in combination mode, where sweeps to be combined do not necessarily belong to a same node of the dendrogram. Executions of the program tailored to more specific needs are provided simply by adding or changing keywords in a keyword file. Output from synthesis or combination runs is collected in directories "merged_files" or "combined_files," respectively. This consists of all log files from POINTLESS and AIMLESS, mtz files for pre- and post-scaling jobs, an ASCII file with the original content of each group of sweeps, an ASCII file with overall merging statistics tabulated, and a plot of R_{meas} vs. Completeness. The user can examine results from any specific group of files either by direct inspection or with CCP4 [59] tools like the program LOGGRAPH.

BLEND can easily be executed with just two command lines. At the same time it has a good amount of flexibility where it enables the user to repeat specific executions in a different way, simply by changing, adding, or deleting certain keywords.

4.2 An Example

A few examples on the use of BLEND are illustrated in reference [54]. In order to target the most important aspects of processing from multiple crystals, here, we will deal with the combination of two sets of data from tetragonal lysozyme. These have been prepared separately for different purposes, but similarities in their cell parameters make them suitable for a well-informed analysis with BLEND. The first set of 17 crystals were soaked in a sodium bromide solution; data were subsequently collected at a wavelength close to the bromide edge, in order to increase the anomalous signal for SAD phasing. The second set of 12 crystals was collected directly from crystallization plates at room temperature; here no heavy atoms have been added. Both were not microcrystal sets, but all data were limited to the first 50 images or, equivalently, to only 10-degree rotation sweeps. For this reason, only combination of crystals can provide a complete dataset, as it normally happens with microcrystals. The space group is P 4_3 2_1 2 and data resolution was limited to 2 Å. A summary of these 29 sweeps, including individual completeness, is provided in Table 1, where the first group, termed G*roup A*, has numbers 1–17 and the second, termed G*roup B*, numbers 18–29.

When BLEND is executed in analysis mode on the above 29 sweeps, the dendrogram of Fig. 4a is returned. Crystals 7 and 9 are very different from the rest and have been considered no further. These crystals are responsible for the high value (13.98 %) of LCV. This drops to 2.20 % when they are removed from the analysis. The other crystals group in two main clusters, essentially corresponding to Group A and Group B. The only exception is crystal 12, falling within the group of lysozyme in plates while belonging to the group of lysozyme with bromide. Why this happens is perfectly understandable if one looks at a plot of cell side *a* vs. cell side c (Fig. 4b); it just occurs that crystal 12 has a size slightly different from the size of the other crystals in the group; the clustering process with ward linkage, then, facilitates the absorption of this crystal in Group B, rather than Group A. This is not a problem for data processing in BLEND because, as we shall shortly see, crystal 12 will be discarded while assembling complete datasets.

Next, BLEND is executed in synthesis mode over all the nodes of the dendrogram by using:

```
Blend -s 20
```

Twenty is a number higher than the last merging height of the dendrogram; any other number higher than 20 would serve the same purpose. The result of this synthesis job is shown in Fig. 4c, where only clusters corresponding to datasets with completeness of around 90 % or better are included and where the corresponding R_{meas} value is typed jointly with the cluster number. Datasets represented by nodes 16, 19, 18, 20, and 12 should yield better-quality electron density maps than those represented by nodes 21, 22, 23, 24, 25, and 27, because they are composed of a smaller number of sweeps and inherently including more isomorphous crystals. But more sweeps mean higher redundancy; this, in turn, yields higher signal-to-noise ratio, higher anomalous signal, and less biased structure factors. Ultimately, the choice of the best merged dataset is a matter of balance between these factors and the need of obtaining sufficient completeness and good merging statistics. Before using any of the merged complete datasets,

Table 1
Cell parameters and completeness from datasets of 29 crystals of tetragonal lysozyme (space group P 4_3 2_1 2)

Crystal number	*a*	*c*	Completeness (%)
1	78.277	38.076	50.3
2	78.238	38.116	36.7
3	78.160	38.004	51.0
4	78.072	37.620	44.8
5	78.494	37.785	56.3
6	78.376	37.715	55.3
7	88.990	42.699	31.5
8	78.439	37.798	46.4
9	82.691	41.061	32.6
10	78.613	37.753	42.4
11	78.341	37.621	59.2
12	79.792	37.788	49.5
13	78.168	37.301	54.9
14	78.343	37.477	54.5
15	78.298	37.651	48.4
16	78.289	37.689	52.2
17	78.293	37.623	56.0
18	78.949	38.416	43.0
19	79.162	38.461	40.6
20	78.595	38.664	44.8
21	78.706	38.355	43.2
22	79.051	38.431	41.8
23	78.784	38.517	40.9
24	78.750	38.480	36.8
25	78.989	38.695	37.3
26	78.961	38.392	41.4
27	78.940	38.455	41.8
28	79.161	38.600	24.9
29	78.826	38.436	41.1

Each dataset includes 10-degree rotation sweeps (50 images of 0.2° each). The first 17 sweeps belong to frozen crystals soaked in a sodium bromide solution (the used wavelength was closed to the bromide edge); the last 12 sweeps were collected from a crystallization plate at room temperature. No heavy atoms had been added to these last 12 crystals. From the completeness column, it is very clear that more sweeps are needed to form a complete dataset

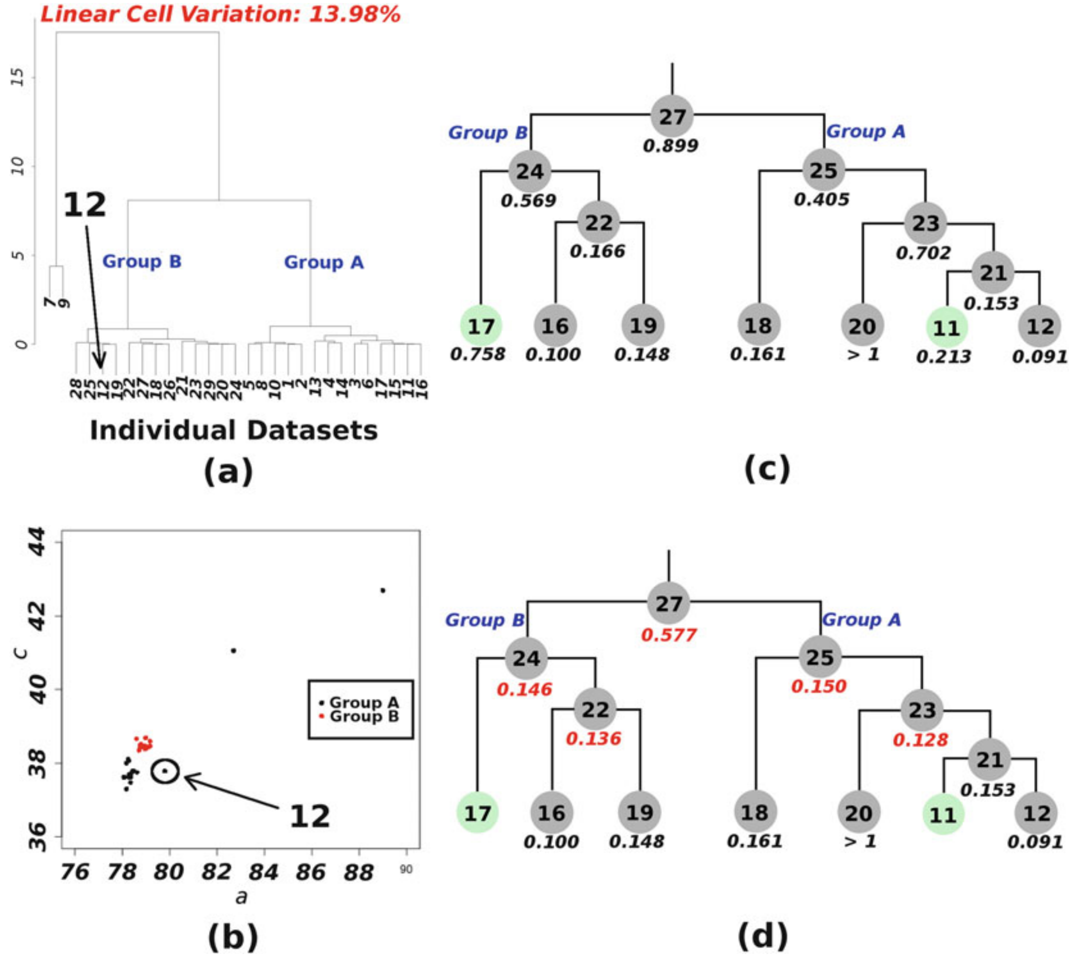
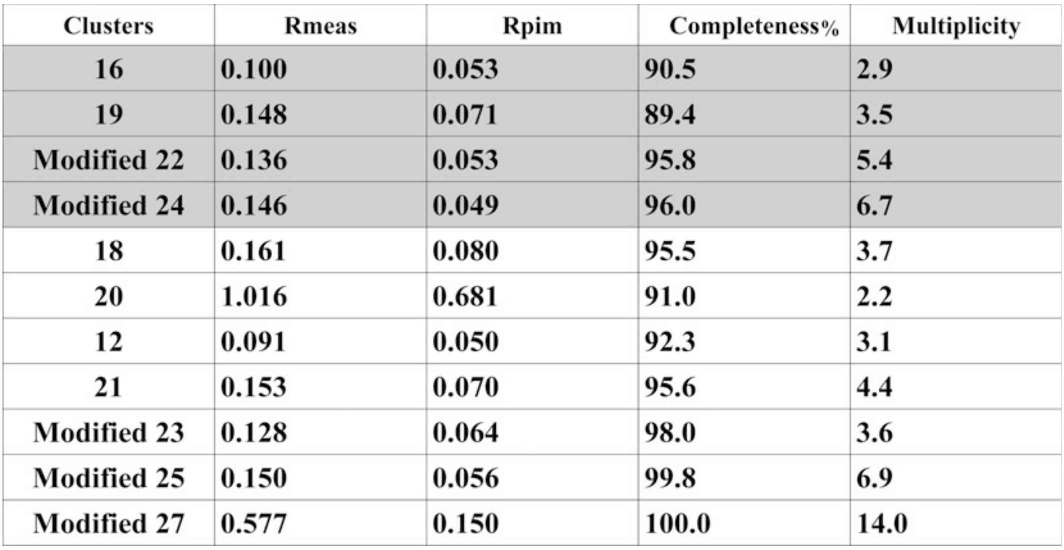


Fig. 4 Data processing with *BLEND* for the lysozyme test case. The 29 sweeps investigated come from two groups of crystals. The first, Group A, includes crystals of lysozyme soaked in bromide solution (sweeps 1–17); the second, Group B, includes crystals of lysozyme with no bromide, but where data were collected in situ at room temperature (sweeps 18–20). (**a**) Clustering is able, in this case, to separate the two groups nearly completely in two large clusters, with the exception of sweeps 7, 9, and 12. (**b**) It is clear from the *a*-*c* plot that 7 and 9 are outlier crystals, while 12 is somewhat off the average value of crystals in Group A and, therefore, has been absorbed in Group B. (**c**) *BLEND* processing has been furthered with datasets having completeness around 90 % or more, with the aim to improve merging statistics. Clusters 17 and 11 are less than 90 % complete, but have been included in the picture because in the following analysis some of the sweeps in these clusters are used. (**d**) All numbers in *red* are improved R_{meas} values obtained when filtering out some sweeps from the composing clusters. The two main clusters, modified cluster 24 corresponding to Group B and modified cluster 25 corresponding to Group A, have respectable merging statistics. Their union into the modified cluster 27, though, presents a high value of R_{meas}, pointing at some non-isomorphism between the two clusters

though, a process of data improvement can be carried out through filtering of bad sweeps and combination of sweeps from different clusters, using BLEND combination mode. For example, by trying out a few combinations of sweeps from both clusters 16 and 19, we soon discover that sweep 23 actually is the only one to deteriorate merging statistics for all resulting datasets. Thus, the R_{meas} for cluster 22 improves from 0.166 to 0.136 if we remove sweep 23. Or cluster 24 improves R_{meas} from 0.569 to 0.146 when, as said before, sweep 12 is removed from cluster 17, and this is merged with the modified cluster 22. A few more datasets can be improved in this way; these have R_{meas} typed in red in Fig. 4d. Final statistics for all groups is shown in Table 2.

From the analysis just carried out, two complete and redundant datasets with respectable statistics can be obtained, modified clusters 24 and 25, respectively, corresponding to Group B and Group A. Running molecular replacement jobs with PHASER [60], using chicken egg white lysozyme as model (PDB code 1AZF, stripped of all waters and bromides), followed by rigid body refinement and restrained refinement with REFMAC [61] and automated addition of waters with COOT [62], yields two structures which are spatially close but do not overlap (Fig. 5a). This is, obviously, to be expected, because we know that crystals in Group A were prepared differently from crystals in Group B. Quite amazingly, the same procedure carried out on data from cluster 27 produces an interpretable electron density map (Fig. 5b), despite the poor merging statistics for this data set.

Table 2
Final merging statistics for the clusters selected by *BLEND*



Clusters	Rmeas	Rpim	Completeness%	Multiplicity
16	0.100	0.053	90.5	2.9
19	0.148	0.071	89.4	3.5
Modified 22	0.136	0.053	95.8	5.4
Modified 24	0.146	0.049	96.0	6.7
18	0.161	0.080	95.5	3.7
20	1.016	0.681	91.0	2.2
12	0.091	0.050	92.3	3.1
21	0.153	0.070	95.6	4.4
Modified 23	0.128	0.064	98.0	3.6
Modified 25	0.150	0.056	99.8	6.9
Modified 27	0.577	0.150	100.0	14.0

Some of the clusters have been modified by subtraction of specific sweeps. These are the result of trials with the program in combination mode. These final statistics greatly improve those of the initial clustering. Datasets inside the *gray-shaded* area belong to Group B; the rest belong to Group A (See main text)

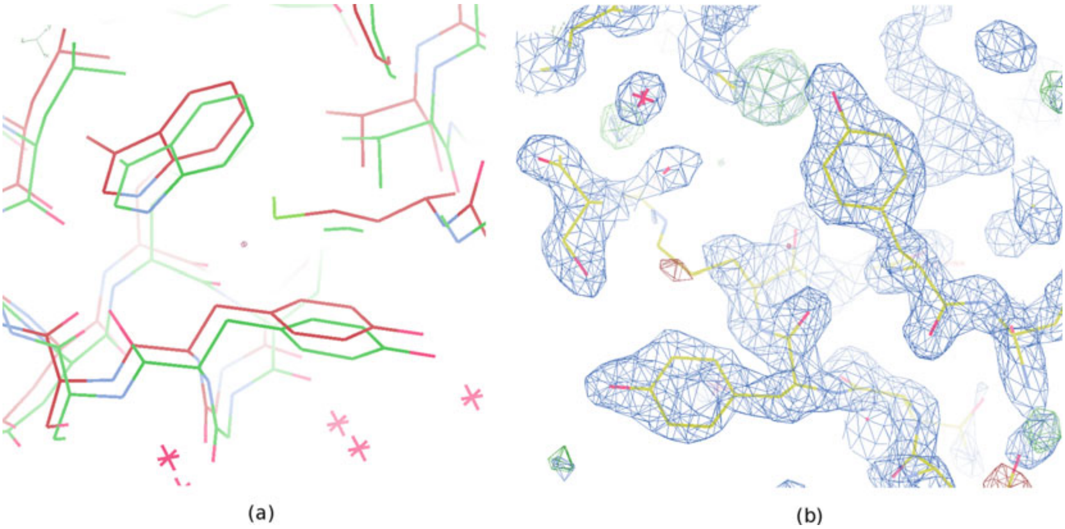


Fig. 5 (**a**) The two datasets corresponding to modified clusters 24 and 25 (Group B, *red bonds*, and Group A, *green bonds*, respectively) yield two non-isomorphous structures. Their non-isomorphism, in this case, is given by a difference in packing (displaced chains) and a difference in the inclusion or absence of bromide atoms. (**b**) The dataset obtained by merging the two clusters just described into modified cluster 27, although the union of two non-isomorphous datasets still shows a very interpretable electron density. Ambiguities obviously arise around heavy atom sites, like the empty one shown at the *top* of this map

4.3 Alternative Procedures

While BLEND deals with unscaled sweeps, other groups researching data processing with multiple crystals prefer to use complete or almost complete datasets to increase multiplicity and, thus, the strength of the anomalous signal in those cases where it would otherwise be weak and mostly undetectable. This is what, for instance, Liu et al. [2, 50, 52] accomplish when phasing both heavy-atom derivatives and native proteins through anomalous signal from sulfur. In their work the accent is in the elimination of non-isomorphism either by working at low resolutions (e.g., 3 Å) or through the use of hierarchical cluster analysis with the single linkage method to monitor isomorphism and, eventually, get rid of unwanted outliers. Giordano et al. [53] achieve and confirm similar results using the average linkage method. Their use of cluster analysis seems to be more systematic and, similarly to what happens in BLEND, is meant to provide a tool for the selection of isomorphous groups. Both Liu et al. [2, 50, 52] and Giordano et al. [53] can use alternative descriptors to cell parameters because the use of complete datasets allows scaling and, accordingly, the introduction of accurate intensities. One of the descriptors used by both research groups, for example, relates to correlation coefficients between intensities at various resolutions. In BLEND there is an option to use descriptors based on unscaled intensity averages in shells of resolution; this is not the default, though, because of a lack of systematically consistent results.

A somewhat different approach to multiple crystal merging is provided by Hanson et al. [63]. Their main aim is to gradually include groups of reflections, rather than whole wedges of data, in a controlled way, in an attempt to filter out radiation damaged intensities. The selection is guided by continuous reference to a medium-resolution complete dataset, collected at low exposure from one of the crystals. Groups of reflections corresponding approximately to rotation of 1° are included or rejected from the main set according to the correlation between estimated and measured peak profiles being above or below a threshold. A second threshold, this time for R merge, commands the acceptance or rejection, in the final dataset, of all reflections previously selected. In case of rejection the first threshold is increased by 5 %, and the whole process is repeated until a group of reflections manage to overcome both thresholds or until all reflections from the specific wedge of data have been examined. With this method Hanson et al. [63] have been able to build a 97.2 % complete data set at 2.8 Å resolution, with 90 % completeness in the last resolution shell.

5 The Strategy for Collecting Data from One Tiny Crystal

In this section, in contrast to the previous one in which we discussed the merging of data from multiple crystals (Sect. 4), we will describe the challenge of collecting a complete dataset from a tiny protein crystal using a microfocused beam.

As described in the introduction of this section, severe radiation damage can be avoided by changing irradiation points during data collection. But what is the extent we should translate for changing irradiation points when using a 100-μm crystal and a 10-μm beam? How should we determine the suitable step length? As in normal experiments, the amount of absorbed dose should be estimated using RADDOSE (Sect. 3). Furthermore, there are additional difficulties when considering the quantity of radiation damage when images are collected from multiple irradiation points. The major problem is the propagation length of radiation damage, referred to as PLRD.

Before describing PLRD, an important experimental method to mitigate radiation damage should be noted. This method is known as “helical data collection” and was proposed by Flot et al. [4]. Normal data collection is performed using only a rotation axis, and helical data collection utilizes additional translation axes for changing the irradiation points. Normally, during the first step of helical data collection, a three-dimensional vector is defined along the longest axis of crystal shape. Irradiation points are distributed along the vector with the same pitch. The name of this method is due to the helical movements of the crystal on the goniometer during data collection. In the study published by Flot et al. [4], a more important concept is included in the proposed "helical data collection" method, which involves maximizing the number of irradiation points during data collection even if the translation step is smaller than the beam footprint. This is done to equalize the radiation damage at each point. For example, it is more appropriate to set the translation step between each irradiation points as 1 μm, even when the beam size is 10 μm. This equalizes the amount of radiation damage at each irradiation point, as described by Flot et al. [4]. A plot of the relative B factors of the frames during the scaling process shows a relatively "flat" shape, compared with the non-helical process. Relative B factor is a good indicator of radiation damage [45]. A flat series of relative B factors clearly shows that the method is capable of mitigating severe radiation damage using a microfocused beam.

Here, we consider the details of accumulated radiation damage in the helical data collection from a theoretical perspective. Let us imagine the case when we are aware of the precise PLRD of the protein crystal. The beam profile normally shows Gaussian shape; then, the propagation of the radiation damage could be Gaussian shaped. Along this assumption, the following things are considered. The distribution of radiation damage is described as "intensity decay" of the diffracting power of a crystal. Here PLRD is defined as the full-width half-maximum value of a function of the intensity decay. This decay curve of diffracting power, referred to as the DCDP, can be used to estimate radiation damage during helical data collection (Fig. 6a). First, the integrated area of this decay curve is set to 1.0. The total amount of decay of a crystal is regarded as 1.0 after exposure. Figure 6b describes a simulated calculation of the accumulated radiation damage in a virtual crystal during helical data collection. For the first exposure, the area of 1.0 is accumulated in the crystal. For the second exposure, simple accumulation of the DCDP is convoluted onto the first one after the exposure position is translated as the crystal is moved by 0.5 μm. This corresponds to the exposure after changing the irradiation point. Performing this process sequentially, in the same manner, can reveal an interesting curve of accumulated damage to the crystal. The graph clearly illustrates that the helical data collection method shows a flat region of radiation damage area to crystal after some translation. The height of the plateau region, representing the maximum radiation damage in this virtual experiment, can be estimated by knowing the precise PLRD and step length between each irradiation point. For example, PLRD reported by Sanishvili et al. [64] corresponded to 2.6 μm × 3.1 μm (H × V) for an energy of 15.1 keV and a beam size of 1.16 μm × 1 μm (H × V). Then, when this PLRD is used for a 1.0-μm-step helical data collection, the height of the plateau region is estimated to be 0.58 using the simulated calculation. This value corresponds to 0.5 when an absorbed dose is 20 MGy. Then, you can control the absorbed dose by setting wider helical step length or by setting

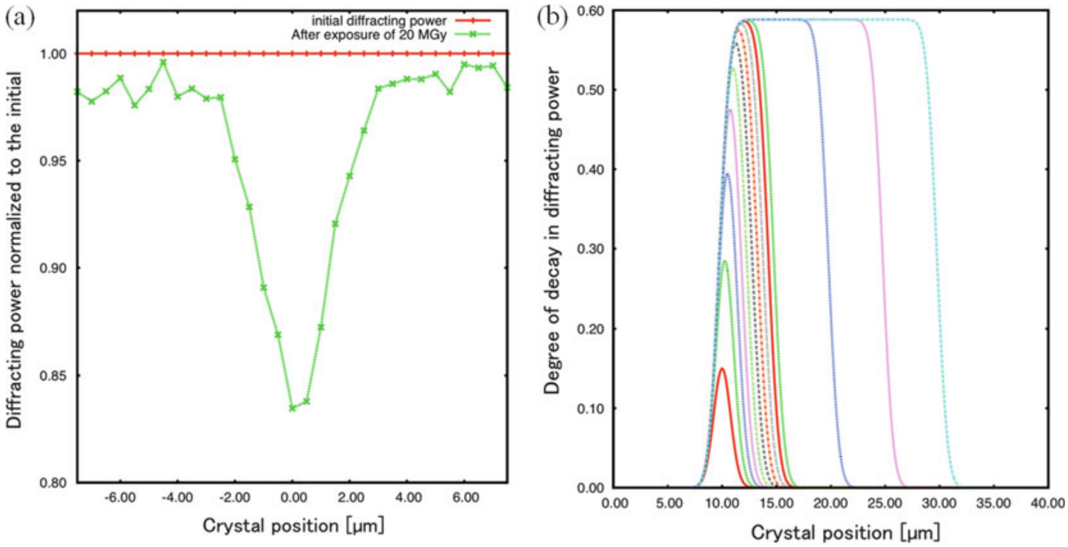


Fig. 6 (**a**) Decay curve of diffracting power (DCDP) observed on the frozen lysozyme crystal by using 1 × 10 μm beam. The peak height, negative value, roughly corresponds to 0.16, and FWHM of DCDP is around 2.5 μm. This can be modeled as a Gaussian distribution. (**b**) Modeled Gaussian function can be utilized for a simulation of "helical data collection." This figure is an example calculation by using 1 × 10-μm beam and 0.5-μm-step helical. The peak height of DCDP and FWHM are set to 0.16 and 2.5 μm, respectively. The first beam is exposed to the crystal position of 10.0 μm, and the first decay curve is summed to the position. The second exposure is on to 0.5 μm right, and the decay value from the model is summed up. This summation is repeatedly conducted until the height of summed decay reaches to the flat region. For this helical data collection, diffracting power of the crystal finally reduces to 0.58 times in entire crystal volume

thicker attenuator. In the simple case, determining the data collection strategy during helical data collection requires the beam size, the crystal size, the X-ray energy, and the intended absorbed dose. The degree of radiation damage can be quantified by a factor compared with the normalized DCDP, 1.0. Measurement of PLRD is also well described by Sanishvili et al. [64]. PLRD in the crystal depends on both the beam size and the energy of the X-rays used. The effect of the X-ray energy is discussed in the following section. Using various beam sizes and energies, PLRD can be determined as a function. The function can be utilized for the estimation of maximum radiation damage for the data collection strategy. Strategy software for helical data collection, named KUMA, has been developed at BL32XU. This software can simulate the radiation damage as previously described. It suggests attenuation factors for each exposure after the user inputs all the required information. Moreover, it enables initial phase determination using tiny membrane protein crystals generated by the lipidic-cubic phase method [65–68].

6 High-Energy Crystallography

The advantages of using X-ray whose energy is higher than 20 keV for MX data collection have been discussed for a long time because a reduction in radiation damage and an improvement in data accuracy could be expected [69, 70]. However, the high-energy X-rays are not widely used. One possible reason is that the cryo-technique began to be routinely used to reduce radiation damage. Another reason is that high-energy X-rays can be used at only at a limited number of beamlines, such as BL41XU at SPring-8 [71].

In the middle of 2000s, attention was drawn to the use of high-energy X-rays because of its usefulness for microcrystallography [72]. Nave and his colleagues used Monte Carlo simulation to track the photoelectrons generated by the irradiation of X-rays [73, 74]. The simulation showed that photoelectrons spread a few micrometers from the original position and escaped from the illuminated volume if the crystal size was small enough. Thus, less energy was deposited in the illuminated volume. This behavior became more significant when higher-energy X-rays were used because higher-energy photoelectrons had a longer spatial spread of trajectories.

The mitigation of radiation damage by photoelectron escape was experimentally exploited by Sanishvili et al. [64]. They measured the damage rate, which is the reduction of diffraction intensity per absorbed dose, using 18.5 keV X-rays of size 1–100 μm. The result showed that the damage rate became smaller as the beam size decreased, if the beam size was smaller than 10 μm. They also observed that the radiation damage propagated to 4 μm by probing the reduction of diffraction power around a central "burn" position. These experimental results revealed that photoelectron escape reduced radiation damage in the illuminated volume.

Fourme et al. [75] investigated the optimum X-ray energy for data collection using both experimental data and simulation. They used the experimental data reported by Shimizu et al. [76], which systematically investigated the energy dependency of radiation damage and concluded that radiation damage solely depended on the absorbed dose regardless of the X-ray energy. Fourme normalized the number of datasets that Shimizu had collected with X-ray energies of 6.5–33 keV by dividing by the detective quantum efficiency of the detector. These results indicated that if a detector had the same efficiency at all X-ray energies, one could collect a larger number of data set using higher energy. They also calculated the intensity-to-dose ratio, I/D, using Mote Carlo simulations, postulating crystal sizes of 1–100 μm an X-ray energy range of 5–80 keV. The results showed that the optimum energy, which maximizes I/D, is located between 24 and 41 keV, depending on the crystal size. The optimum energy of a large crystal was higher than that of a small crystal and the energy dependency of a large crystal was less significant compared with that of a small crystal.

They explained that this behavior of I/D was caused by a decrease in X-ray absorption and photoelectron escape.

Although the advantages of high energy have been described above, it also has some disadvantages. One of the disadvantages is the reduction in diffraction intensity, which is inversely proportional to the square of the X-ray energy. However, it can be overcome by using a high-intensity synchrotron beamline. The other disadvantage is the detector efficiency. The detector used at an MX beamline has an optimum efficiency at around 12.4 keV, and it has lower efficiency for high-energy X-rays because of the large transmission. Therefore, development of a new detector suitable for high-energy X-rays is essential to completely utilize the merits.

7 XFEL New Source for Protein Microcrystallography

The X-ray free-electron laser (XFEL) is a new X-ray source for protein crystallography. The development of XFELs, which have extremely high-peak intensity and ultrashort pulse duration, allows data collection from biological samples with significantly reduced radiation damage. Radiation damage is presently a major factor limiting the attainable resolution in the imaging of biological materials, particularly when using X-rays.

The first structure obtained at an XFEL facility was of photosystem-I, large membrane protein complex. Its structure was revealed at the Linac Coherent Light Source at Stanford [77]. Utilizing the repetition of pulsed X-rays, 30 Hz, in this report, the new method "serial femtosecond crystallography" was established. Many micron-sized crystals flowed into the XFEL path by a liquid jet system, and diffraction patterns are detected with a high frame rate CCD detector. The use of this method, which is now generally used for protein crystallography at the XFEL facility, has resulted in the output of some innovative phenomena that cannot be acquired at a synchrotron facility [78, 79]. Using the new source has presented a number of challenges for developing the required hardware and software. At the Japanese XFEL facility, named SACLA (SPring-8 Angstrom free-electron LAser), serial femtosecond crystallography and radiation damage-free structural analysis are also conducted. Radiation damage-free structural analysis using large frozen protein crystals overcomes the resolution limit, compared with SFX, in which smaller crystals are utilized to flow the sample crystal into a capillary with a diameter of few tens of microns [80, 81]. Recently, time-resolved protein crystal structure analyses had been reported [82, 83]. Recent impacts on structural biology from XFEL are reviewed on the papers [84, 85].

XFEL will provide considerable information for structural biology by making the maximum use of its ultrashort pulse and high brilliance.

References

1. Smith JL, Fischetti RF, Yamamoto M (2012) Micro-crystallography comes of age. Curr Opin Struct Biol 22:602–612
2. Liu Q, Zhang Z, Hendrickson WA (2010) Multi-crystal anomalous diffraction for low-resolution macromolecular phasing. Acta Crystallogr D Biol Crystallogr 67:45–59
3. Riekel C, Burghammer M, Schertler G (2005) Protein crystallography microdiffraction. Curr Opin Struct Biol 15:556–562
4. Flot D, Mairs T, Giraud T et al (2010) The ID23-2 structural biology microfocus beamline at the ESRF. J Synchrotron Rad 17: 107–118
5. Tanaka H, Adachi M, Aoki T et al (2006) Stable top-up operation at SPring-8. J Synchrotron Rad 13:378–391
6. Yamauchi K, Mimura H, Inagaki K, Mori Y (2002) Figuring with subnanometer-level accuracy by numerically controlled elastic emission machining. Rev Sci Instrum 73: 4028
7. Yumoto H, Mimura H, Matsuyama S et al (2005) Fabrication of elliptically figured mirror for focusing hard x rays to size less than 50 nm. Rev Sci Instrum 76:063708
8. Flot D, Gordon EJ, Hall DR et al (2005) The care and nurture of undulator data sets. Acta Crystallogr D Biol Crystallogr 62:65–71
9. Igarashi N, Ikuta K, Miyoshi T et al (2008) X-ray beam stabilization at BL-17A, the protein microcrystallography beamline of the photon factory. J Synchrotron Rad 15:292–295
10. Groves MR, Müller IB, Kreplin X, Müller-Dieckmann J (2007) A method for the general identification of protein crystals in crystallization experiments using a noncovalent fluorescent dye. Acta Crystallogr D Biol Crystallogr 63:526–535
11. Gill HS (2010) Evaluating the efficacy of tryptophan fluorescence and absorbance as a selection tool for identifying protein crystals. Acta Cryst F66:364–372
12. Madden JT, DeWalt EL, Simpson GJ (2011) Two-photon excited UV fluorescence for protein crystal detection. Acta Cryst D67: 839–846
13. Hirata K, Kawano Y, Ueno G et al (2013) Achievement of protein micro-crystallography at SPring-8 beamline BL32XU. J Phys Conf Ser 425:012002
14. Broennimann C, Eikenberry EF, Henrich B et al (2006) The PILATUS 1M detector. J Synchrotron Rad 13:120–130
15. Hasegawa K, Hirata K, Shimizu T et al (2009) Development of a shutterless continuous rotation method using an X-ray CMOS detector for protein crystallography. J Appl Cryst 42: 1165–1175
16. Ben-Shem A, Garreau de Loubresse N, Melnikov S et al (2011) The structure of the eukaryotic ribosome at 3.0 A resolution. Science 334:1524–1529
17. Sanishvili R, Nagarajan V, Yoder D et al (2008) A 7 microm mini-beam improves diffraction data from small or imperfect crystals of macromolecules. Acta Cryst D64:425–435
18. Hikima T, Hashimoto K, Murakami H et al (2013) 3D manipulation of protein microcrystals with optical tweezers for X-ray crystallography. J Phys Conf Ser 425:012011
19. Wagner A, Duman R, Stevens B, Ward A (2013) Microcrystal manipulation with laser tweezers. Acta Cryst D69:1297–1302
20. Soares AS, Engel MA, Stearns R et al (2011) Acoustically mounted microcrystals yield high-resolution X-ray structures. Biochemistry 50: 4399–4401
21. Bingel-Erlenmeyer R, Olieric V, Grimshaw JPA et al (2011) SLS crystallization platform at beamline X06DA – a fully automated pipeline enabling in Situ X-ray diffraction screening. Cryst Growth Des 11:916–923
22. Jacquamet L, Ohana J, Joly J et al (2004) A new highly integrated sample environment for protein crystallography. Acta Cryst D60: 888–894
23. le Maire A, Gelin M, Pochet S et al (2011) In-plate protein crystallization, in situ ligand soaking and X-ray diffraction. Acta Cryst D67: 747–755
24. Landau EM, Rosenbusch JP (1996) Lipidic cubic phases: a novel concept for the crystallization of membrane proteins. Proc Natl Acad Sci U S A 93:14532–14535
25. Hilgart MC, Sanishvili R, Ogata CM et al (2011) Automated sample-scanning methods for radiation damage mitigation and diffraction-based centering of macromolecular crystals. J Synchrotron Rad 2011:18
26. Cherezov V, Hanson MA, Griffith MT et al (2009) Rastering strategy for screening and centring of microcrystal samples of human membrane proteins with a sub-10 m size X-ray synchrotron beam. J R Soc Interface 6: S587–S597
27. Zhang Z, Sauter NK, van den Bedem H et al (2006) Automated diffraction image analysis

and spot searching for high-throughput crystal screening. J Appl Cryst 39:112–119

28. Stepanov S, Makarov O, Hilgart M et al (2011) JBluIce-EPICS control system for macromolecular crystallography. Acta Cryst D67:176–188
29. Rasmussen SGF, Choi H-J, Rosenbaum DM et al (2007) Crystal structure of the human β2 adrenergic G-protein-coupled receptor. Nature 450:383–387
30. Nave C, Garman EF (2005) Towards an understanding of radiation damage in cryocooled macromolecular crystals. J Synchrotron Rad 12:257–260
31. Garman EF (2010) Research papers. Acta Cryst D66:339–351
32. Holton JM, Frankel KA (2010) The minimum crystal size needed for a complete diffraction data set. Acta Cryst D66:393–408
33. Paithankar KS, Owen RL, Garman EF (2009) Absorbed dose calculations for macromolecular crystals: improvements to RADDOSE. J Synchrotron Rad 16:152–162
34. Paithankar KS, Garman EF (2010) Know your dose: RADDOSE. Acta Cryst D66:381–388. doi:10.1107/S0907444910006724 1–8
35. Weik M, Ravelli RB, Silman I et al (2001) Specific protein dynamics near the solvent glass transition assayed by radiation-induced structural changes. Protein Sci 10:1953–1961
36. Leal RMF, Bourenkov GP, Svensson O et al (2011) Experimental procedure for the characterization of radiation damage in macromolecular crystals. J Synchrotron Rad 18:381–386. doi:10.1107/S0909049511002251 1–6
37. Owen RL, Axford D, Nettleship JE et al (2012) Research papers. Acta Cryst D68:810–818. doi:10.1107/S0907444912012553 1–9
38. Owen RL, Yorke BA, Gowdy JA, Pearson AR (2011) Revealing low-dose radiation damage using single-crystal spectroscopy. J Synchrotron Rad 18:367–373
39. Corbett MC, Latimer MJ, Poulos TL et al (2007) Research papers. Acta Cryst D63: 951–960
40. Burmeister WP (2000) Structural changes in a cryo-cooled protein crystal owing to radiation damage. Acta Cryst D56:328–341
41. Ravelli RB, McSweeney SM (2000) The "fingerprint" that X-rays can leave on structures. Structure 8(3):315–328
42. Weik M, Ravelli RB, Kryger G et al (2000) Specific chemical and structural damage to proteins produced by synchrotron radiation. Proc Natl Acad Sci U S A 97:623–628
43. Ramagopal UA, Dauter Z, Thirumuruhan R et al (2005) Radiation-induced site-specific damage of mercury derivatives: phasing and implications. Acta Cryst D61:1289–1298
44. Owen RL, Rudiño-Piñera E, Garman EF (2006) Experimental determination of the radiation dose limit for cryocooled protein crystals. Proc Natl Acad Sci U S A 103:4912–4917
45. Kmetko J, Husseini NS, Naides M et al (2006) Quantifying X-ray radiation damage in protein crystals at cryogenic temperatures. Acta Crystallogr D Biol Crystallogr 62:1030–1038
46. Murray JW, Garman EF, Ravelli RBG (2004) X-ray absorption by macromolecular crystals: the effects of wavelength and crystal composition on absorbed dose. J Appl Crystallogr 37:513–522
47. Krojer T, von Delft F (2011) Assessment of radiation damage behaviour in a large collection of empirically optimized datasets highlights the importance of unmeasured complicating effects. J Synchrotron Rad 18: 387–397
48. Zeldin OB, Gerstel M, Garman EF (2013) Optimizing the spatial distribution of dose in X-ray macromolecular crystallography. J Synchrotron Rad 20:49–57
49. Zeldin OB, Gerstel M, Garman EF (2013) RADDOSE-3D: time- and space-resolved modelling of dose in macromolecular crystallography. J Appl Cryst 46:1225–1230
50. Liu Q, Zhang Z, Hendrickson WA (2011) Multi-crystal anomalous diffraction for low-resolution macromolecular phasing. Acta Crystallogr D Biol Crystallogr 67:45–59
51. Liu Q, Dahmane T, Zhang Z et al (2012) Structures from anomalous diffraction of native biological macromolecules. Science 336:1033–1037
52. Liu Q, Liu Q, Hendrickson WA (2013) Robust structural analysis of native biological macromolecules from multi-crystal anomalous diffraction data. Acta Crystallogr D Biol Crystallogr 69:1314–1332
53. Giordano R, Leal RMF, Bourenkov GP et al (2012) Research papers. Acta Cryst D68: 649–658
54. Foadi J, Aller P, Alguel Y et al (2013) Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 69:1617–1632
55. Everitt BS, Landau S, Leese M, Stahi D (2011) Cluster analysis, 5th edn. Wiley, New York
56. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv (CSUR) 31(3):264–323
57. Leslie AGW, Powell HR (2007) Processing diffraction data with MOSFLM. Evolving

Methods Macromol Crystallogr 245:41–45, 978-1-4020-6314-5

58. Kabsch W (2010) Integration, scaling, space-group assignment and post-refinement. Acta Cryst D66:133–144
59. Winn MD, Ballard CC, Cowtan KD et al (2011) Overview of the CCP4 suite and current developments. Acta Cryst D67:235–242
60. McCoy AJ, Grosse-Kunstleve RW, Adams PD et al (2007) Research papers. J Appl Cryst 40:658–674
61. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst D53:240–255
62. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Research papers. Acta Cryst D66: 486–501
63. Hanson MA, Roth CB, Jo E et al (2012) Crystal structure of a lipid G protein-coupled receptor. Science 335:851–855
64. Sanishvili R, Yoder DW, Pothineni SB et al (2011) Radiation damage in protein crystals is reduced with a micron-sized X-ray beam. Proc Natl Acad Sci 108:6127–6132
65. Kato HE, Zhang F, Yizhar O et al (2012) Crystal structure of the channelrhodopsin light-gated cation channel. Nature 482: 369–374
66. Tanaka Y, Hipolito CJ, Maturana AD et al (2014) Structural basis for the drug extrusion mechanism by a MATE multidrug transporter. Nature 496:247–251
67. Nishizawa T, Kita S, Maturana AD et al (2013) Structural basis for the counter-transport mechanism of a H^+/Ca^{2+} exchanger. Science 341:168–172
68. Kumazaki K, Chiba S, Takemoto M et al (2014) Structural basis of Sec-independent membrane protein insertion by YidC. Nature 509:516–520
69. Arndt UW (1984) Optimum X-ray wavelength for protein crystallography. J Appl Cryst 17: 118–119
70. Helliwell JR (1984) Synchrotron X-radiation protein crystallography: instrumentation, methods and applications. Rep Prog Phys 47: 1403–1497
71. Hasegawa K, Shimizu N, Okumura H et al (2013) Diffraction structural biology. J Synchrotron Rad 20:910–913
72. Moukhametzianov R, Burghammer M, Edwards PC et al (2008) Protein crystallography with a micrometre-sized synchrotron-radiation beam. Acta Cryst D64: 158–166
73. Cowan JA, Nave C (2008) The optimum conditions to collect X-ray data from very small samples. J Synchrotron Rad 15:458–462
74. Nave C, Hill MA (2005) Radiation damage. J Synchrotron Rad 12:299–303
75. Fourme R, Honkimaki V, Girard E et al (2012) Reduction of radiation damage and other benefits of short wavelengths for macromolecular crystallography data collection. J Appl Cryst 45:652–661
76. Shimizu N, Hirata K, Hasegawa K et al (2006) Dose dependence of radiation damage for protein crystals studied at various X-ray energies. J Synchrotron Rad 14:4–10
77. Chapman HN, Fromme P, Barty A et al (2012) Femtosecond X-ray protein nanocrystallography. Nature 469:73–77
78. Redecke L, Nass K, DePonte DP et al (2013) Natively inhibited trypanosoma brucei cathepsin B structure determined by using an X-ray laser. Science 339:227–230
79. Yu C, Zhang YL, Pan WW et al (2013) CRL4 complex regulates mammalian oocyte survival and reprogramming by activation of TET proteins. Science 342:1518–1521
80. Hirata K, Shinzawa-Itoh K, Yano N et al (2014) Determination of damage-free crystal structure of an X-ray–sensitive protein using an XFEL. Nat Meth 11:734–736
81. Cohen AE, Soltis SM, Gonzalez A, et al (2014) Goniometer-based femtosecond crystallography with X-ray free electron lasers. In: Proceedings of the National Academy of Sciences. doi:10.1073/pnas.1418733111
82. Kupitz C, Basu S, Grotjohann I et al (2014) Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. Nature 513:261–265
83. Tenboer J, Basu S, Zatsepin N et al (2014) Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. Science 346:1242–1246
84. Schlichting I (2015) Feature articles. IUCrJ 2:246–255. doi:10.1107/S205225251402702X
85. Kern J, Yachandra VK, Yano J (2015) Metalloprotein structures at ambient conditions and in real-time: biological crystallography and spectroscopy using X-ray free electron lasers. Curr Opin Struct Biol 34:87–98

Chapter 15

Structural Biology and Electron Microscopy

Kazuhiro Mio, Masahiko Sato, and Chikara Sato

Abstract

Like X-ray crystallography and NMR, electron microscopy (EM) is now widely applied to determine the structure of proteins and their macromolecular complexes. Single-particle analysis (SPA), which reconstructs the three-dimensional (3D) structure of a protein from its EM images using image processing, has an advantage when the target molecule is difficult to crystallize or only a small amount of protein can be obtained. The technique is based on the theory that two-dimensional EM images of a protein contain sufficient information to reconstruct the original 3D structure. SPA was developed when this theory was applied to ribosomes and the coat protein of icosahedral or helical symmetrical viruses. Because SPA does not require protein crystallization, it is widely applicable to the analysis of solubilized membrane proteins or supermolecular complexes. It allows conformational changes undergone by proteins to be documented. Many other EM-based structural analysis techniques are available in addition to SPA. Electron tomography reconstructs the 3D structure of a protein complex or a cell from a series of images recorded by tilting the specimen in the EM column. Electron crystallography can yield the high-resolution structure of proteins crystallized in two dimensions in a lipid bilayer. Atmospheric scanning electron microscopy directly observes cells in aqueous solution and has realized high-throughput immuno-EM of cells without hydrophobic treatment. It can also visualize protein microcrystals in the crystallization buffer.

Keywords Electron microscope, Macromolecular complex, Single-particle analysis, Three-dimensional structure

1 Introduction

The number of structures registered in the Protein Data Bank (PDB; http://www.rcsb.org/pdb/) is increasing rapidly and reached more than 100,000 in 2015. The acceleration observed over the years is the result of improved experimental protocols, the development of new techniques, and the funding of many national projects. The development of crystallization robots was a particular milestone, as these only require a few microliters of sample for each crystallization condition search. Today, researchers aim to obtain snapshots of conformational dynamics or the structure of protein complexes by the co-crystallization of the interacting components. In spite of the advances, the structural analysis of many important

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks, DOI 10.1007/978-4-431-56030-2_15, © Springer Japan 2016

proteins is lagging behind due to difficulties in purification and/or crystallization.

In 1968, De Rosier and Klug demonstrated that three-dimensional (3D) structures can be reconstructed from the two-dimensional (2D) projections obtained in the transmission electron microscope (TEM) [1]. This methodology was applied to analyze ribosomes [2, 3] and spherical viruses [4, 5], the analysis profiting from the high contrast of the RNA in ribosomes and the high symmetry of the viruses. The technique has developed into the method known as single-particle analysis (SPA) [6]. The resolution attainable by SPA of cryo-electron microscopy (cryo-EM) images is now reaching near-atomic or atomic level as a result of improved direct detection camera (DDC) and image processing methods and increased computational ability [7–9].

2 Materials

2.1 Electron Microscope (EM)

The three leading EM manufacturers are:

- JEOL Ltd., 1-2, Musashino 3-chome Akishima, Tokyo 196-8558, Japan (http://www.jeol.co.jp/en/)
- Hitachi High-Technologies Corporation, 24-14, Nishi-Shimbashi 1-chome, Minato-ku, Tokyo 105-8717, Japan (http://www.hitachi-hitec.com)
- FEI Company, North America NanoPort 5350 NE Dawson Creek Drive Hillsboro, Oregon 97124, USA (http://www.fei.com)

2.2 Peripheral Devices and Materials for Electron Microscopy (EM)

1. Glow discharge system, carbon coater, and other devices essential for imaging proteins on carbon film supported by a mesh EM grid can be obtained from companies specialized in EM equipment.
2. Various types of EM grids are commercially available. Quantifoil holey carbon film grids (Quantifoil Micro Tools GmbH) are popular for cryo-EM.
3. Plunge freezer (“Vitrobot” from FEI or “EM GP Automatic Plunge Freezer” from Leica). This enhances the reproducibility of freezing a sample in a thin layer of buffer for cryo-EM.
4. Liquid nitrogen and liquid ethane (and/or propane) for cryo-EM sample preparation, cooling the specimen and cooling the sample holder.
5. Highly sensitive photographic films (e.g., Kodak electron microscope film SO-163) and developing systems and a high-performance film digitizer (scanner) or another data recording system, i.e., a charge-coupled device (CCD) detector or a

direct detection camera (DDC) using complementary metal–oxide–semiconductor (CMOS) detectors.

2.3 Data Analysis

Computational power can limit the efficiency of the analysis, especially when the size of the target is very large, the size of the dataset is large, or the size of each image is large, e.g., for super-resolution analysis. Several software packages are available for SPA:

IMAGIC (https://www.imagescience.de/imagic.html)

Spider (http://spider.wadsworth.org/spider_doc/spider/docs/spider.html)

EMAN1 and EMAN2 (http://blake.bcm.edu/emanwiki/EMAN)

XMIPP (http://xmipp.cnb.csic.es/twiki/bin/view/Xmipp/WebHome)

Eos (http://www.yasunaga-lab.bio.kyutech.ac.jp/Eos/index.php/Main_Page)

FREALIGN (http://grigoriefflab.janelia.org/frealign)

RELION (http://www2.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page)

3 Methods: Sample Preparation, EM, and Single-Particle Reconstruction

3.1 Protein Purification

The isolated protein employed should be pure with minimum protein deformation and degradation. It is recommended to carry out a size-exclusion chromatography (SEC) step immediately before the sample is adsorbed to the EM grid to ensure that it is homogeneous. This step efficiently removes molecules with different mobility due to partial denaturation or subunit dissociation. In this section, we discuss the purification of membrane proteins and give several tips.

Membrane proteins such as ion channels, transporters, pumps, and cell surface receptors have at least one, and frequently more than six, membrane spanning regions with extracellular and intracellular domains. Membrane proteins are extracted from the cell membrane using detergents. The isolated proteins are generally unstable and denature easily. As the best detergent for protein extraction is not always the best to ensure stability, researchers have to find the optimum target protein–detergent combinations for each step of the process. To minimize denaturation, solubilized proteins should be handled in aqueous solutions containing detergent above the critical micelle concentration (CMC). After extraction, they can be enriched by a combination of different purification procedures, including affinity chromatography, ion exchange chromatography, and SEC. These remove impurities. SEC also indicates the condition of the protein; additional peaks appear if there is significant denaturation or aggregation. As the peaks obtained by SEC frequently become ambiguous when the buffer contains detergent, the presence of a fluorescence tag on the target protein is advantageous, allowing fluorescence-detection SEC (FSEC) to be used [10]. The condition of the protein, especially the degree of aggregation, can be also monitored by negative stain TEM. Here, the protein is adsorbed to the carbon film of an EM grid and surrounded by a high scattering salt, which gives a negative contrast in the microscope [11]. Those who are not familiar with the negative stain TEM of detergent-rich protein samples are advised to directly observe protein aggregation in buffer using the atmospheric scanning electron microscope (ASEM) [12] in combination with metal staining as described in Sect. 4.3.

In most cases, we obtained high-quality proteins using a combination of affinity chromatography and SEC [13], both of which are especially effective for membrane proteins. The affinity chromatography step should be carried out at an early stage of the purification process. The addition of glycerol or sucrose (up to 50 %) to the protein sample sometimes helps to minimize absorption loss and unstability of proteins, especially during the chromatography and ultrafiltration.

3.2 Generation of Antibodies for Affinity Purification

Polyclonal antibodies against cytoplasmic tails or linkers in the target proteins are ideal for immuno-affinity chromatography. In our experience, finding monoclonal antibody clones appropriate for affinity-column purification is sometimes not easy, because the affinity of most monoclonal antibodies to the target protein is lower than expected. Instead, it could be better to raise a polyclonal antibody against a synthetic peptide (20–25 amino acids) that corresponds to part of the protein. The selected sequence should not contain too many cysteine residues to avoid various conformers and should not include transmembrane or extracellular segments, but rather the cytoplasmic terminal or a linker sequence of the target membrane protein.

This approach was used to purify the voltage-gated sodium channel of electric eels [14, 15]. Antiserum was raised by immunizing a New Zealand White rabbit with the selected oligopeptide conjugated to keyhole limpet hemocyanin (Sigma Chemical Co., St. Louis, MO) [16]. Antibodies were purified by affinity chromatography, using an affinity gel prepared by conjugating the oligopeptide to Actigel (Sterogene Bioseparations Inc., Arcadia, CA). Antibodies that bind to such gels are usually eluted using low or high pH buffer. In this case, low pH buffer (pH 2.5) was employed. Finally, the affinity gel required to purify the membrane protein was synthesized by conjugating the eluted antibodies (e.g., 1–100 mg) to Actigel (1–10 ml). Such affinity gels can be stored for several years in a buffer containing 0.1 % NaN_3. Sometimes the affinity of some antibodies against synthesized sequences is too high to allow protein elution under moderate conditions. Such high-affinity antibodies can be used for the quantitative analysis such as Western blotting procedure and/or ELISA. Peroxidase-conjugated Fabs can be prepared for this purpose [17].

3.3 Sample Preparation for Observation by EM

3.3.1 Negative Staining

Biological macromolecules are primarily comprised of light atoms, such as hydrogen (H), carbon (C), nitrogen (N), and oxygen (O). Most electrons penetrate these light atoms without scattering, resulting in very low-contrast images of the protein above the background scattering of the supporting carbon. Negative stains can be used to increase contrast when samples are dried on carbon film for EM. Various negative stains are available; all contain heavy metals, such as uranium (U), tungsten (W), and molybdenum (Mo).

For negative-stain TEM, a thin carbon film on a copper mesh EM grid is rendered hydrophilic by glow discharge in a low pressure of air immediately before a few microliters of the protein sample is added. After the proteins have adsorbed to the film, excess solution is removed by filter paper. The adsorbed proteins are washed several times with water or a buffer containing ammonium carbonate ($(NH_4)_2CO_3$). The heavy metal solution is then added to the grid and excess stain solution is sucked up by filter paper. This can be repeated a couple of times (Fig. 1a, left). The heavy metal remains on the carbon film and frequently at the molecular edge and within surface indentations and internal cavities of each protein particle. Electrons are scattered by the residual heavy metal and penetrate unstained regions. As proteins are generally more hydrophobic than the glow-discharged carbon, their surface remains mainly unstained, and they are visualized brightly surrounded by a border of dark stain, except where the stain has filled cavities (Fig. 1b, left). Since it gives reverse contrast like a film negative, the method is called negative staining [18]. Negative-stain EM can provide relatively rich structural information without the need to stabilize the sample by fixation or to embed it in a resin.

3.3.2 Metal Shadowing

Metal shadowing is used to observe proteins and nucleic acids. Metals are heated by current at high voltage under vacuum, vaporize, and adhere to the molecular surface of proteins. Since the metal coat scatters electrons, the surface structure can be clearly observed at high contrast. Metals with high atomic number, such as platinum, gold, tungsten, and a Pt-Pd alloy, can be vapor deposited in this way.

3.3.3 Cryo-Embedding

Crystalline ice produces a diffraction pattern on the fluorescent screen of a TEM when the electron beam is applied. However, when water is rapidly cooled using liquid ethane slush at liquid nitrogen temperature (boiling point −195.8 °C) or using a copper

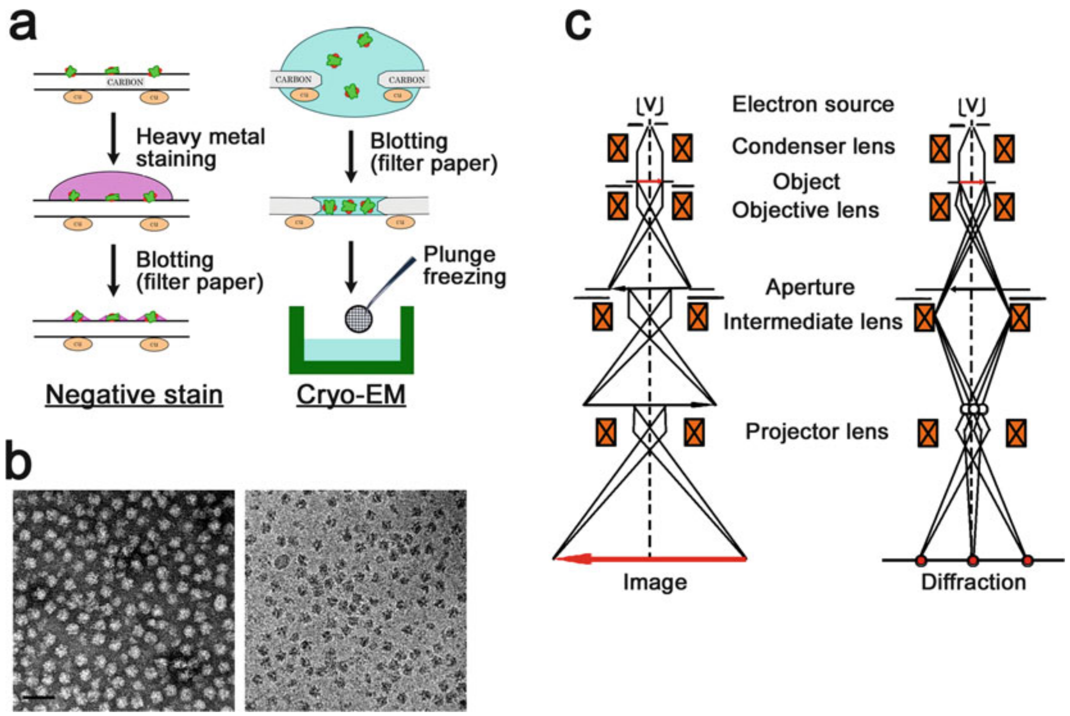


Fig. 1 Electron microscopy for structural biology. (**a**) Sample preparation for negative-stain EM (*left*) and cryo-EM (*right*). In this figure, membrane proteins (*green*) are associated with membrane lipids (*red*). (**b**) EM images of ribosomes purified from *E. coli*: *left*, negative-stain TEM, and *right*, image cryo-EM. Data were provided by Takeshi Yokoyama. *Scale bar* represents 100 nm. (**c**) *Diagram* indicating the use of the TEM in the imaging (*left*) and diffraction (*right*) mode

metal button cooled by liquid helium (boiling point −268.9 °C), "vitrified ice" is generated. Since this has no crystal structure, it does not diffract the electron beam.

To prepare samples in vitrified ice, a solution containing target proteins is applied to an EM grid coated with a perforated carbon film (holey grid), and excess liquid is blotted away using a filter paper. At this stage, the residual protein suspension spans the large numbers of small holes in the perforated carbon film. This thin aqueous layer (should be <500 nm) is rapidly cooled by plunging the grid into liquid nitrogen-cooled ethane slush either manually or in a more controlled manner using a plunge freezer (Fig. 1a, right). The proteins become embedded in vitrified ice in a close-to-native state [19, 20]. As the density of protein is slightly higher than that of vitrified ice, the particle images appear dark in a lighter vitrified ice background (Fig. 1b, right).

The vitrified ice functions as a "supporting film" for the target proteins, which require neither fixation nor staining. However, the contrast of particles in vitrified ice is very low, and, in most cases, large-scale image alignment/classification and averaging are required to obtain a clearer view. Recent progress in computational ability and algorithm development has facilitated the handling of the large amount of image data required for high-resolution analysis.

The cryo-embedded samples are transferred to the cryo-EM at liquid nitrogen temperature using a cryo-transfer instrument. Each grid is observed in the cooled stage of the microscope at liquid nitrogen or liquid helium temperature [21, 22].

3.4 EM

3.4.1 Comparison of the LM and the TEM

The electron beam, generated in the electron gun at the top of a TEM, is accelerated down the microscope column, passes through the sample, and impinges on the detector below (Fig. 1c). Because electrons are scattered by air, the microscope column is kept at high vacuum, 10^{-5}–10^{-7} Pa. Electrons that irradiate the specimen either simply pass through it or are scattered. The latter causes them to diverge at various angles. Magnetic lenses positioned along the microscope column are used to condense or disperse the electrons. The enlarged image of the specimen obtained can either be viewed using the fluorescence plate at the bottom of the column or captured on photographic film or several types of detectors (Sect. 2.2 and below).

The wavelength of an electron beam depends on the acceleration voltage employed. In the TEM the accelerating voltage is usually 100–300 kV, which corresponds to wavelengths of 0.00370–0.00196 nm. This is much shorter than the wavelengths used for light microscopy (LM), making it possible to visualize specimens at much higher resolution. Because the lenses of an EM are electromagnets, the magnification can be tuned by changing their voltage. The image obtained is modulated by the contrast transfer function (CTF), which is determined by the parameters of the microscope including the acceleration voltage used and the spherical aberration constant, Cs. The combination of lens compensating aberrations is different in an LM. The cryo-TEM was developed to observe samples embedded in vitrified ice at the temperature of liquid nitrogen or liquid helium.

3.4.2 Data Recording

Photographic film has been used for more than half century to record TEM images. Sheet film is commonly employed. The film is highly sensitive and allows a larger area to be recorded than other detector systems. However, the response of photographic film to the number of incident electrons is not linear and developing the film takes time. Moreover, before image processing, the negatives have to be digitized by a high-performance image scanner.

Phosphor-coupled CCD detectors can be used online and are replacing photographic film. When electrons irradiate the fluorescent scintillator, they generate light, which is then transferred via a lens or fiber optic to the CCD sensor. In the CCD, electrons are first accumulated in the depletion region (potential wells), then transferred to successive wells, and read out as electric signals. To suppress the dark current, the CCD is usually cooled to −30 °C or less. CCD detectors with a larger pixel dimension (4 × 4 K) are preferentially used for data acquisition in SPA.

DDCs were recently developed on the basis of complementary metal–oxide–semiconductor (CMOS) technology. Each pixel contains both a photodetector and an active amplifier that is addressed and read out individually. Unlike CCDs, readout of CMOS devices does not require pixel-to-pixel charge transfer. As a result of their improved active pixel sensor (APS) and radiation resistance, DDCs are now being built into cryo-EMs. They have a high detective quantum efficiency (DQE), a very low point spread function (PSF), and rapid readout [23–26]. Further improvement of the dynamic range and endurable electron dose of DDCs is expected to make these cameras even more sensitive increasing the contrast of recorded cryo-EM images [23–25].

3.5 Data Processing

Several software packages for SPA, such as IMAGIC [27, 28], Spider [29, 30], EMAN [31, 32], XMIPP [33–35], Eos [36], FREALIGN [37], and RELION [38], facilitate progress and broaden the use of this method. There are differences in the analytical policy, so the details should be obtained from the original papers or home pages. Here, we outline the basic concept and analytical flow of SPA (Fig. 2).

3.5.1 Step 1: Data Pretreatment

Some proteins in a sample may be partly denatured or truncated by proteolysis. Their images cannot be used for the reconstruction and ideally should be excluded by inspection before processing begins. The defocus value of each micrograph should first be measured, and the CTF should be corrected, as this is critical for the interpretation of spatial frequencies beyond the first zero of the CTF. The CTF correction differs slightly from program to program, e.g., EMAN2 [32] prefers to adjust the CTF for each picked particle, while IMAGIC V [27] recommends to correct the CTF for each micrograph. In case of RELION and FREALING, CTF correction is performed in the algorithms for classification and refinement. Some modulations caused by the CTF, e.g., those resulting from maladjusted astigmatism and sample drift, prevent precise reconstruction. Inspection of the Thon rings is thus an effective way to identify the micrographs that are affected by such modulations.

3.5.2 Step2: Particle Picking

Sufficient particles have to be collected from the recorded images. Currently, hundreds of thousands or even a million particles are used for structure determination at near-atomic or atomic resolution from cryo-EM images and several thousands for the reconstruction of a negatively stained sample. If the target protein has high point symmetry, e.g., virus particles, the number of particles required is less. Interactive particle pickup is generally used to

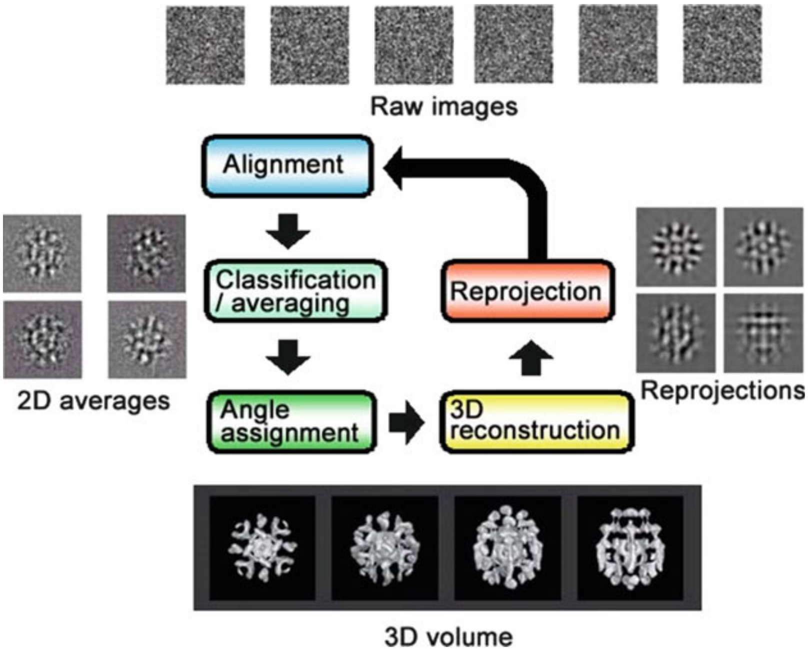


Fig. 2 Workflow for 3D reconstruction by SPA. The reconstruction of TRPC3 channel from cryo-EM images is used as an example. In this experiment, TRPC3 molecules in vitrified ice were recorded by cryo-electron microscopy with a stable liquid helium stage. From 142,553 particles selected by the neural network system, a 3D map was reconstructed at 15 Å resolution by the FSC >0.5 criterion [13]

obtain data from the digitized micrographs. This step can be a bottleneck in the processing and sometimes it is difficult to distinguish the faint projections from the background. Images taken with a relatively small defocus contain the high-frequency information essential for high-resolution analysis, but the contrast is low. Thus the “focal pair” technique is sometimes employed to facilitate particle pickup. Here, pairs of images are recorded, the first close to focus and the second with a larger under focus. Particles are picked up from the first image using the coordinates obtained from the higher contrast second image. Automated pickup programs are also being eagerly developed. We have developed a neural network (NN)-based particle pickup program [39] and a program based on iterative multi-reference alignment (MRA) [40]. Motion correction of captured images with DDC enables uses of much clearer particle images [25].

3.5.3 Step3: Alignment and 2D Averages of Particles

To reduce background noise and improve the signal-to-noise ratio, the selected particle images are aligned by MRA techniques and sorted into classes of homogeneous 2D images by automatic multivariate statistical classification procedures or NN-based classification algorithms [41]. These characteristic views are used for the 3D reconstruction.

If a low-resolution volume of the target protein or a homologue is already available, template-matching can be used for MRA and further classification. Low-resolution projections are generated from the 3D volume or crystal data and used as templates. To avoid bias from the model, the resolution of the template should be much less than the expected resolution of the final EM reconstruction, or the data should be analyzed using multiple templates and the resulting 3D reconstructions compared.

3.5.4 Step 4: Reconstruction

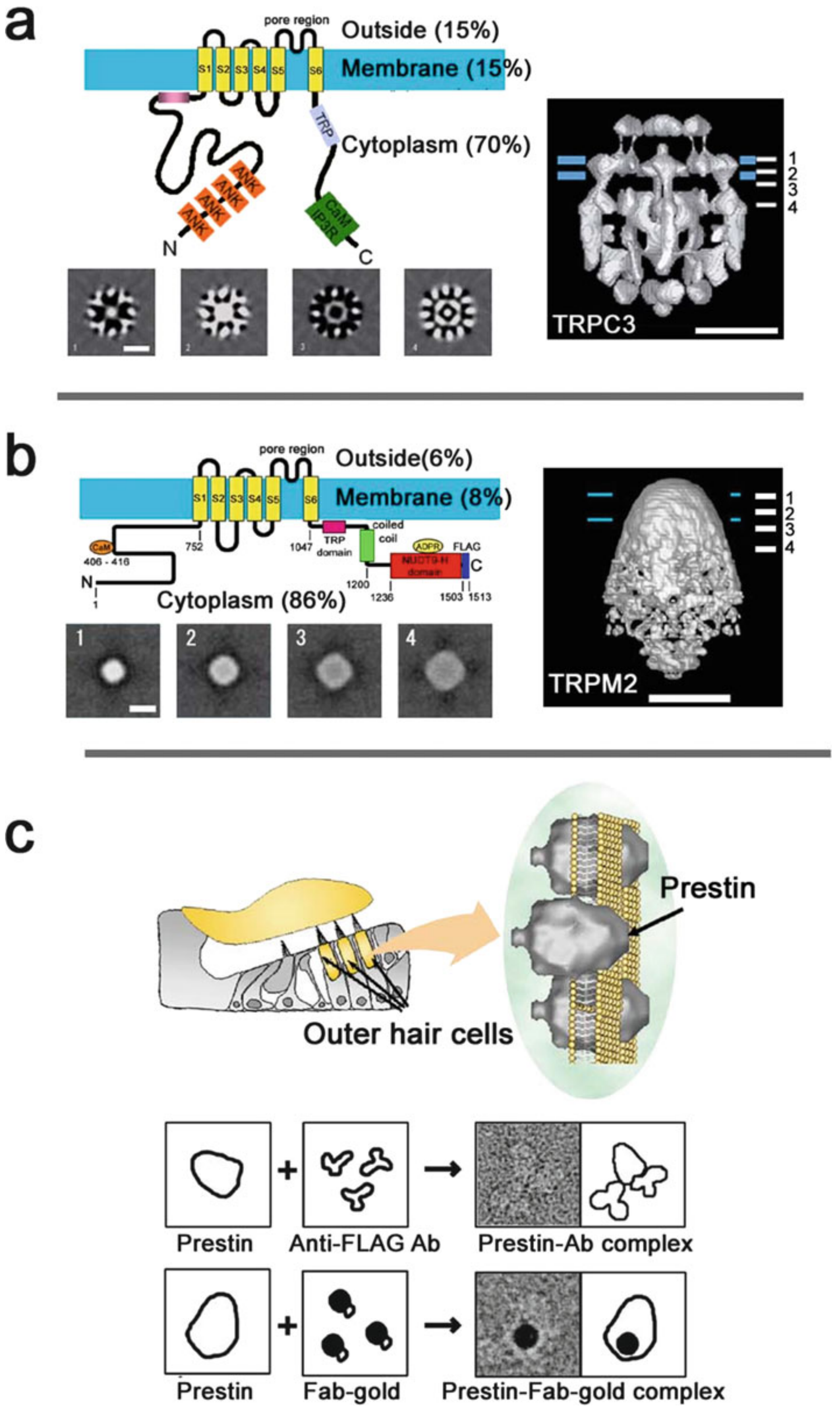
To create the initial 3D structure, the Euler angle of each particle average needs to be determined. Several methods have been developed to do this posteriorly. The common line method [42, 43] uses the central section theorem, which states that any two projections of a given structure share a common central line in Fourier space. Based on this theorem, the sinogram approach was successfully applied to structures with high point symmetry, but the robustness is relatively low for noisy images of asymmetric or heterogeneous molecules.

The random conical tilt method is also used to generate initial models [44, 45]. Tilted (30–60°) and untilted EM images are taken of the same grid area. The untilted images are used to sort the data into classes representing characteristic views of the molecule and to determine the azimuths, while the tilted images are used for reconstruction. The initial model generated becomes the starting model for the second round of alignment. Such cycles of alignment and reconstruction are repeated until the images converge (Fig. 2). Although this technique allows the 3D structures of proteins that have a strong preferred orientation on the support film to be elucidated, the reconstruction is often incomplete due to missing data around 90° tilt (the missing cone).

For posteriori Euler angle assignment, we have developed a novel reference-free 3D reconstruction system using simulated-annealing algorithms [46]. This starts from an initial 3D volume that is generated by back-projecting the randomly oriented 2D averages on a sphere. The structure is then optimized by evaluating the correlation coefficient between the reprojections of the volume and the average images. The method can be applied to asymmetric proteins of unknown structure and can overcome local minimums that appear during the volume optimization step. Membrane proteins reconstructed by single-particle analysis are exhibited in Fig. 3 [13, 47, 48].

3.5.5 Step 5: Interpretation of the Reconstructed Volume

The most popular way to assess the resolution of an EM reconstruction is to compare the two reconstructions generated when the dataset is randomly divided in half and the two halves independently analyzed. The reconstructions from, e.g., the even- and odd-numbered images, are compared in Fourier space and the differences


a
pore region
Outside (15%)
S1 S2 S3 S4 S5 S6
Membrane (15%)
TRP
Cytoplasm (70%)
ANK
ANK
ANK
ANK
CaM IP3R
N
C
1
2
3
4
TRPC3
b
pore region
Outside(6%)
S1 S2 S3 S4 S5 S6
Membrane (8%)
CaM
406 - 416
752
1047
TRP domain
coiled coil
ADPR
FLAG
NUDT9-H domain
C
1200
1236
1503
1513
N
1
Cytoplasm (86%)
1
2
3
4
TRPM2
c
Prestin
Outer hair cells
Prestin
+
Anti-FLAG Ab
Prestin-Ab complex
Prestin
+
Fab-gold
Prestin-Fab-gold complex

determined over different shells. A Fourier shell correlation (FSC) threshold of 0.5 is conservative and the most common criteria used to evaluate SPA [4]. A threshold of 0.143 is also used [49]. This value comes from the corresponding threshold value used in X-ray crystallography.

If it is available, fitting the high-resolution X-ray substructure or a partial structure into the EM volume enables validation of the reconstructed structure and also interpretation of the structure based on quasi-atomic modeling. In addition to manual fitting, several programs are available for rigid body fitting and flexible fitting [50]. Labeling technique using specific antibodies or gold conjugates will provide domain information of the target macromolecules (Fig. 3c).

4 Other Applications

4.1 Tomography

Electron tomography (ET) is a 3D reconstruction technique; a series of EM images are taken by tilting the sample stage in the microscope column, and the 3D structure is reconstructed from these images (Fig. 4a). The principle of this technique is the same as for SPA and X-ray computer tomography (CT) and magnetic resonance imaging (MRI), both of which are popular in the medical field. In ET, images should be ideally taken from all directions; otherwise limitation of tilting angles will cause the "missing wedge problem." Cryo-electron tomography, where the sample is maintained at liquid nitrogen or liquid helium temperature, has been developed for the precise analysis of tissue, cells, and macromolecules [53]. Because scanning transmission electron microscope (STEM) tomography can change scanning focus depth depending on the tilt of the sample plane, it can avoid the focal gap caused by tilting the sample for TEM tomography [51].

4.2 2D Crystallization and Electron Crystallography

The plasma membrane of cells is mainly formed by a lipid bilayer and proteins. Cells use receptor proteins integrated in the membrane to transmit external signals to their interior and membrane channel or pump proteins to perform transmembrane transport.

High-resolution structures of functionally important membrane proteins have been obtained by electron crystallography of 2D crystals [54] (Fig. 4b, c). If present in sufficient quantity, the

Fig. 3 Single-particle reconstruction of membrane proteins. (**a**) Reconstruction of the TRPC3 channel at 15 Å resolution from cryo-embedded specimens [13]. (**b**) Structure of the TRPM2 channel at 28 Å resolution from negatively stained specimens [47]. (**c**) Reconstructed structure of the motor protein prestin, which amplifies sound signal in the inner ear. Labeling with specific antibodies or gold-conjugated Fab determines the topology of the molecules [48]

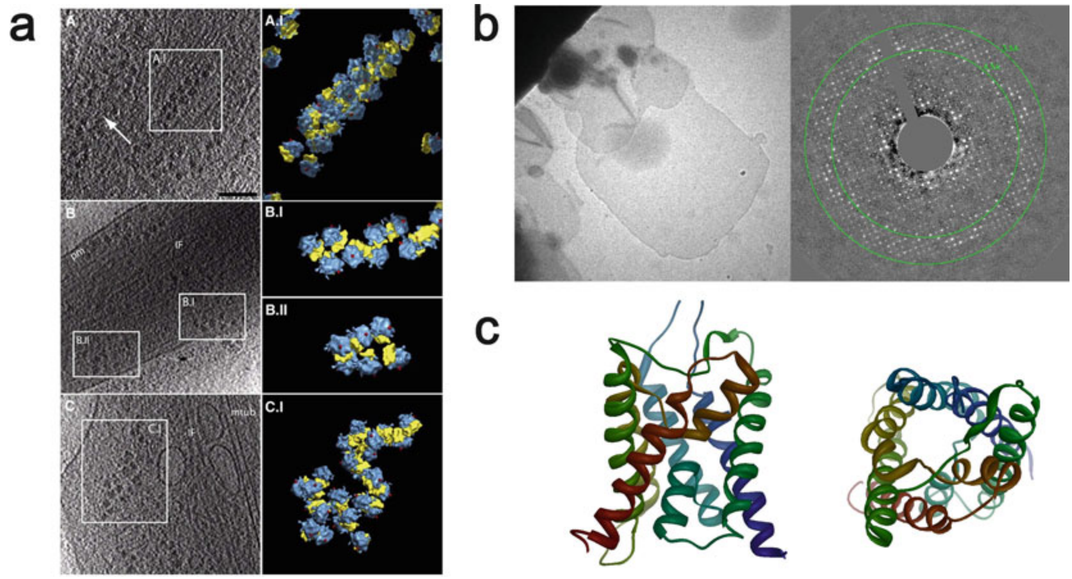


Fig. 4 ET and electron crystallography. **a** Polyribosome in human glioblastoma cell line visualized by ET. *Left panels* (**a**–**c**) are subtomograms showing various shapes. *Right panels* (*A.I–C.I*) are isosurface models derived from the tomographic average. From Ref. [51] with permission. **b** Low-magnification image of 2D crystals of aquaporin-4 (AQP4) (*upper left*) and the diffraction pattern (*upper right*). Data were provided by Kaoru Mitsuoka. (**c**) Ribbon diagrams of AQP4, resolved at 2.8 Å resolution from two-dimensional crystals [52]

membrane protein was extracted from the natural source using a suitable detergent [55]. Alternatively, it was overexpressed in a suitable expression system (i.e., bacteria cells, insect cells, yeast cells, or mammalian cells) and similarly extracted. After purification, membrane proteins are mixed in various ratios with detergent-solubilized lipid. Detergent is then removed from the sample, usually by dialysis, the sample being incubated to reconstitute the target membrane proteins into a lipid bilayer. 2D crystals grow if the conditions are suitable. Key parameters are the kind of detergent, the pH, the temperature during dialysis, and the lipid-to-protein ratio [55].

Some membrane proteins are highly expressed in nature and even form 2D crystals. The purple membranes of Archaea are 2D crystals of bacteriorhodopsin [56]. Extracted purple membrane patches have been examined by electron crystallography. They were treated with an ionic detergent first, which caused them to fuse enlarging the analytical area available [57, 58].

4.3 In-Liquid Observation of Protein Localization and Crystals by ASEM

In the standard EM, the sample must be observed in vacuum, which means that it has to be dried or coated or frozen. ASEM was developed to realize direct observation of cells or protein complexes in aqueous solution under open atmosphere. The ASEM is an inverted scanning electron microscope (SEM) (Fig. 5a). The column is under vacuum being sealed at the top by

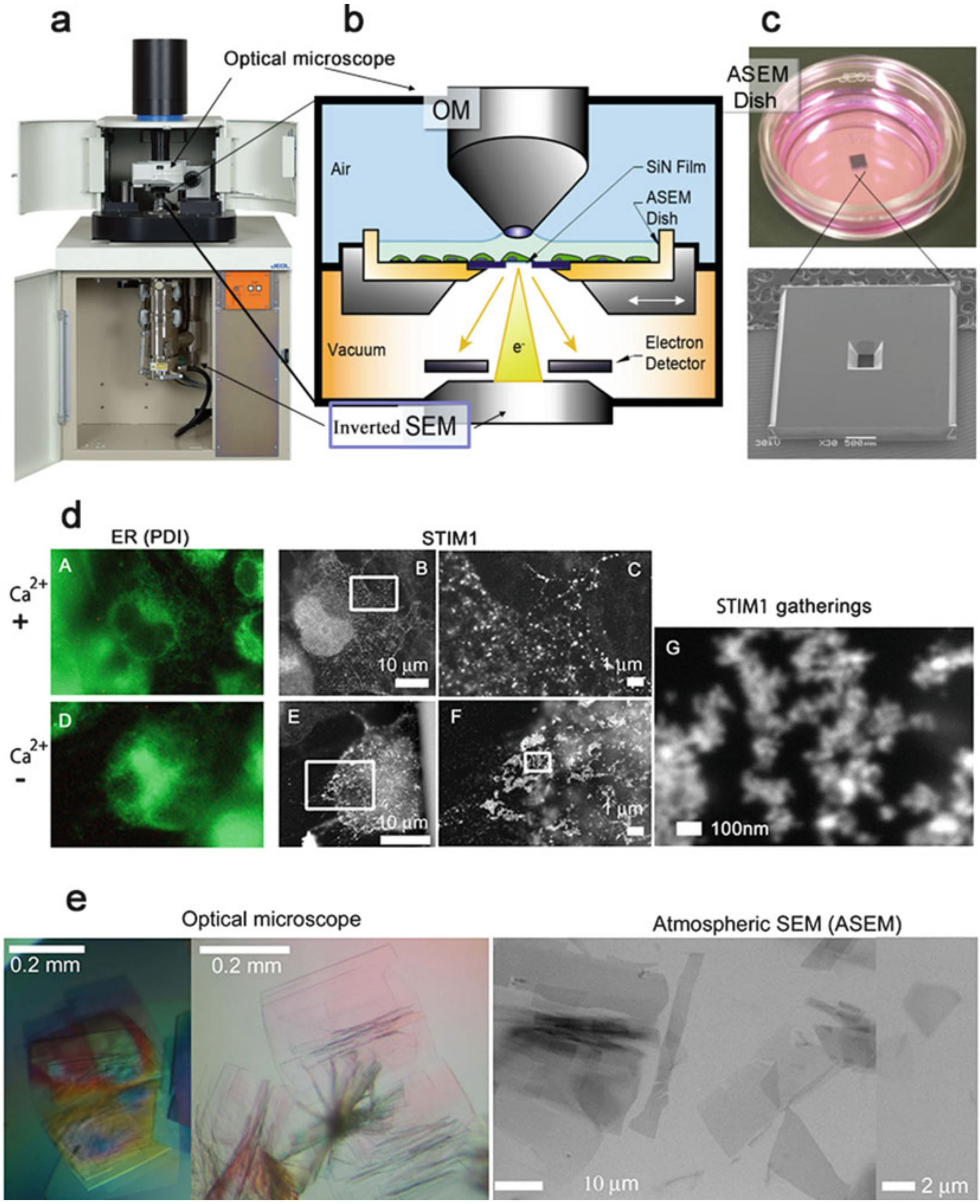


Fig. 5 ASEM observation of protein and cells. (**a**–**c**) The ASEM as used for CLEM [12]. (**d**) Dynamic rearrangement of STIM1 in response to Ca^{2+} store depletion. STIM1-expressed COS7 cells with (*lower*) and without thapsigargin treatment (*upper*) were immuno-labeled for STIM1 [59]. (**e**) Direct observation of protein 3D microcrystals in crystallization buffer without staining [60]

the SiN film window in the base of the open ASEM specimen dish (Fig. 5c). The sample in this dish is at atmospheric pressure and can be in liquid. The electron beam of the inverted SEM is projected up the column through the SiN film onto the sample and the backscattered electrons are collected (Fig. 5b) [12]. The observable sample depth is 2–3 μm and the resolution obtained when imaging a sample in solution is 8 nm near the SiN film. The inverted SEM and a LM positioned above the sample (Fig. 5a, b) are aligned, allowing correlative microscopy.

ASEM realizes high-throughput immuno-EM of cells, because it does not require the time-consuming hydrophobic treatment and resin embedding otherwise necessary for samples to endure the vacuum of an EM. Various kinds of primary cells including neurons, megakaryocyte, and ES cells have been cultured in the ASEM dish and labeled in situ [61]. As illustrated by Fig. 5d, correlative light-electron microscopy (CLEM) using ASEM allowed molecular supercomplex formation by the Ca^{2+} sensor STIM1 of the endoplasmic reticulum in response to Ca^{2+} store depletion to be visualized in situ [59]. Moreover, ASEM can be used to observe wet tissue placed on the ASEM dish [62].

ASEM can also be used to directly observe micro-protein 3D crystals in crystallization buffer without or with staining [60], to detect and study aggregation, and to detect crystals that cannot be resolved by OM. Further, an ASEM dish with a standard crystallization chamber has been developed to allow the in situ observation of crystallization [60].

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas, Structural basis of cell-signaling complexes mediating signal perception, transduction, and responses, and by grants from CREST; from the Ministry of Education, Culture, Sports, Science, and Technology; from Canon; and from AIST.

References

1. De Rosier DJ, Klug A (1968) Reconstruction of three dimensional structures from electron micrographs. Nature 217:130–134
2. Frank J, Zhu J, Penczek P et al (1995) A model of protein synthesis based on cryo-electron microscopy of the *E. coli* ribosome. Nature 376:441–444
3. Stark H, Mueller F, Orlova EV et al (1995) The 70S *Escherichia coli* ribosome at 23 Å resolution: fitting the ribosomal RNA. Structure 3:815–821
4. Bottcher B, Wynne SA, Crowther RA (1997) Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. Nature 386:88–91
5. van Heel M, Gowen B, Matadeen R et al (2000) Single-particle electron cryo-microscopy: towards atomic resolution. Q Rev Biophys 33:307–369
6. Frank J (2006) Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native

state. Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state. Oxford University Press, New York

7. Bartesaghi A, Merk A, Banerjee S et al (2013) Electron microscopy. 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. Science 348:1147–1151
8. Jiang J, Pentelute BL, Collier RJ, Zhou ZH (2015) Atomic structure of anthrax protective antigen pore elucidates toxin translocation. Nature 521:545–549
9. Paulsen CE, Armache JP, Gao Y, Cheng Y, Julius D (2015) Structure of the TRPA1 ion channel suggests regulatory mechanisms. Nature 520:511–517
10. Kawate T, Gouaux E (2006) Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. Structure 14:673–681
11. Bremer A, Henn C, Engel A, Baumeister W, Aebi U (1992) Has negative staining still a place in biomacromolecular electron microscopy? Ultramicroscopy 461:85–111
12. Nishiyama H, Suga M, Ogura T et al (2010) Atmospheric scanning electron microscope observes cells and tissues in open medium through silicon nitride film. J Struct Biol 169:438–449
13. Mio K, Ogura T, Kiyonaka S et al (2007) The TRPC3 channel has a large internal chamber surrounded by signal sensing antennas. J Mol Biol 367:373–383
14. Sato C, Sato M, Iwasaki A, Doi T, Engel A (1998) The sodium channel has four domains surrounding a central pore. J Struct Biol 121:314–325
15. Sato C, Ueno Y, Asai K et al (2001) The voltage-sensitive sodium channel is a bell-shaped molecule with several cavities. Nature 409:1047–1051
16. Yuuki H, Hasunuma Y, Komazawa K et al (1996) A sensitive enzyme immunoassay specific for salmon calcitonin. Biomed Res 17:257–259
17. Ishikawa E, Yoshitake S, Imagawa M, Sumiyoshi A (1983) Preparation of monomeric Fab'-horseradish peroxidase conjugate using thiol groups in the hinge and its evaluation in enzyme immunoassay and immunohistochemical staining. Ann N Y Acad Sci 420:74–89
18. Brenner S, Horne RW (1959) A negative staining method for high resolution electron microscopy of viruses. Biochim Biophys Acta 34:103–110
19. Taylor KA, Glaeser RM (1976) Electron microscopy of frozen hydrated biological specimens. J Ultrastruct Res 55:448–456
20. Adrian M, Dubochet J, Lepault J, McDowall AW (1984) Cryo-electron microscopy of viruses. Nature 308:32–36
21. Fujiyoshi Y, Mizusaki T, Morikawa K et al (1991) Development of a superfluid-helium stage for high-resolution electron-microscopy. Ultramicroscopy 38:241–251
22. Henderson R (2004) Realizing the potential of electron cryo-microscopy. Q Rev Biophys 37:3–13
23. Jin L, Milazzo AC, Kleinfelder S et al (2008) Applications of direct detection device in transmission electron microscopy. J Struct Biol 161:352–358
24. Milazzo AC, Moldovan G, Lanman J et al (2010) Characterization of a direct detection device imaging camera for transmission electron microscopy. Ultramicroscopy 110:744–747
25. Li X, Mooney P, Zheng S et al (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat Methods 10:584–590
26. Grigorieff N (2013) Direct detection pays off for electron cryo-microscopy. Elife 2:e00573
27. van Heel M, Harauz G, Orlova EV, Schmidt R, Schatz M (1996) A new generation of the IMAGIC image processing system. J Struct Biol 116:17–24
28. van Heel M, Portugal R, Rohou A et al (2011) Four-dimensional cryo electron microscopy at quasi atomic resolution: IMAGIC 4D. In: Arnold E, Himmel DM, Rossmann MG (eds) Crystallography of biological macromolecules. Wiley, New York, pp 624–628
29. Frank J, Radermacher M, Penczek P et al (1996) SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J Struct Biol 116:190–199
30. Shaikh TR, Gao H, Baxter WT et al (2008) SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nat Protoc 3:1941–1974
31. Ludtke SJ, Baldwin PR, Chiu W (1999) EMAN: semiautomated software for high-resolution single-particle reconstructions. J Struct Biol 128:82–97

32. Tang G, Peng L, Baldwin PR et al (2007) EMAN2: an extensible image processing suite for electron microscopy. J Struct Biol 157:38–46
33. Scheres SH, Nunez-Ramirez R, Sorzano CO, Carazo JM, Marabini R (2008) Image processing for electron microscopy single-particle analysis using XMIPP. Nat Protoc 3:977–990
34. Sorzano CO, Marabini R, Velazquez-Muriel J et al (2004) XMIPP: a new generation of an open-source image processing package for electron microscopy. J Struct Biol 148:194–204
35. Marabini R, Masegosa IM, San Martin MC et al (1996) Xmipp: an image processing package for electron microscopy. J Struct Biol 116:237–240
36. Yasunaga T, Wakabayashi T (1996) Extensible and object-oriented system Eos supplies a new environment for image analysis of electron micrographs of macromolecules. J Struct Biol 116:155–160
37. Grigorieff N (2007) FREALIGN: high-resolution refinement of single particle structures. J Struct Biol 157:117–125
38. Scheres SH (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol 180:519–530
39. Ogura T, Sato C (2004) Automatic particle pickup method using a neural network has high accuracy by applying an initial weight derived from eigenimages: a new reference free method for single-particle analysis. J Struct Biol 145:63–75
40. Kawata M, Sato C (2013) Multi-reference-based multiple alignment statistics enables accurate protein-particle pickup from noisy images. Microscopy 62:303–315
41. Ogura T, Iwasaki K, Sato C (2003) Topology representing network enables highly accurate classification of protein images taken by cryo electron-microscope without masking. J Struct Biol 143:185–200
42. Crowther RA (1971) Procedures for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs. Philos Trans R Soc Lond B Biol Sci 261:221–230
43. van Heel M (1987) Angular reconstitution – a posteriori assignment of projection directions for 3-D reconstruction. Ultramicroscopy 21:111–123
44. Frank J, Goldfarb W, Eisenberg D, Baker TS (1978) Reconstruction of glutamine synthetase using computer averaging. Ultramicroscopy 3:283–290
45. Radermacher M, Wagenknecht T, Verschoor A, Frank J (1986) A new 3-D reconstruction scheme applied to the 50S ribosomal subunit of E. coli. J Microsc 141:RP1–RP2
46. Ogura T, Sato C (2006) A fully automatic 3D reconstruction method using simulated annealing enables accurate posterioric angular assignment of protein projections. J Struct Biol 156:371–386
47. Maruyama Y, Ogura T, Mio K et al (2007) Three-dimensional reconstruction using transmission electron microscopy reveals a swollen, bell-shaped structure of transient receptor potential melastatin type 2 cation channel. J Biol Chem 282:36961–36970
48. Mio K, Kubo Y, Ogura T et al (2008) The motor protein prestin is a bullet-shaped molecule with inner cavities. J Biol Chem 283:1137–1145
49. Rosenthal PB, Henderson R (2003) Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. J Mol Biol 333:721–745
50. 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
51. Brandt F, Carlson LA, Hartl FU, Baumeister W, Grunewald K (2010) The three-dimensional organization of polyribosomes in intact human cells. Mol Cell 39:560–569
52. Tani K, Mitsuma T, Hiroaki Y et al (2009) Mechanism of aquaporin-4′s fast and highly selective water conduction and proton exclusion. J Mol Biol 389:694–706
53. Baumeister W (2002) Electron tomography: towards visualizing the molecular organization of the cytoplasm. Curr Opin Struct Biol 12:679–684
54. Gonen T, Cheng Y, Sliz P et al (2005) Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. Nature 438:633–638
55. Jap BK, Zulauf M, Scheybani T et al (1992) 2D crystallization: from art to science. Ultramicroscopy 46:45–84
56. Lanyi JK (2004) Bacteriorhodopsin. Annu Rev Physiol 66:665–688
57. Kimura Y, Vassylyev DG, Miyazawa A et al (1997) Surface of bacteriorhodopsin revealed by high-resolution electron crystallography. Nature 389:206–211

58. Subramaniam S, Henderson R (2000) Molecular mechanism of vectorial proton translocation by bacteriorhodopsin. Nature 406:653–657
59. Maruyama Y, Ebihara T, Nishiyama H, Suga M, Sato C (2012) Immuno EM-OM correlative microscopy in solution by atmospheric scanning electron microscopy (ASEM). J Struct Biol 180:259–270
60. Maruyama Y, Ebihara T, Nishiyama H et al (2012) Direct observation of protein microcrystals in crystallization buffer by atmospheric scanning electron microscopy. Int J Mol Sci 13:10553–10567
61. Hirano K, Kinoshita T, Uemura T et al (2014) Electron microscopy of primary cell cultures in solution and correlative optical microscopy using ASEM. Ultramicroscopy 143:52–66
62. Memtily N, Okada T, Ebihara T et al (2015) Observation of tissues in open aqueous solution by atmospheric scanning electron microscopy: applicability to intraoperative cancer diagnosis. Int J Oncol 46:1872–1882

Chapter 16

Structure Determination Software for Macromolecular X-Ray Crystallography

Min Yao

Abstract

Because of the phase problem in crystallography, electron density maps can only be calculated based on the substructure of heavy atoms (experimental phasing) or known homology structure (molecular replacement) to determine the macromolecular structure. Such phasing methods include various errors and are limited by the observed diffraction resolution of crystals. Therefore, various mathematic methods and excellent software packages have been developed for structure determination. Specially, structural genomics projects have advanced the development of powerful and automated methods for macromolecular crystallography during the past decade. In this chapter, typical software often used for structure determination will be introduced. We begin with an overview of the structure determination process and simple mathematic methods in each section. After introducing software packages used in each step, we will mention the strategy/practice for each process of structure determination.

Keywords Crystal structure determination, Phasing, Refinement

1 Introduction

The theory of X-ray crystallography was constructed early for mineral and small-molecule structural analysis at the beginning of the twentieth century. Laue and Bragg, parent–child, were awarded the Nobel Prize for discovering that crystals diffracted [1] following Bragg's law ($2d\sin\theta = n\lambda$) [2–4] by exposure to X-ray in 1914 and 1915, respectively. Two decades later, Bernal and Crowfoot first observed the diffraction photon of a protein crystal [5]. However, the diffraction patterns of protein crystals were very complicated, and it seemed impossible to directly elucidate the three-dimensional structure of the protein at that time.

Basically, the diffraction of X-rays by crystals is a physical phenomenon of Fourier transform (*FT*) due to X-rays being light described as a wave spectrum using sine/cosine:

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks,
DOI 10.1007/978-4-431-56030-2_16, © Springer Japan 2016

$$F(k) = FT(\rho(r)) = \int \rho(r) \exp 2\pi i(kr)dr = |F(k)| \exp(i\alpha_k) \qquad (1)$$

Here, complex number $\boldsymbol{F}(\boldsymbol{k})$ is the structure factor of $\boldsymbol{k} = (h, k, l)$ with contributions from all atoms in the crystal. $\boldsymbol{\rho}(\boldsymbol{r})$ is the electron density of the atom at position $\boldsymbol{r} = (x, y, z)$. $\boldsymbol{\rho}$ $(\boldsymbol{r})$ and $\boldsymbol{F}(\boldsymbol{k})$ is a Fourier transform pair and exists in different coordinate systems of real (x, y, z) and reciprocal (h, k, l) space, respectively. If we can obtain the structure factor by exposing a crystal to X-rays, $\boldsymbol{\rho}(\boldsymbol{r})$ can be calculated from diffracted signals of $\boldsymbol{F}(\boldsymbol{k})$ by inversed Fourier transform (FT^{-1}):

$$\rho(r) = FT^{-1}(F(k)) = \int F(k) \exp 2\pi i(-kr)dk \qquad (2)$$

Unfortunately, only intensity, which is proportional to the square of the complex amplitude as $I \approx |F(k)|^2$ diffracted by a crystal, can be measured; thus, electron density of atoms cannot be calculated directly using expression 2. This is a famous phase problem in crystallography, which makes structure determination of macromolecules very difficult, and the most developments of protein crystallography in the past half century are primarily around phasing. The first protein structures of myoglobin and hemoglobin were solved in 1961 [6, 7].

In the past decade and a half, structural genomics projects have advanced the development of macromolecular crystallographic methods and techniques in both hard- and software, including sample expression, preparation, crystallization, diffraction data collection, and structure determination. Such powerful and sophisticated developments with the use of synchrotron radiation and the amazing progress of computer hardware have dramatically reduced difficulties and the time required to solve structures. Today, undergraduate students can solve protein structures with little training in crystallographic techniques, whereas two decades ago, the successful determination of a de novo protein structure may take the period from a Master's to PhD degree course and warrant a high-impact publication.

Figure 1 shows an overview of structural analysis after obtaining a crystal which is applicable to a diffraction experiment. Data collection depends on the X-ray source (X-ray beam), diffraction meter, and collection conditions such as temperature, wavelength, exposure time, oscillation angle, and distance between the crystal and detector. The use of third-generation synchrotron radiation and significant progress of detector advance the measurement of diffraction data, consequently allow us to collect good-quality data with high

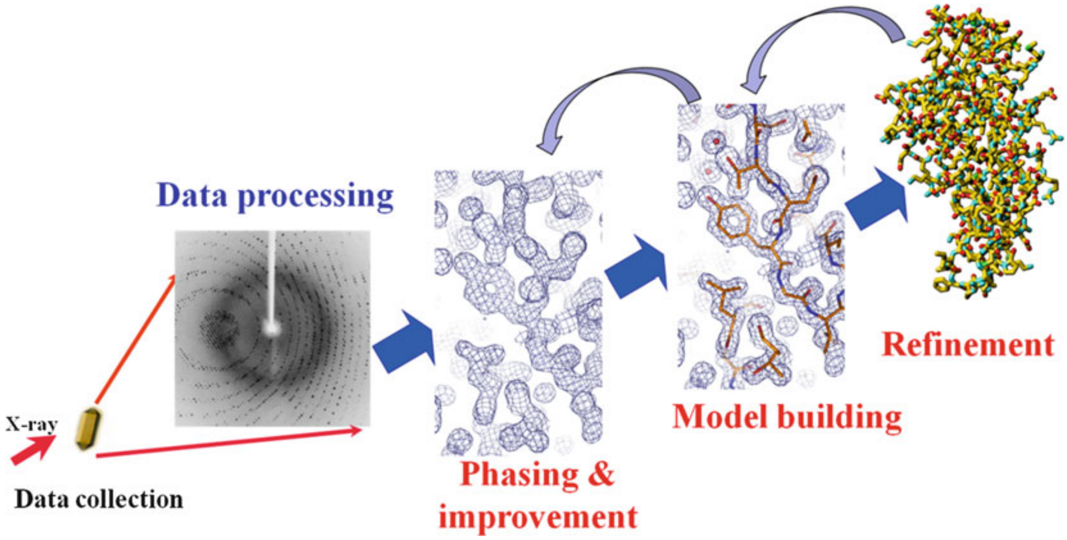


Fig. 1 Process of X-ray structural analysis after obtaining crystals

signal/noise (S/N) ratio in several minutes, resulting in increased success rate of structure determination.

The last four steps (data procession, phase calculation and improvement, model building, and refinement) of structure determination shown in Fig. 1 are performed on a computer using software programs. A variety of powerful and sophisticated programs have been developed. Among them, the typical, famous, and commonly used program is *c*ollaborative *c*omputational *p*roject *n*umber **4** (CCP4) suite, which collects many programs, associated data, and subroutine libraries [8].

As mentioned above, structural genomics projects have advanced the development of methods and software systems that have accelerated and automated various stages of structural analysis. Each step of structure determination on a computer (Fig. 1) can be performed almost automatically. Even the last step, structural refinement, which is time consuming and requires a great deal of expertise with crystallography, can be performed semi-automatically, recently [9, 10]. For example, the **CCP4** suite has been expanded with a computer graphics user interface (**CCP4i**) and several automation pipelines [11, 12]. Other excellent software packages such as **HKL3000** [13], **PHENIX** [14, 15], **Coot** [16], and **XDS** [17] have been developed. Here, advanced software packages will be introduced following the process stage of structure determination shown in Fig. 1.

2 Diffraction Data Processing

High-quality diffraction data are required for all structure determination calculations. Data are collected using the oscillation method, which is the most basic and important step during structural analysis. The data process includes indexing reflections, integrating intensity detected on frames, scaling integrated reflections recorded on many frames, and finally merging equipollent reflections to a set of unique reflections.

The used software histogram of PDB (Protein Data Bank) statistics shows that the data processing programs used for most deposited structures are **HKL2000** [18], **Mosflm** [19], and **XDS** [17]. The **HKL2000** program performs all data processing described above, and **Mosflm** and **XDS** are usually used for index and integration steps. The **SCALA** [20] program in the **CCP4** suite or **XSCALE** [17] is used for scaling and merging reflections. **HKL2000** and **Mosflm** (**iMosflm** [21]) data processing can be monitored in real time with GUI during the integration of all reflections, and **XDS** processes data nearly automatically without manual interruption. **HKL2000** is a powerful program for indexing using one frame, but the user must decide carefully the sport size and reflection range for integration. **Mosflm** and **XDS** dynamically and automatically determine the integrating range and mosaicity and also perform 3D profile fitting for each frame with excellent algorithms. Even if reflection patterns appear as if they are connected, as shown in Fig. 2, it is able to obtain good-quality data by using **XDS**. It is better to process data using one or more programs in order to obtain quality data for difficult data processing cases. The resolution range of processed data for the next calculation should be determined by comprehensively considering the parameters of completeness, I/σ(I) (at least >1.5), Rsym, and redundancy/multiplicity. Moreover, the twin situation must be checked.

3 Phasing

Electron density $\boldsymbol{\rho}(\boldsymbol{r})$ cannot be obtained directly from expression 2 because the phases (α_k) of the structure factor are lost during measurement. The first calculation step for structure determination is the phase reconstruction of the structure factor after data processing. The phase reconstruction methods can be divided into the following two categories:

1. Molecular replacement (MR) method
2. Substructure determination method using signal of heavy atoms:
   Multiple isomorphous replacement (MIR) (first generation)

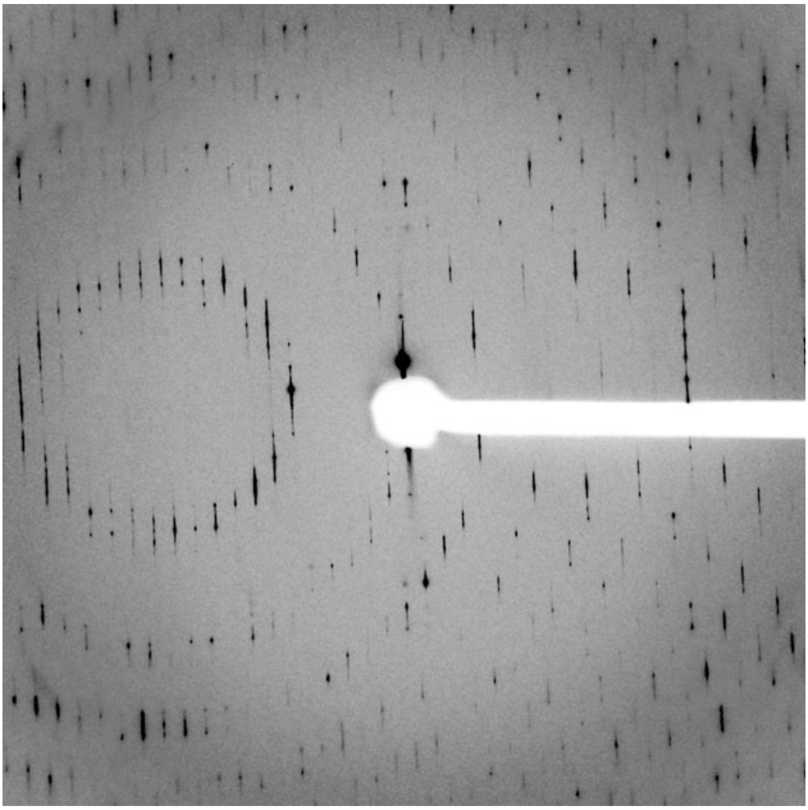


Fig. 2 Diffraction image of an octameric pore crystal of staphylococcal γ-hemolysin. The crystal is twin and belongs to space group $C222_1$ with unit cell parameters of $a = 206.5$, $b = 206.1$, and $c = 190.3$. Diffraction data were collected at beamline BL41XU of SPring-8 and processed using XDS. The structure was solved by molecular replacement and refined with twin factor $\alpha = 0.44$ at 2.49 Å resolution [22]

Multiple-wavelength anomalous diffraction (MAD) (second generation)

Single-wavelength anomalous diffraction (SAD) (third generation)

3.1 Phasing by Molecular Replacement

MR is a method to determine the target protein structure by using known structural homology based on the assumption that the sharp and inner structures are similar between the target and known structure. As shown in Fig. 3, the basic idea of this method is to find orientation ($\theta1, \theta2, \theta3$) and position ($t1$, $t2$, $t3$) of molecules in the target crystal using known structure as a search model. The phasing probability of the MR method is dependent on the similarity of the search model to the target protein. Generally, it can solve the structure using the MR method if the sequence identity is >30 %. The conventional MR method uses the Patterson function $\mathrm{P}(\boldsymbol{u})$ (expressions 3, 4, 5 and 6) to search for the orientation ($\theta1$, $\theta2$, $\theta3$) and position ($t1$, $t2$, $t3$) of molecules in the target crystal [23]:

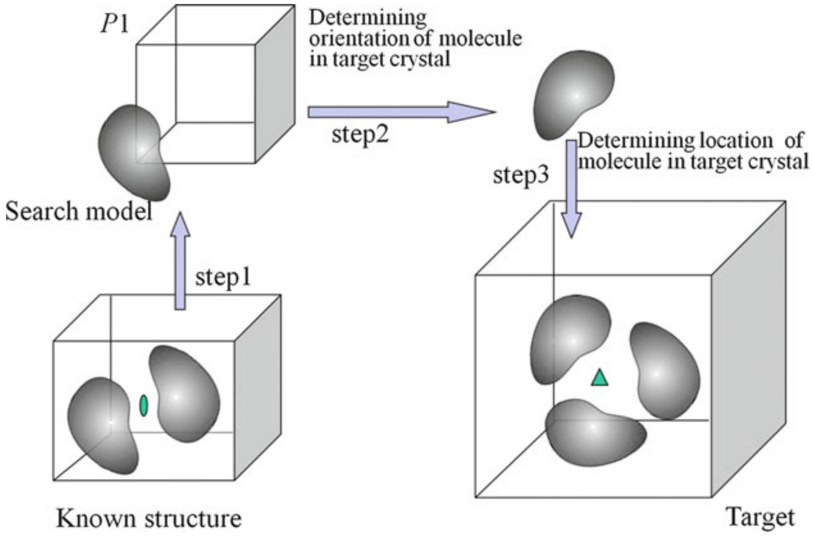

P1
Determining orientation of molecule in target crystal
step2
Determining location of molecule in target crystal
step3
Search model
step1
Known structure
Target

Fig. 3 Process of MR. *Step 1* is to build a search model from known structure and make a model crystal with space group of P1. *Step 2* is to search rotation parameter (θ1, θ2, θ3), and *step 3* is to find the position of molecules in target crystal with contact check

$$R(\theta 1, \theta 2, \theta 3) = \left(\frac{1}{V}\right) \int_U P(u) P_M(ru) du \tag{3}$$

$$T(t1, t2, t3) = \left(\frac{1}{V}\right) \int_U P(u) P_{\text{rot\_anw}}(u + t) du \tag{4}$$

Here, the Patterson function P($\boldsymbol{u}$) is a convolution described by a different space system (u, v, w) that is a vector set between atoms and can be calculated from observed intensity $\boldsymbol{I}(\boldsymbol{k})$ or model $\boldsymbol{\rho}(\boldsymbol{r})$ as shown below:

$$P(u) = \left(\frac{1}{V}\right) \int_U \rho(x) \rho(x + u) dx \tag{5}$$

$$\begin{aligned} P(u) &= \rho(r) \bigotimes \rho(-r) \\ &= FT^{-1}(F(k)) \bigotimes FT^{-1}(F(-k)) \\ &= FT^{-1}(F(k) \cdot F(-k)) = FT^{-1}\left(F(k) \cdot F^*(k)\right) \\ &= FT^{-1}\left(|F(r)|^2\right) = FT^{-1}(I(k)) \end{aligned} \tag{6}$$

$$
\begin{aligned}
P(u) &= \left(\frac{1}{V}\right)\sum\nolimits_k I(k)\exp(-2\pi ku) \\
&= \left(\frac{2}{V}\right)\sum\nolimits_k I(k)\cos\left(-2\pi ku\right)
\end{aligned}
\tag{7}
$$

Several algorithms and programs of MR method have been developed. The typical MR method software using the Patterson function is **AMoRe** [24] in the **CCP4** package. The advantages of **AMoRe** are fast calculation and easy to adjust parameters. The more powerful **MOLREP** [25] and **PHASER** [26] software in the **CCP4** package, which combine the Patterson function with the maximum-likelihood method, are often used. A search model is automatically optimized based on sequence arrangement in **MOLREP**. **PHASER** calculates phases automatically and more effectively by searching for the small domain. **PHENIX** developed a new useful program called **SCULPTOR** [27] in which the answer model can be modified after the rotation and translation parameters are found to overcome the conformation changing problem of domain or a partial structure between the search model and the target protein.

Figure 4 shows an example of structure determination using MR method. Translational initiation factor eIF5B is a multi-domain protein consisting of four domains with a high flexibility property. The structure of the eIF5B-1A complex was attempted to be solved by the MR method using the whole structure of eIF5B [28] with **MOLREP** and **PHASER** in the **CCP4** package and **PHENIX**. However, no answers were obtained. Finally structure was solved by a domain search using the following steps: (1) domains I–II were used as a search model and the solution was found; (2) then, domain III was used as a further search model with the answer from step (1) as a fixed molecule in the target crystal structure; (3) finally, the eIF5B domain IV was used as a search model with the answers of domains I–II and III as fixed molecules.

Strategy of Molecular Replacement

1. Prepare a similar search model using sequence information.
   (a) Search for homologous protein structures from the Protein Data Bank (PDB).
   (b) Construct a model using 3D structure prediction software on a web server such as *Phyre2* (http://www.sbg.bio.ic.ac.uk/phyre2) [29], *SWISS-MODEL* (http://swissmodel.expasy.org/) [30], and Rosetta [31].
   (c) If the target protein consists of multiple domains, it may be better to use each homology domain as search model and calculate respectively.
2. Search for the orientation and position of the target protein.
   (a) Use different programs and different models.

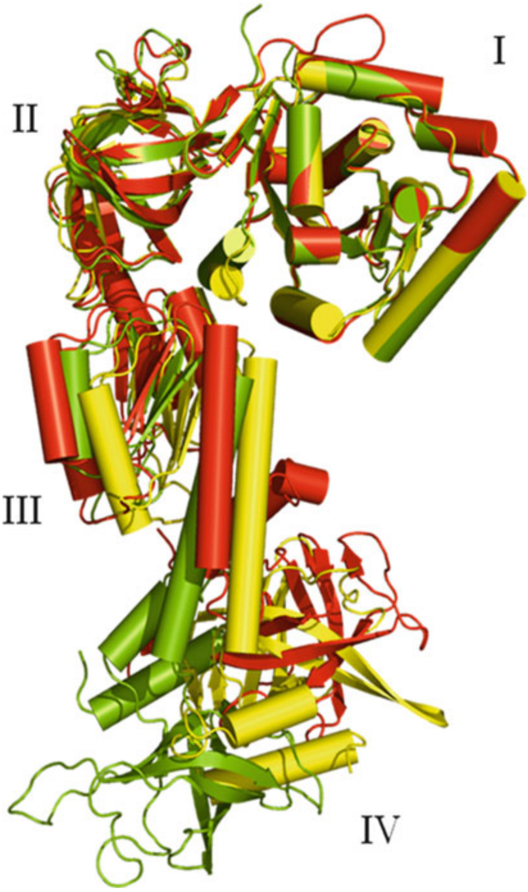


Fig. 4 Structures of eIF5B [28]. The structures of eIF5B were superposed by using domain G (I). I, II, III, and IV are marks of domains G, II, III, and IV, respectively. *Red*, eIF5B (3WBI); *Green* and *yellow*, two eIF5B molecules in eIF5B-1A complex (3WBK)

(b) Use separate domains to search one by one. If one domain or one molecule is found, then fix the answer model and continue to find the other parts.

(c) If the target is bigger, for example, >500 residues, use a search model that only contains the main chain and Cβ atoms as it may be effective.

3. Modify the answer model during rigid-body refinement.

   If the answers seem to be found, however rigid-body refinement does not give a low R free factor (>45–50 %), **SCULPTOR** program packaged in the **PHENIX** software system may be effective to modify the answer structure [32].

4. Change resolution range for calculation.

   Generally, the low-resolution part of diffraction data contributes to the connection of a peptide chain in electron density map, whereas the high-resolution part presents detailed information for atoms such as bond lengths, bond angles, and torsion angles (Ramachandran). Therefore, a low-resolution region of <3.0 Å is generally used for the MR method.

5. Remove model bias by averaging the molecules in AU.
   Because the MR method uses a search model to calculate phases, the problem of model bias is introduced. The electron density map seems to fit each part of the model well in most cases even if wrong answers are used. Therefore, averaging molecules is a powerful method to distinguish wrong answers if there are multiple molecules in an asymmetric unit (AU).

3.2 Phasing Based on Substructure Determination (Experimental Phasing)

Methods using the substructure of heavy atoms (here, we define a heavy atom as the atomic number larger than oxygen) are key to solving the phase problem of the macromolecular structure by experimental measurements. Phase calculations can be considered in two parts: substructure determination of heavy atoms and phase calculation using the substructure. Methods such as multi- and single-isomorphous replacement (MIR and SIR), MAD (Multiple-wavelength Anomalous Diffraction) and SAD (Single-wavelength Anomalous Diffraction), direct method, or their combinations (e.g.,, SIRAS method) were developed. In all of these methods, the substructure of heavy atoms is determined first by using their scattering single for special wavelength, which is much larger than that of the main protein compounds (O, C, N, and H atoms) (Fig. 5, expressions 8 and 9).

Here, $\boldsymbol{F_{PH}}(\boldsymbol{k})$ is the structure factor of a protein with a heavy atom, called the derivative, whereas $\boldsymbol{F_P}(\boldsymbol{k})$ is the structure factor of the protein. f_j is the scattering factor of atom j and is proportional to its electron number. α_P can be calculated based on expression 8, 9, 10 and 11 (Figs. 5a and 6) if the site ($\boldsymbol{r}_H$) can be estimated:

$$F_P(k) = \sum\nolimits_j f_j \exp 2\pi i k r_j \tag{8}$$

$$\begin{aligned} F_{PH}(k) &= F_P(k) + F_H(k) \\ &= \sum\nolimits_j f_j \exp 2\pi i k r_j + f_H \exp 2\pi i k r_H \end{aligned} \tag{9}$$

$$\begin{aligned} F_{PH}^2 &= (|F_P|\exp(i\alpha_P) + |F_H|\exp(i\alpha_H))^2 \\ &= F_P^2 + F_H^2 + 2F_P F_H \cos\left(\alpha_p - \alpha_H\right) \end{aligned} \tag{10}$$

$$\alpha_p = \alpha_H \pm \cos^{-1}\left(\frac{F_{PH}^2 - F_P^2 - F_H^2}{2F_P F_H}\right) \tag{11}$$

Moreover, when heavy atom contributes to the scattering with an anomalous signal for special wavelength, f_H should be described

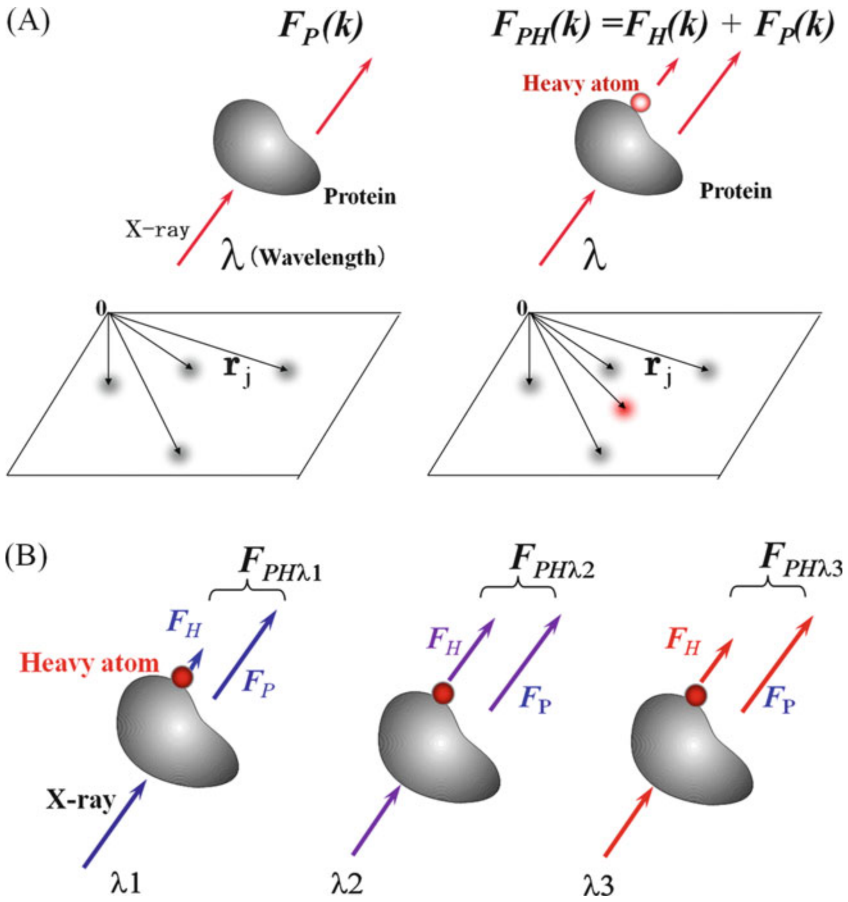


Fig. 5 Concept of substructure determination. (**A**) is for MIR method, and (**B**) is for MAD method

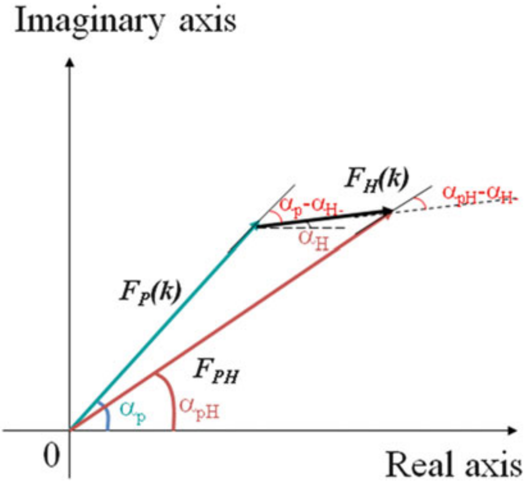


Fig. 6 The structure factor triangle for native protein and derivative

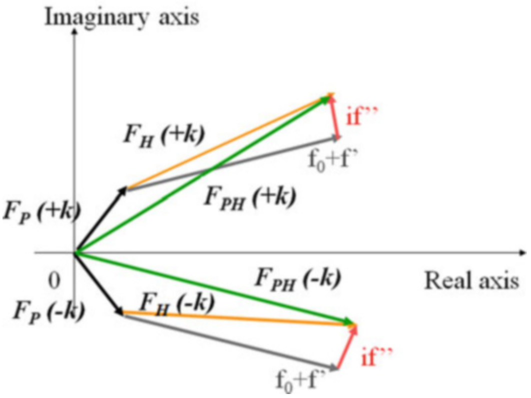


Fig. 7 The structure factor with anomalous signal. ***F*(*k*)** and ***F*(−*k*)** is called Friedel pair with $|F(k)| \neq |F(\text{-}k)|$

by three parts f_{H0}, f'_H, *and* f''_H as expression 12 (Fig. 7). f_{H0} is a scattering at signal-independent wavelength, which is only proportional to the atomic number and is used in the isomorphous replacement method (the last two parts are much smaller and can be ignored). f'_H and f''_H represent the anomalous scattering at signal-dependent wavelength. Due to the development of the beamline at synchrotron radiation facilities, f'_H and f''_H can be measured with a high S/N ratio and the SAD method (Fig. 5b) has become the mainstream method at pressent.

$$F_H(k) = f_H \exp 2\pi i k r_H = \left(f_{H0} + f'_H + i f''_H\right) \exp 2\pi i k r_H \quad (12)$$

When f'_H and f''_H are taken as contributors to f_H expression 11 can be transformed into

$$\alpha_p = \alpha_H \pm \cos^{-1}\left(\frac{F^2_{PH\lambda 1} - F^2_{PH\lambda 2} - F^2_H}{2F_{PH\lambda 1}F_H}\right) \quad (13)$$

and

$$\alpha_p = \alpha_H \pm \cos^{-1}\left(\frac{F^2_{PH\lambda}(+k) - F^2_{PH\lambda}(-k) - F^2_H}{2F_{PH\lambda}F_H}\right) \quad (14)$$

for the MAD and SAD methods, respectively.

3.2.1 Substructure Determination

Substructure determination is typically performed using the Patterson function P(***u***). Different from the MR method (expression 7), the Patterson function P(***u***) uses the differences in scattering signal (intensity I) between the native and derivative for MIR (expression 16), between wavelengths for MAD (expression 16), or an anomalous signal (Friedel pair) at a special wavelength for MAD/SAD (expression 17), as the Fourier transform coefficient:

$$P(u) = \left(\frac{1}{V}\right)\sum\nolimits_k (I_{PH}(k) - I_P(k))\exp(-2\pi k u) \tag{15}$$

$$P(u) = \left(\frac{1}{V}\right)\sum\nolimits_k (I_{\lambda 1}(k) - I_{\lambda 2}(k))\exp(-2\pi k u) \tag{16}$$

$$P(u) = \left(\frac{1}{V}\right)\sum\nolimits_k (I_{\lambda}(+k) - I_{\lambda}(-k))\exp(-2\pi k u) \tag{17}$$

Such Patterson functions produce strong peaks contributed by heavy atoms compared with those contributed by the main protein compounds, allowing the identification of the heavy atom sites. Harker found that the self-vectors of heavy atoms in the Patterson map appear on a special section that are independent of the site coordinates but are derived from crystal symmetry. For example, in the case of the $P2_1$ space group, there are two equivalent molecules: $x_1 = (x, y, z)$ and $x_2 = (-x, y + 1/2, -z)$, the self-vector is $x_1 - x_2 = (2x, -1/2, 2z)$, and the self-vector peak always appears on the section of $v = 1/2$ in the Patterson map P ($\boldsymbol{u} = \{u, v, w\}$). Such special sections are named Harker sections. The site coordinates (x, z) can be estimated as $x = u/2$, $z = w/2$ using the self-vector peak, and the y of the first heavy atom is free (generally, the y value is set to 0).

The **SHELX** [33] program, which was developed for small-molecule structure determination at the early stage, has been expanded to **SHELXCDE** [34, 35] (**C**, data preparation; **D**, substructure determination; **E**, phase calculation including phase improvement) for macromolecules. The advantage of **SHELXCDE** is that it determines substructure sites (heavy atoms) by using the Patterson function $P(u)$ combined with the direct method. Such a combined method is effective to avoid misfounding of sites in which cross vectors of atoms appear strongly on the Harker section, as the first site of a heavy atom is estimated using a stronger peak on the Harker section.

3.2.2 Phase Calculation Including Phase Improvement and Model Building

The phases can be calculated as expression 11 after obtaining the substructure of the heavy atoms; however, two answers will be obtained. Generally, two or more kinds of heavy atom derivatives are necessary to obtain a unique answer for MIR. The concept of the SIRAS/MAD/SAD method is similar to that of the MIR method, but it uses different scattering signals such as anomalous diffraction ($I(\boldsymbol{k}) \neq I(-\boldsymbol{k})$) according to the special wavelength or scattering differences of different wavelengths. SIRAS is a combination of SIR (difference between native and derivative) and anomalous diffraction. While different wavelengths are used to discriminate answers in MAD, the SAD method attempts to use mathematic power to address two-answer problem.

Generally, the derivative is not actually isomorphous of native crystal, as the crystal is damaged in a soaking experiment, and the resolution of the macromolecular crystal diffraction data is limited (normally 2.5–3.5 Å). Moreover, various errors are included in diffraction data, such as X-ray damage, systematic errors from the diffraction meter, arrangement of the 2D detector, and data processing. Thus, the structure factor triangle shown in Fig. 6 should have an error part, ε (Fig. 8). Therefore, the phases calculated by expression 11 are very noisy, and even electron density calculated by such phases cannot be interpreted without improving the phase (Fig. 9). Methods of phase improvement such as solvent flatting, histogram matching, non-crystallographic symmetry averaging, and averaging the inter-crystal form were developed two decades ago. The new idea developed recently is to combine these methods with model building and using built partial model.

SHARP [37] and **PHENIX** (**PHENIX_AutoSol** [38]) have been developed recently for experimental phasing. These two software packages implement excellent algorithms combined with the maximum-likelihood method for calculating phases after determination of substructure with phase improvement. The advantage of **SHARP** is that it not only uses the maximum-likelihood method in the phase calculation, but also the amplitude of the structure factor through error, ε, is considered in the calculation. The **SHARP** program is very effective for the MIR and MAD methods. **Auto_SHARP** makes an automatic determination pipeline, including **SHELXCD** as a substructure determination part, **Solomon**

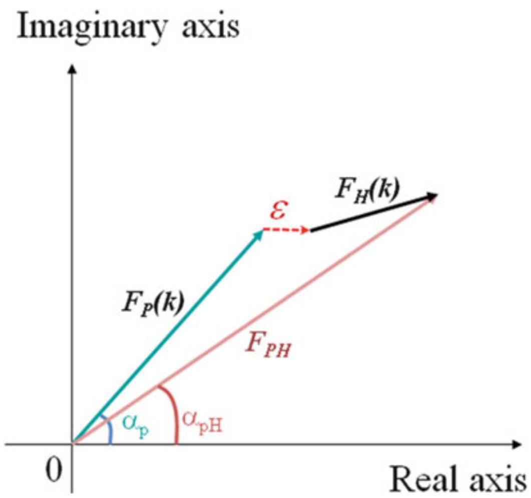


Fig. 8 The structure factor triangle with error part for native protein and derivative

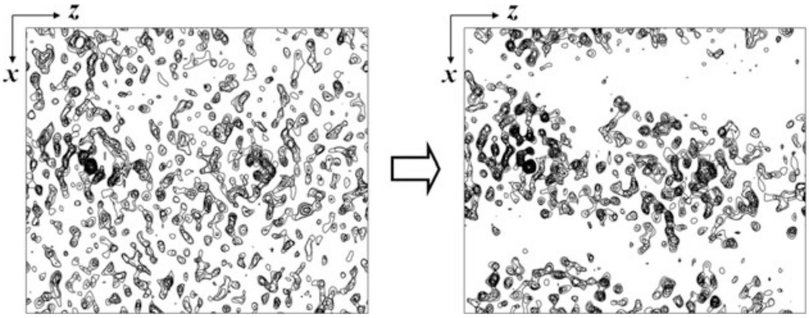


Fig. 9 Phase improvement of IDH [36]. 2D projected electron density map at initial (*left*) and improved (*right*) stage

[39] as a phase improvement part using solvent flatting, and the **ARP/wARP** [40] program for model building. **PHENIX_AutoSol** is a program package that determines structures automatically including substructure determination, phase calculation, and phase improvement with modeling step.

The Strategy of Phasing Based on Substructure

1. Substructure determination

   SHELXCDE is used to determine the substructure for all methods described above, and the parameters CC and CCweak in the *.res output file should be checked. If CCweak is >15 %, the structure can be solved, and the model may be built automatically to ~70 % (from the data of author's laboratory). **SHELXCDE** chooses the resolution range automatically depending on a single of $\Delta d/I > 0.8$. Moreover, success rate may be increased by adjustment of parameters. For example, if the substructure sites are >10, the NTRY parameter should be set to more than 500. If the substructure sites are >20, NTRY should be set to at least 1,000.
2. Phase calculation

   It is a way to calculate phases with the **SHARP** or **PHENIX_AutoSol** programs after substructure determination. The substructure coordinates can be inputted into **SHARP** graphics interface, and the PDB file of the heavy atom sites is used in **PHENIX_AutoSol**. Generally, the coordinate, occupancy, and the B-factor of the heavy atom sites should be refined during phasing. In difficult cases, it may be better to refine the coordinate and B-factor with fixed occupancy of the heavy atom sites.

4 Structural Refinement

The phases of the structure factor cannot be measured directly but can be estimated indirectly using the MR method with structural homology or using the heavy atom substructure. Such phases contain various errors from diffraction data processing, substructure determination, and phase calculation. Although the phases can be improved as described above, phase errors and resolution limitations of the measured diffraction data result in a poor-quality electron density map which is insufficient to build a full and exact model. Therefore, the structure determination process requires another step called structural refinement in which close agreement is achieved between the observed and calculated amplitudes of the structure factors by removing conformation errors, completing missed fragments remaining in the initial model, and adding other molecules that exist in the crystallization reagents or samples.

4.1 Mathematics Used for Refinement

The R factor, which evaluates whether the calculated amplitude of the structure factor ($|\boldsymbol{F}_{\mathrm{cal}}(\boldsymbol{k})|$) is in agreement with that of observed ($|\boldsymbol{F}_{\mathrm{obs}}(\boldsymbol{k})|$), is shown in expression 18:

$$R = \frac{\sum_{k} \left| |F_{\mathrm{obs}}(k)| - c|F_{\mathrm{cal}}(k)| \right|}{\sum_{k} |F_{\mathrm{obs}}(k)|} \tag{18}$$

If the current structure is close to a crystal structure, the R factor should become smaller. The purpose of refinement is to make the R factor smaller. Various calculation methods and programs such as **SHELXL** [41], **TNT/BUSTER** [42, 43], **X-PLOR** [44]/**CNS** [45], **REFMAC5** [46], and **PHENIX** [47] have been developed for refinement. Among methods of these programs, the traditional mathematics used for refinement is the least squares method to minimize Q which represents the difference between the actual measured amplitude of the structure factor ($\boldsymbol{F}_{\mathrm{obs}}$) and the calculated value ($\boldsymbol{F}_{\mathrm{cal}}$) (expression 19). Here, $w(\boldsymbol{k})$ is weight estimated from the standard error $\sigma(I(\boldsymbol{k}))$, which is calculated from diffraction data, and $\boldsymbol{m}$ is a scale factor:

$$Q(X) = \sum_{k} w(k)\left(|F_{\mathrm{obs}}(k)| - m|F_{\mathrm{cal}}(k, X)|\right)^2 \tag{19}$$

In the simple least squares method (expression 19), the basic adjustment parameters, X which represents position (x, y, z), the B-factor (B, vibration around the central position of an atom), and occupancy (occ) of each atom, are used to calculate the structure factor $\boldsymbol{F}_{\mathrm{cal}}$. Compared with small molecules, refinement of macromolecules using the simple least squares method is much more difficult due to large errors in both phases and coordinates and the low ratio (3–5) of the number of diffraction data via refined parameters caused by resolution limitations of macromolecular crystals. Therefore, other methods have been developed to expand the convergence radius of the least squares method.

Ideal stereochemistry parameter values, such as bond distances, bond angles, and torsion angles of peptides, can be estimated based on the results of small-molecule structural analysis. Adding ideal stereochemistry parameters into the simple least squares method as a restriction has been considered (expression 20) [48]. Moreover, methods to reduce the refinement parameters, define a peptide as a rigid body, and only refine the torsion angles (ϕ, ψ) of peptides have also been used:

$$Q(X)=\sum_{k} w(k)(|F_{\text{obs}}(k)|-m|F_{\text{cal}}(k,X)|)^2+\text{restrictions} \tag{20}$$

Rather than using ideal stereochemistry parameter values in the restrictions, Jack and Levitt improved the restricted least squares method and proposed a new least squares method with a potential energy minimization function in 1978 [49]. This potential energy includes the potential bond distance stretch, bending of bond angles, torsional potential, and van der Waals interactions. Using this method, an energy, E, term has been added to the minimization function, as shown in expression (21). Here, w_x is weight estimated from the standard error $\sigma(I(k))$ of the diffraction data:

$$\begin{aligned} Q(X) = w_x \sum_{k} w(k)(|F_{\text{obs}}(k)| - m|F_{\text{cal}}(k, X)|)^2 \\ + (1 - w_x) * E \end{aligned} \tag{21}$$

Brunger further developed the **X-PLOR/CNS** program with a molecular dynamics algorithm called simulated annealing to make the convergence radius of refinement large and to avoid falling into a local minimum [44, 45]. In this algorithm, the energy is minimized by simulating atomic movements with a high enough temperature to exceed the energy barrier followed by slow cooling.

After the 1990s, the maximum-likelihood method was introduced for refinement, such as in the **REFCAC5** program [46]. The basic principle of maximum likelihood is that the best model should produce the maximum probability P to obtain current diffraction data. In other words, this method modifies the parameters of the model to obtain maximum probability P based on diffraction data:

$$P(|F_{\text{obs}}(k)|; F_{\text{cal}}(k, X)) = \int_{0}^{2\pi} p(|F_{\text{obs}}(k)|, \alpha, F_{\text{cal}}(k, X)) d\alpha \tag{22}$$

The least squares method is a special case of the maximum-likelihood method. The advantages of the maximum-likelihood method are that it converges more correctly, and that it has less computational complexity than that of the molecular dynamics algorithm. An energy minimization item has recently been added to the maximum-likelihood method.

4.2 Free R Factor

It is expected that the R factor will become smaller as refinement progresses. However, a smaller R factor does not necessarily guarantee that refinement has been done in correct way. As the ratio of observed data to refined parameters is small, and restrictions are added to the refinement algorithms, it may lead to an incorrect structure with a low R factor (over-refining). To objectively evaluate refinement results, Brunger proposed a free R factor based on a cross-validation statistical method, which is a technique to estimate the performance of a predictive model [50]. The R factor calculation is divided into two parts: R work and R free factors. Both factors are calculated in the same way (expression 18) but use different datasets W and T, respectively. After data processing, the diffraction data are divided into two datasets with a random choice: 90–95 % working (W) measured diffraction data for all calculations of determination and the remaining 5–10 % data (T) used only for testing. Although R work is calculated from working data, R free is calculated from test data. As calculating the free R factor is independent of refinement, the refinement result can be evaluated more correctly using the R free factor.

4.3 Refinement Programs

Various refinement methods and programs have been developed, and each program has its own characteristics based on the computational algorithm. **SHELXL** has the advantage of very high-resolution refinement with anisotropic B-factor, but AUTO_**BUSTER** may be more effective for a good fit between $|\boldsymbol{F}_{\mathrm{obs}}(\boldsymbol{k})|$ and $|\boldsymbol{F}_{\mathrm{cal}}(\boldsymbol{k})|$, for low-resolution refinement (<3.5 Å) [51, 52]. Using the molecular dynamics algorithm (simulated annealing), **CNS** is useful to reduce model bias in cases in which the structure is solved by MR. **REFMAC5** in the **CCP4** package is widely used. **PHENIX_refine** is a refinement program developed more recently, which includes improved valid methods and new ideas described below.

Strategy of Refinement

1. Using bulk solvent correction

   One of the advancements in refinement method development is the introduction of the bulk solvent correction in the calculated structure factor (expression 23) for the effect of a disordered solvent [53], as seen in **CNS**, **REFMAC5**, and **PHENIX_refine**:

$$F_{\mathrm{cal}}(k) = F_p(k)^p\left(1 - c_{\mathrm{sol}}\exp\left(-B_{sol}S^2/4\right)\right) \qquad (23)$$

   Here, c_{sol} is the ratio of average electron density of a solvent/protein. Introduction of the bulk solvent correction into refinement actually improves agreement between the observed data and the model, resulting in lower R_free/R_work, especially for a low-resolution range (<10 Å). The **REFMAC** program includes a bulk solvent correction in the σ_A estimate.

2. Using TLS parameters

   The B-factor of an atom represents the vibration around its central position, and it can be defined isotropically (as a sphere by parameter B) or anisotropically (in three dimensions by six parameters u11, u22, u33, u12, u13, and u23). The anisotropic B-factor is appropriate to describe the vibration of atom around a central position; however, it increases the refinement parameters enormously. Therefore, the isotropic B-factor is usually used at resolutions <1.8 Å. Schomaker and Trueblood proposed a description of anisotropic motion using fewer parameters called Translation, Libration, and Screw (TLS) [54]. The **REFMAC5** program implements early TLS parameters (20 parameters/groups) of grouped atoms (total molecule, domain, or fragment) into refinement to cover the anisotropic motion problem with a decrease in the total number of parameters [55].

3. Using the H atoms

   Data in the PDB show that the most macromolecular crystals (80 %) diffract to resolution of 1.7–3.2 Å and hydrogen atoms are unclearly shown on an electron density map (Fo-Fc or 2Fo-Fc map). Consequently, refinement is usually performed without hydrogen atoms. Moreover, the low number ratio of diffraction data via the refined parameters is also a reason for excluding hydrogen atoms from the refined parameters. A new idea using hydrogen atoms to avoid the crash between atoms has been added to the refinement method and is effective for adjusting position parameters (x, y, z) of atoms during the refinement calculation. In this case, hydrogen atoms are added to the coordinate file with occ of zero.

4.4 Automatic Refinement Process

A huge calculation is required to refine a protein structure as many parameters and much data are involved. With the development of computers, the computation time of one cycle (one big cycle including refinement of atomic coordinates and B-factors) of refinement has been shortened from several days to minutes or hours depending on the size of the protein and data resolution. Moreover, even if there are various outstanding refinement programs, the convergence range of the atomic coordinates of refinement is narrow, as the ratio of the protein crystal diffraction data to refinement parameters is small. If the model does not match the electron density map well or if there are some missing parts, manual fitting and building with a computer graphics program is necessary after refinement calculation (Fig. 10). This manual intervention requires a great deal of expertise in crystallography. Furthermore,

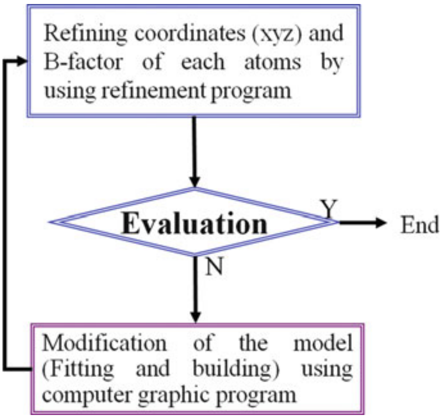

Refining coordinates (xyz) and B-factor of each atoms by using refinement program
Evaluation
Y
End
N
Modification of the model (Fitting and building) using computer graphic program

Fig. 10 Refinement process

each protein/nucleic acid residue must be checked one by one in each cycle of refinement; therefore, refinement is a time-consuming step in the structure determination process.

The Coot [16] computer graphics program has been developed to become powerful for fitting and partly model building. However, perfecting manual operation is considerably dependent on the operator's skill level, particularly in the cases of relatively low-resolution and huge molecules. If manual intervention can be automated, it will certainly speed up and save labor for structural analysis. We have developed an automatic refinement software package called LAFIRE (local correlation-coefficient-based automatic fitting for refinement) to realize manual intervention-free refinement [9, 10]. This software package is designed to perform the whole process of protein/nucleic acid structural refinement automatically with the **PHENIX_refine**, **REFMAC5**, **BUSTER**, or **CNS** refinement programs from an initial model that can be approximate, fragmentary, or even only a main chain. A fully or semi-automatic refinement process can be realized within a few hours or days using **LAFIRE**.

Acknowledgments

We thank Professor Isao Tanaka for helpful advices and provision of some figures, Drs. Yong Zhou and Yamashita for developing program LAFIRE, and all members of X-ray structure biology laboratory, Faculty of Advanced Life Science, Hokkaido University, JAPAN for providing the information of structure determinations. This work was partly supported by Grants-in-Aid for Scientific Research in a Priority Area and the National Project on Protein Structural and Functional Analysis, from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Official Nobel Prize site http://www.nobelprize.org/nobel_prizes/physics/laureates/1914/
2. Bragg WH (1912) X-rays and crystals. Nature 90:360–361
3. Bragg WL (1912) The specular reflection of x-rays. Nature 90:410–410
4. Official Nobel Prize site http://www.nobelprize.org/nobel_prizes/physics/laureates/1915/
5. Bernal JD, Crowfoot D (1934) X-ray photographs of crystalline pepsin. Nature 133:794–795
6. Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Dvies DR (1960) Structure of myoglobin: a three-dimensional fourier synthesis at 2 Å resolution. Nature 185:422–427
7. Perutz MF, Rossmann MG, Cullis ANNF, Muirhead H, Will G, North ACT (1960) Structure of hæmoglobin: a three-dimensional fourier synthesis at 5.5-Å. Resolution obtained by X-ray analysis. Nature 185:416–422
8. Collaborative Computational Project, Number 4 (1994) The CCP4 suite: programs for protein crystallography. Acta Crystallogr D50:760–763
9. Yao M, Zhou Y, Tanaka I (2006) LAFIRE: software for automating the refinement process of protein-structure analysis. Acta Crystallogr D62:189–196
10. Yamashita K, Zhou Y, Tanaka I, Yao M (2013) New model-fitting and model-completion programs for automated iterative nucleic acid refinement. Acta Crystallogr D69:1171–1179
11. Potterton E, Briggs P, Turkenburg M, Dodson E (2003) A graphical user interface to the CCP4 program suite. Acta Crystallogr D59:1131–1137
12. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, 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 D67:235–242
13. Minor W, Cymborowski M, Otwinowskib Z, Chruszcz M (2006) *HKL*-3000: the integration of data reduction and structure solution – from diffraction images to an initial model in minutes. Acta Crystallogr D59:45–49
14. Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002) *PHENIX*: building new software for automated crystallographic structure determination. Acta Crystallogr D58:1948–1956
15. Adams PD, Afonine PV, Bunkóczi 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 D66:213–221
16. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr D60:2126–2132
17. Kabsch W (2010) XDS. Acta Crystallogr D66:125–132
18. Otwinowski Z, Minor W (1997) Processing of x-ray diffraction data collection in oscillation mode. Methods Enzymol 276:307–326
19. Leslie AGW (1993) Auto-indexing of rotating diffraction images and parameter refinement. In: Sawyer L, Isaacs N, Bailey S (eds) Proceeding of the CCP4 study weekend. Daresbury Laboratory, Daresbury, pp 44–51
20. Evans PR (1997) Scaling of MAD data. In: Wilson KS, Davies G, Ashton AW, Bailey S (eds) Proceedings of CCP4 study weekend. Daresbury Laboratory, Daresbury, pp 97–102
21. Battye TGG, Kontogiannis L, Johnson O, Powell HR, Leslie AGW (2011) *iMOSFLM*: a new graphical interface for diffraction-image processing with *MOSFLM*. Acta Crystallogr D67:271–281
22. Yamashita K, Kawai Y, Tanaka Y, Hirano N, Kaneko J, Tomita N, Ohta M, Kamio Y, Yao M, Tanaka I (2011) Crystal structure of the octameric pore of staphylococcal γ-hemolysin reveals the β-barrel pore formation mechanism by two components. Proc Natl Acad Sci U S A 108:17314–17319
23. Brunger AT (1997) Patterson correlation searches and refinement. Methods Enzymol 276:558–580
24. Navaza J (1994) AmoRe: an automated package for molecular replacement. Acta Crystallogr A50:157–163
25. Vagin A, Teplyakov A (1997) MOLREP: an automated program for molecular replacement. J Appl Cryst 30:1022–1025
26. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) *Phaser* crystallographic software. J Appl Cryst 40:658–674
27. Bunkoczi G, Read JRJ (2011) Improvement of molecular-replacement models with Sculptor. Acta Crystallogr D67:303–312

28. Zheng A, Yu J, Yamamoto R, Ose T, Tanaka I, Yao M (2014) X-ray structures of eIF5B and eIF5B-eIF1A complex: conformational flexibility of eIF5B restricted on the ribosome by interaction with eIF1A. Acta Crystallogr D70:3090–3098

29. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the web: a case study using the phyre server. Nat Protoc 4:363–371

30. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res V12: W252–W258

31. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93

32. Fujiwara T, Saburi W, Inoue S, Mori H, Matsui H, Tanaka I, Yao M (2013) Crystal structure of Ruminococcus albus cellobiose 2-epimerase: structural insights into epimerization of unmodified sugar. FEBS Lett 587:840–846

33. Sheldrick GM (2008) A short history of SHELX. Acta Crystallogr A64:112–122

34. Schneider TR, Sheldrick GM (2002) Substructure solution with SHELXD. Acta Crystallogr D58:1772–1779

35. Sheldrick GM (2010) Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. Acta Crystallogr D66:479–485

36. Yasutake Y, Watanabe S, Yao M, Takada Y, Fukunaga N, Tanaka I (2003) Crystal structure of the monomeric isocitrate dehydrogenase in the presence of NADP$^+$: insight into the cofactor recognition, catalysis, and evolution. J Biol Chem 278:36897–36904

37. de La Fortelle E, Bricogne G (1997) Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. Methods Enzymol 276:472–494

38. Terwilliger TC, Adams PD, Read RJ, McCoy AJ, Moriarty NW, Grosse-Kunstleve RW, Afonine PV, Zwart PH, Hung LW (2009) Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. Acta Crystallogr D65:582–601

39. Abrahams JP, Leslie AGW (1996) Methods used in the structure determination of bovine mitochondrial F1 ATPase. Acta Crystallogr D52:30–42

40. Morris RJ, Perrakis A, Lamzin VS (2003) ARP/wARP and automatic interpretation of protein electron density maps. Method Enzymol (Carter C, Sweet B (eds)) 374:229–244

41. Sheldrich GM, Schneider TR (1997) SHELXL: high-resolution refinement. Methods Enzymol 277:319–343

42. Tronrud DE, Ten Eyck LF, Matthews BW (1987) An efficient general-purpose least-squares refinement program for macromolecular structures. Acta Crystallogr A43:489–501, http://www.globalphasing.com/buster/

43. Bricogne G, Irwin J (1996) Proceedings of the CCP4 study weekend. In: Dodson E, Moore M, Ralph A, Bailey S (eds) Macromolecular refinement. Daresbury Laboratory, Warrington, pp 85–92

44. Brünger AT, Kuriyan J, Karplus M (1987) Crystallographic *R* factor refinement by molecular dynamics. Science 235:458–460

45. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) *Crystallography & NMR system*: a new software suite for macromolecular structure determination. Acta Crystallogr D54:905–921

46. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D53:240–255

47. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012) Towards automated crystallographic structure refinement with phenix. refine. Acta Crystallogr D68:352–367

48. Engh RA, Huber R (1991) Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr A47:392–400

49. Jack A, Levitt M (1978) Refinement of large structures by simultaneous minimization of energy and *R* factor. Acta Crystallogr A34:931–935

50. Brünger AT (1992) Refinement of large structures by simultaneous minimization of energy and *R* factor. Nature 355:472–475

51. Nakamuraa A, Nemoto T, Heinemann IU, Yamashita K, Sonoda T, Komoda K, Tanaka I, Söll D, Yao M (2013) Structural basis of reverse nucleotide polymerization. Proc Natl Acad Sci U S A 110:20970–20975

52. Liu YC, Nakamura A, Nakazawa Y, Asano N, Ford KA, Hohn MJ, Tanaka I, Yao M, Söll D

(2014) Ancient translation factor is essential for tRNA-dependent cysteine biosynthesis in methanogenic archaea. Proc Natl Acad Sci U S A 111:10520–10525

53. Kostrewa D (1997) Bulk solvent correction: practical application and effects in reciprocal and real space. CCP4 Newsl. Protein Crystallogr 34:9–22

54. Schomaker V, Trueblood KN (1968) On the rigid-body motion of molecules in crystals. Acta Crystallogr B24:63–76

55. Winn M, Isupov M, Murshudov GN (2001) Use of TLS parameters to model anisotropic displacements in macromolecular refinement. Acta Crystallogr D57:122–133

Chapter 17

NMR Structural Biology Using Paramagnetic Lanthanide Probe

Tomohide Saio and Fuyuhiko Inagaki

Abstract

We describe the recent development in nuclear magnetic resonance (NMR) equipped with paramagnetic lanthanide probe. Paramagnetic lanthanide probe provides long-range (~40 Å) distance and angular information that can be exploited in structure determination of large proteins and their complexes, dynamics, ligand screening, and structure-based resonance assignment. Application of the paramagnetic lanthanide probe is not limited to metal-binding proteins but becoming general by the use of lanthanide-binding tags. We here illustrate the practical aspects of the experiments and analyses for the use of paramagnetic lanthanide probe. Applications to protein-protein and protein-ligand structure determination and ligand screening are also shown.

Keywords Paramagnetic lanthanide ion, Pseudocontact shift, Paramagnetic relaxation enhancement, Residual dipolar coupling, Long-range restraint, Nuclear magnetic resonance, Protein structure, Ligand screening, Protein-ligand complex, Lanthanide-binding peptide tag, Lanthanide-chelating tag

1 Introduction

One of the most important features of biomolecular NMR is the structural information in atomic resolution. Nuclei in the molecule placed in the magnetic field provide characteristic resonances to magnetic environment, which is determined by chemical structure and higher-order structure, and dynamics. Thus, NMR can be exploited for three-dimensional structure determination and for the analysis of interaction, protein folding, posttranslational modification, structural change, and dynamics, at atomic resolution [1, 2]. Decades ago, application of biological NMR was limited to molecules less than 20 kDa. Nowadays, however, proteins over several hundreds of kDa [3–5] and membrane proteins [6–10] are within the range of the target, owing to the improvement of the NMR instrument as well as method development including deuterium (^{2}H) labeling [11], transverse relaxation-optimized spectroscopy

Toshiya Senda and Katsumi Maenaka (eds.), *Advanced Methods in Structural Biology*, Springer Protocols Handbooks, DOI 10.1007/978-4-431-56030-2_17, © Springer Japan 2016

(TROSY) [12], and paramagnetic lanthanide probe method that we describe in this chapter.

Paramagnetic lanthanide ions fixed in proteins provide plenty of structural information that can improve both quality and efficiency of the structural analysis by NMR. While nuclear Overhauser effect (NOE), one of the most important methods used in protein structural analysis, gives short-range (~5 Å) distance information, paramagnetic lanthanide probe provides long-range (~40 Å) quantitative distance and angular information. This long-range information is powerful in the structure determination of larger proteins and their complexes. Structural analysis of the large proteins and complexes is generally time-consuming and requires much effort, due to difficulty in collecting a sufficient number of NOE restraints for high-quality structure determination. Especially shortage of intermolecular and/or inter-domain restraints can be a major issue in the analysis, but paramagnetic lanthanide probe provides a solution for this issue. Geometrical information by paramagnetic lanthanide probe can be alternatively used to speed up the analysis [13–15] or can be combined with local information from NOE to obtain more accurate and precise structure [16–20]. Recent development in the computational methods achieved de novo protein structure determination by paramagnetic restraints without any known structure or NOE restraint [21–24]. Structure determination is not an only application of the paramagnetic lanthanide probe. Quantitative long-range information of the paramagnetic lanthanide probe by simple and rapid analysis enables its application to ligand screening [25–27], dynamics analysis [28, 29], characterization of structural changes [30], and structure-based resonance assignment [31, 32]. Despite its fruitful information, paramagnetic lanthanide probe is not yet widely applied, because of special techniques required in sample preparation, NMR measurement, and analysis. Here we describe the practical aspects needed in the application of the paramagnetic lanthanide probe to protein structural analysis by NMR.

1.1 What Kind of Information Does Paramagnetic Lanthanide Probe Provide?

By the use of paramagnetic lanthanide probe, one can obtain various kinds of long-range structural information: distance and angular information from pseudocontact shift (PCS), angular information from residual dipolar coupling (RDC), and distance information from paramagnetic relaxation enhancement (PRE) [33]. These paramagnetic effects can be observed simultaneously once a paramagnetic lanthanide ion is attached to the target protein [34]. Among them the most useful effect is PCS, since PCS provides accurate long-range distance and angular information by simple and quick NMR experiments. PCS is a chemical shift change that depends on the relative location of the observed nucleus to the lanthanide ion (Fig. 1). PCS arises from through-space interactions with the unpaired electrons of the paramagnetic lanthanide ion and

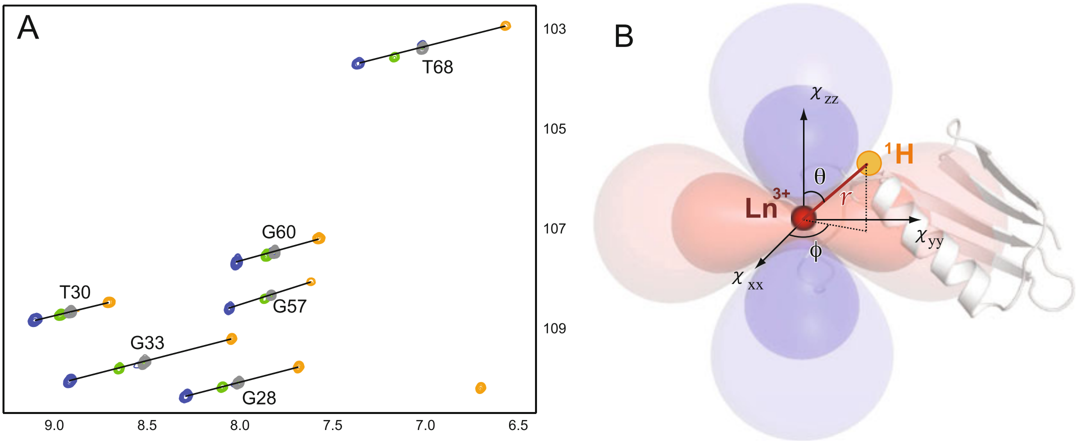


Fig. 1 (**a**) Overlay of ^{1}H-^{15}N-HSQC spectra of ^{15}N-labeled LBT-GB1 (two-point anchored) in complex with La^{3+} (*gray*), Er^{3+} (*green*), Tm^{3+} (*blue*), and Tb^{3+} (*orange*). Chemical shift difference between paramagnetic ion (Er^{3+}, Tm^{3+}, or Tb^{3+}) and diamagnetic lanthanide ion (La^{3+}) is PCS. (**b**) Schematic representation of structural information provided by PCS, drawn on the PCS isosurface depicting the PCSs of ±2.5 and ±0.6 ppm, induced by Tb^{3+} fixed in LBT-GB1 (2rpv.pdb [20]). *Blue* and *red* surfaces indicate the special locations of positive and negative PCSs, respectively

can be observed within the range of 40 Å from the ion [17]. PCS isosurface, which visualizes PCSs as shells of a constant PCS value, shows that the PCS values depend on the spatial location of the observed nuclei relative to the paramagnetic lanthanide ion (Fig. 1b). Thus, PCS values contain distance and angular information of the nuclei, as represented by Eq. (1)

$$\Delta\delta^{PCS} = \frac{1}{12\pi r^3}\left[\Delta\chi_{ax}(3\cos^2\theta - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\theta\cos 2\phi\right] \quad (1)$$

where $\Delta\delta^{PCS}$ is the pseudocontact shift; r, θ, and ϕ are the polar coordinates of the nucleus with respect to the principal axes of the magnetic susceptibility tensor; and $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ are the axial and rhombic components of the magnetic susceptibility tensor (Table 1), as defined by Eq. (2):

$$\Delta\chi_{ax} = \chi_{zz} - \frac{\chi_{xx} + \chi_{yy}}{2}, \text{and } \Delta\chi_{rh} = \chi_{xx} - \chi_{yy} \quad (2)$$

Paramagnetism of lanthanide ions comes from the unpaired electrons in 4f orbital (Table 1). The 4f electrons are located inside of the 5s and 5p electrons and shielded from ligands. Consequently, the 4f electrons take no part in bonding, and thus the chemical properties of the lanthanides ions are similar to each other. On the other hand, magnetic properties of the lanthanide ion vary among the ions [35, 36]. Different lanthanide ions indicate various

Table 1
Electron configurations and ionic radii of the lanthanides

	Electron configuration		Radius (pm)		Paramagnetic effects	
	Atom	Ln^{3+}	Atom	Ln^{3+}	χ^{b}	$\Delta\chi_{\alpha x}$ / $\Delta\chi_{rh}$ [b]
La	[Xe][a] $5d^1$ $6s^2$	[Xe]	187.7	103.2	–	–
Ce	[Xe] $4f^1$ $5d^1$ $6s^2$	[Xe] $4f^1$	182.5	101.0	5.6	2.1/0.7
Pr	[Xe] $4f^3$ $6s^2$	[Xe] $4f^2$	182.8	99.0	11.2	3.4/2.1
Nd	[Xe] $4f^4$ $6s^2$	[Xe] $4f^3$	182.1	98.3	11.4	1.7/0.4
Pm	[Xe] $4f^5$ $6s^2$	[Xe] $4f^4$	181.0	97.0	–	–
Sm	[Xe] $4f^6$ $6s^2$	[Xe] $4f^5$	180.2	95.8	0.6	0.2/−0.1
Eu	[Xe] $4f^7$ $6s^2$	[Xe] $4f^6$	204.2	94.7	~6	−2.3/−1.6
Gd	[Xe] $4f^7$ $5d^1$ $6s^2$	[Xe] $4f^7$	180.2	93.8	55.1	0/0
Tb	[Xe] $4f^9$ $6s^2$	[Xe] $4f^8$	178.2	92.3	82.7	42.1/11.2
Dy	[Xe] $4f^{10}$ $6s^2$	[Xe] $4f^9$	177.3	91.2	99.2	34.7/20.3
Ho	[Xe] $4f^{11}$ $6s^2$	[Xe] $4f^{10}$	176.6	90.1	98.5	18.5/5.8
Er	[Xe] $4f^{12}$ $6s^2$	[Xe] $4f^{11}$	175.7	89.0	80.3	−11.6/−8.6
Tm	[Xe] $4f^{13}$ $6s^2$	[Xe] $4f^{12}$	174.6	88.0	50.0	−21.9/−20.1
Yb	[Xe] $4f^{14}$ $6s^2$	[Xe] $4f^{13}$	194.0	86.8	18.0	−8.3/−5.8
Lu	[Xe] $4f^{14}$ $5d^1$ $6s^2$	[Xe] $4f^{14}$	173.4	86.1	–	–

[a] [Xe] = $1s^2$ $2s^2$ $2p^6$ $3s^2$ $3p^6$ $3d^{10}$ $4s^2$ $4p^6$ $4d^{10}$ $5s^2$ $5p^6$

[b] χ, $\Delta\chi_{ax}$, and $\Delta\chi_{rh}$ values are in 10^{-32} [m^3] [35, 36]

magnitudes and signs of the $\Delta\chi$-tensor due to the difference in the number of 4f electrons (Table 1). The lanthanide ions with smaller magnitudes of the $\Delta\chi$-tensor generate smaller PCS, but at the same time they generate less PRE, which is suitable to obtain the structural information close to the ion. The lanthanide ions with larger magnitudes of $\Delta\chi$-tensor provide stronger PRE as well as PCS, where signals near the ion become too broad to be detected, but PCS can reach to the nuclei far away from the ion. There are also diamagnetic lanthanide ions (La^{3+} and Lu^{3+}) to serve diamagnetic references. This is important because all paramagnetic effects are measured as the difference between the data sets measured in the paramagnetic and diamagnetic states. In contrast to other paramagnetic lanthanide ions having faster electron relaxation time (10^{-12}–10^{-13} s), gadolinium ion (III) has the longest electron relaxation time (10^{-8}–10^{-9} s) and generates strong PRE through dipole-dipole relaxation mechanism while provides no PCS. PRE caused by Gd^{3+} is stronger than those caused by nitroxide spin labels and as strong as those arising from Mn^{2+}, reaching up to

35 Å distance from the ion. A wide variety of paramagnetic effects can be observed by the use of several kinds of lanthanide ions. This is one of the advantages of the lanthanide probe method over the methods using other paramagnetic ions or spin labels.

1.2 How Can Paramagnetic Lanthanide Probe Be Applied to Non-metalloproteins?

For the application of the lanthanide probe, lanthanide ion has to be fixed in a protein frame. Due to lack of an efficient method to attach the ion onto a protein, the paramagnetic lanthanide probe method was limited to metal-binding proteins, by replacing their natural metals such as calcium and magnesium to a lanthanide. The studies on metalloproteins have established a number of useful applications to protein structural/interaction analysis, for example, PCS-based structure refinement [16–19], structure determination of a protein-protein complex [13] and protein-ligand complex [27], conformational and dynamical analysis of multidomain proteins [28, 29], and structure-based NMR signal assignment [31, 34]. Application of the paramagnetic lanthanide probe to non-metalloproteins requires a rigid lanthanide-binding tag, because mobility of the tag reduces the anisotropic paramagnetic effect of the lanthanide ion, losing accuracy and reliability of the structural information. For this purpose, several lanthanide-binding tags have been developed. They are classified into two types: lanthanide-binding peptide tags and lanthanide-chelating reagents. Peptide tags can be attached to a target protein through N- or C-terminal fusion [37–40], through a disulfide bond [41, 42], or through double anchoring via N- or C-terminal fusion and a disulfide bridge [14, 20, 25, 43]. Chelating reagents can be attached through disulfide bond(s) [44–55] or by the introduction of unnatural amino acid of p-azido-L-phenylalanine (AzF) conjugated to the tag via triazole [56]. We here describe the details about two major tags: Caged Lanthanide NMR Probe 5 (CLaNP-5) [54, 55] and two-point anchored lanthanide-binding peptide tag (LBT) [14, 20, 25, 43].

1. CLaNP-5

   In addition to the mobility, another major problem for synthetic tags is peak splitting due to enantiomeric conformer of the lanthanide-substituted tag. Keizers et al. [54, 55] have successfully overcome these issues with CLaNP-5, where two pyridine-N-oxides are introduced to the DOTA-based chelating tag having two arms for disulfide bridges with protein (Fig. 2a). Double linkage of the tag in C-2 symmetric architecture enables strong paramagnetic effect without peak splitting. When two cysteine mutations are properly designed on the protein, CLaNP-5 tag is efficiently ligated to the target protein

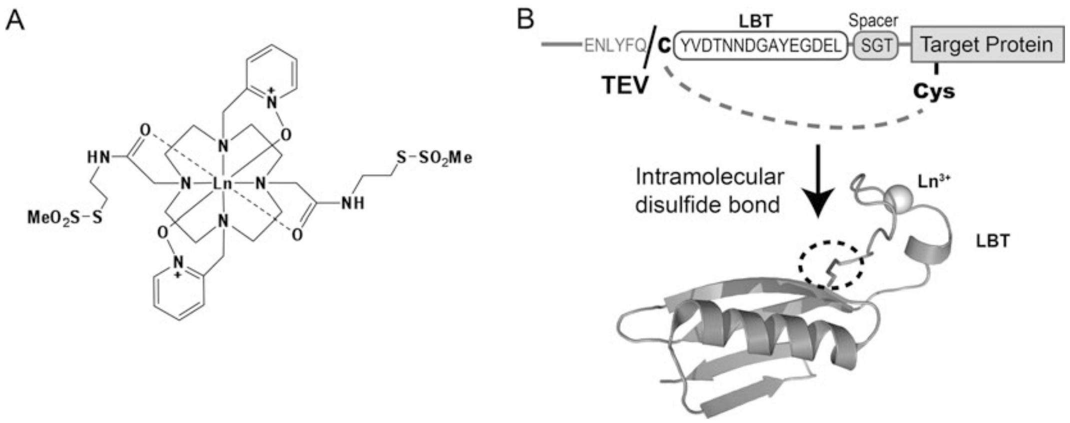


Fig. 2 (**a**) Chemical structure of CLaNP-5 [54, 55]. (**b**) Scheme of the two-point attachment of the LBT [20]. LBT is fused to the target protein with a spacer consisting of three to five amino acids, and second anchoring point is made by disulfide bond between Cys residues at the N-terminus of LBT and on the surface of the protein

by mixing the reagent with the protein solution. Another advantage of CLaNP-5 is flexibility in the position of the tag: the tag can be introduced basically anywhere on the surface of the protein.

2. Two-point anchored LBT

Despite the attractive feature of the synthetic tags, they are not widely used so far, mainly due to the limited availability of the tag. The compound derived from multiple steps of synthetic reactions is not always easy for molecular biologists to prepare by themselves. Saio et al. [20] reported a rigid and easily available lanthanide-binding tag: two-point anchored lanthanide-binding peptide tag where lanthanide-binding peptide (CYVDTNNDGAYEGDEL) derived from the EF-hand motif and optimized for lanthanide binding [41, 42, 57, 58] is attached to the target protein via two anchoring points, a disulfide bridge and an N- or C-terminal fusion (Fig. 2b). The sample preparation is simple and efficient. The LBT sequence is fused to N-/C-terminus of a target protein with a spacer consisting of three to five amino acids [20, 43], and one cysteine residue is introduced to the surface of the protein by mutagenesis. After protein preparation in the reduced condition, the disulfide bridge between the cysteines at the terminus of the tag and on the surface of the protein is efficiently formed by the addition of 5,5′-dithiobis(2-nitrobenzoic acid) (DTNB). Two-point anchoring of LBT suppresses the mobility of the tag, providing strong paramagnetic anisotropic effects that can be used for structural analysis of protein-protein [14, 43] and protein-ligand complexes [25] and resonance assignment. The K_d between lanthanide ion and LBT is ~50 nM [57], which is strong enough to generate significant paramagnetic effects but at the same time is not too strong to exchange the lanthanide ion between experiments.

Paramagnetic lanthanide probe now can be applied to non-metalloproteins by the use of the lanthanide-binding tags including CLaNP-5 and the two-point anchored LBT. In this chapter we describe practical aspects of the experiments and analyses for the use of paramagnetic lanthanide probe with paramagnetic lanthanide tag, especially two-point anchored LBT. Though the process of the attachment of the tag is different depending on the tag, other procedures including measurement of PCS, tensor analysis, and structure calculation are common to all applications.

2 Materials

2.1 Stock Solutions of Lanthanide Ions

1. Lanthanide chloride is dissolved in water or NMR buffer at a concentration of 5 mM.

2.2 Minimal Media for Isotope Labeling

1. M9 salts: 6.8 g Na_2HPO_4, 3.0 g KH_2PO_4, 0.5 g NaCl, 1 g $^{15}NH_4Cl$.
2. 1 M $MgSO_4$.
3. 0.1 M $CaCl_2$.
4. 5 mg/mL thiamin.
5. [$^2H^{13}C$] glucose.
6. [^2H] glucose.
7. [3-Methyl-^{13}C; 3,3-$^2H_2$] α-ketobutyrate (Cambridge Isotope Laboratories, Andover, MA, cat. no. CDLM7318).
8. [3-Methyl-^{13}C; 3,4,4,4-$^2H_4$] α-ketoisovalerate (Cambridge Isotope Laboratories, cat. no. CDLM7317).
9. [Methyl-^{13}C] L-methionine (Cambridge Isotope Laboratories, cat. no. CLM206).
10. [3-^{13}C; 2-^2H] L-alanine (Cambridge Isotope Laboratories, cat. no. CDLM8649).
11. 1 M isopropyl β-D(-)-thiogalactopyranoside (IPTG).

2.3 Software for Tensor Calculation

1. Numbat [59]: http://www.nmr.chem.uu.nl/~christophe/numbat.html
2. Echidna [60]: http://www.nmr.chem.uu.nl/~christophe/echidna.html
3. Olivia: http://fermi.pharm.hokudai.ac.jp/olivia/ (Yokochi et al.)
4. FANTASIAN [61, 62]

2.4 Software for Matrix Calculation

1. MATLAB: MathWorks, Natick, MA
2. FreeMat: http://freemat.sourceforge.net

2.5 Tools for the Preparation of Xplor Input Files

1. HIC-Up server [63]: http://xray.bmc.uu.se/hicup/
2. PRODRG2 server [64]: http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg
3. VEGA ZZ [65]: http://nova.colombo58.unimi.it/cms/index.php?Software_projects:VEGA_ZZ
4. VEGA ZZ server: http://nova.colombo58.unimi.it/vegawe.htm

3 Methods

Here, we describe a procedure to utilize paramagnetic lanthanide probe method with non-metalloproteins, lacking metal-binding site, by the use of two-point anchored LBT. The protocol consists of construct design and optimization, sample preparation, NMR measurement, and analysis. We also describe an application of the paramagnetic lanthanide probe to structure determination of protein-protein complex and to ligand screening and drug design.

3.1 Construct Design for the Attachment of the Lanthanide-Binding Tag

The LBT sequence consisting of 16 amino acids, CYVDTNNDGAYEGDEL, is fused to N- or C-terminus of the target protein where surface-exposed cysteine is also introduced by site-directed mutagenesis (Fig. 2b).

1. Fuse LBT sequence to N- or C-terminus of the target protein (*see* Note 1), with a spacer consisting of 3–5 amino acids by megaprimer method [66] (*see* Note 2). The optimal length of the spacer tends to be 3–4 a.a. when the Cα distance between the terminal residue and cysteine is around 5 Å and 4–5 a.a. when the Cα distance is around 10 Å (Table 2) [43] (*see* Note 3). The amino acid composition of the spacer is arbitrary.
2. Pick one residue whose side chain is exposed but backbone forms a rigid structure, e.g., forms secondary structure, and mutate it to cysteine by site-directed mutagenesis. If the

Table 2
Spacer length between the two-point anchored LBT and target proteins and the distance between the Cα atoms of N-terminus residue of the target and the anchoring residue disulfide bond

	Anchoring point	Cα atom distance (Å)	Minimal spacer length	References
GB1	M1-E19C	6.1	3	[20]
p62 PB1 domain	S3-C26	6.0	3	[14]
FKBP12	V2-T75C	5.6	3	[43]
Grb2 SH2 domain	W60-M73C	9.9	4	[25]

protein originally contains exposed cysteines (*see* Note 4), these need to be mutated into another amino acid.

3.2 Preparation of LBT-Attached Protein

1. The protein containing LBT needs to be purified in reduced conditions (*see* Note 5). Purify the protein in the presence of ~1 mM dithiothreitol (DTT) or 2-mercaptoethanol, except at final gel filtration step. To avoid contamination of metal ions from the medium that may bind to LBT, add EDTA to the sample and/or buffer before final gel filtration.
2. After gel filtration using a buffer lacking reducing reagent or EDTA, dilute the protein to 20 μM or less and incubate the protein with 1 mM DTNB for 2 h at room temperature to form an intramolecular disulfide bond.
3. Dialyze the protein to remove free DTNB, followed by buffer exchange into NMR buffer (*see* Note 6) using concentrator.
4. Formation of intramolecular disulfide bridge as well as the absence of intermolecular disulfide bridge can be verified by SDS-PAGE in nonreducing condition and by NMR (*see* Note 7).
5. Add 1 equivalent of lanthanide ion from 5 mM stock solution to the sample (*see* Note 8).
6. Validate the construct design based on NMR spectra for diamagnetic and paramagnetic state. If the construct is properly designed, all of the diamagnetic resonances except from the region around the anchoring points should match to those of the original protein. Also, you should observe only a single set of paramagnetic resonances without peak splitting or global severe peak broadening.

3.3 Preparation of Stable Isotope-Labeled Protein

Use of the appropriate isotope labeling is inevitable for advanced NMR analysis. Especially in the analysis of paramagnetic lanthanide probe, residue-specific and/or atom-specific labeling is quite useful to reduce spectral complexity and ambiguity. We here describe a standard protocol for (i) uniform ^{2}H/^{15}N/^{13}C-labeling, (ii) residue-specific ^{15}N-labeling or inverse labeling, and (iii) methyl-specific protonation in deuterated background.

(i) *Preparation of uniform ^{2}H/^{15}N/^{13}C-labeled proteins*

For backbone resonance assignment, the sample needs to be labeled with ^{13}C and ^{15}N. Triple labeling (^{2}H/^{15}N/^{13}C) is useful for protein above 20 kDa. The protocol for uniform ^{2}H/^{15}N/^{13}C-labeling is described below, but this can be extended to uniform ^{15}N/^{13}C- or ^{15}N-labeling by replacing $^{2}H_2O$ and ^{2}H/^{13}C-glucose with $^{1}H_2O$ and ^{1}H/^{13}C-glucose, or $^{1}H_2O$ and ^{1}H/^{12}C-glucose, respectively.

1. Inoculate a sterile 50 mL tube containing 5 mL of LB medium in 70 % $^{2}H_2O$ and incubate with shaking at 37 °C for 4–6 h.

2. Transfer 100–200 μL of the media to a sterile 250 mL flask containing 50 mL of ^{2}H/^{13}C/^{15}N-M9 medium and incubate with shaking for ~16 h.
3. Pellet the cells at 3,000 g for 5 min.
4. Resuspend the cells in 1 L of M9 medium containing ^2H_2O and M9 salt supplemented with 1 mM $MgSO_4$, 0.1 mM $CaCl_2$, 5 mg/L of thiamin, 2 g/L of ^{2}H/^{13}C-glucose, and appropriate antibiotic.
5. Incubate with shaking at 37 °C for 4–6 h until OD_{600} = ~0.4.
6. Refrigerate the medium at 15–25 °C 30 min before the induction.
7. Add IPTG at the final concentration of 0.1–1.0 mM, and continue the culture with shaking at 15–25 °C for ~16 h.

(ii) *Preparation of residue-specific ^{15}N-labeled or inversely labeled proteins*

NMR spectra of the protein attached with paramagnetic lanthanide ion show large PCSs (Fig. 1a), and sometimes it is difficult to track all of the shifts. Residue-specific ^{15}N-labeling, where only selected amino acid types give resonances on ^{1}H-^{15}N-HSQC spectra, or inverse labeling, where the resonances from selected types of the amino acids are suppressed on the spectra, simplifies the spectra, thus making it easier to assign the shifted resonances.

1. Inoculate a sterile 50 mL tube containing 5 mL of LB medium and incubate with shaking at 37 °C for 4–6 h.
2. Transfer 100–200 μL of the media to a sterile 250 mL flask containing 50 mL of M9 medium and incubate with shaking for ~16 h.
3. Pellet the cells at 3,000 g for 5 min.
4. Resuspend the cells in 1L of M9 medium containing M9 salt supplemented with 1 mM $MgSO_4$, 0.1 mM $CaCl_2$, 5 mg/L of thiamin, 2 g/L of glucose, and appropriate antibiotic. For residue-specific ^{15}N-labeling, use 1 g/L of $^{14}NH_4Cl$, 50 mg/L of ^{15}N-labeled amino acid(s), and 500 mg/L of each unlabeled amino acid. Unlabeled amino acids are added from the beginning of the culture, and labeled amino acid(s) are added 1 h before protein induction. For inversely labeled protein, use 1 g/L of $^{15}NH_4Cl$ and 500 mg/L of unlabeled amino acid(s). Unlabeled amino acids are added 1 h before protein induction.
5. Incubate with shaking at 37 °C for 4–6 h until OD_{600} = ~0.4.
6. Refrigerate the medium at 15–25 °C 30 min before the induction.

7. Add IPTG at the final concentration of 0.1–1.0 mM, and continue the culture with shaking at 15–25 °C for ~12 h (*see* Note 9).

(iii) *Preparation of deuterated proteins with methyl-specific protonation*

1. Inoculate a sterile 50 mL tube containing 5 mL of LB medium in 70 % ^2H_2O and incubate with shaking at 37 °C for 4–6 h.
2. Transfer 100–200 μL of the media to a sterile 250 mL flask containing 50 mL of $^2H^{15}N$-M9 medium and incubate with shaking for ~16 h.
3. Pellet the cells at 3,000 g for 5 min.
4. Resuspend the cells in 1 L of M9 medium containing M9 salt supplemented with 1 mM $MgSO_4$, 0.1 mM $CaCl_2$, 5 mg/L of thiamin, 2 g/L of ^2H-glucose, and appropriate antibiotic.
5. Incubate with shaking at 37 °C until OD_{600} = ~0.4.
6. Add 50 mg/L of α-ketobutyrate (3-methyl-^{13}C, 3, 3-$^2H_2$), 85 mg/L of α-ketoisovalerate (3-methyl-^{13}C, 3,4,4,4-$^2H_4$), and 50 mg of [methyl-^{13}C] L-methionine 1 h before the induction. Add 50 mg/L of [2-^2H, 3-^{13}C] L-alanine 30 min before the induction.
7. Refrigerate the medium at 15–25 °C 30 min before the induction.
8. Add IPTG at the final concentration of 0.1–1.0 mM, and continue the culture with shaking at 15–25 °C for ~16 h.

3.4 Measurement of Paramagnetic Effects

We here describe how to measure PCSs that give the most useful information in protein structural analysis. All of the paramagnetic effects are measured by the comparison between paramagnetic and diamagnetic state. As discussed in introduction, two ions in lanthanide group, La^{3+} and Lu^{3+}, are diamagnetic thus are used as a reference. PCS is chemical shift change induced via through-space interaction between observed nuclei and electrons in lanthanide ion. The most popular way to measure PCS is to use two-dimensional NMR spectra (Fig. 1a):

1. Prepare NMR samples of LBT-attached protein, containing 1 equivalent of a lanthanide ion. At least two samples are required: one containing paramagnetic lanthanide ion and one containing diamagnetic lanthanide ion as a reference. It is useful to use multiple paramagnetic lanthanide ions, including ones having weaker as well as stronger paramagnetic effect (Fig. 1a) (Table 1) (*see* Note 10).
2. Acquire 2D NMR spectra for each sample. ^1H-^{15}N-HSQC is often used for the observation of backbone PCSs.
3. Assign shifted resonances based on the resonance assignment made for diamagnetic protein. On the overlaid spectra, the

shifted resonances from the same nucleus align in a straight line (Fig. 1a) (*see* Note 10).

4. Subtract the chemical shift of diamagnetic resonance from that of paramagnetic resonance, to obtain PCS. In the case of backbone HSQC spectra, PCS values both for $^1H^N$ and ^{15}N are obtained.

3.5 Determination of $\Delta\chi$-Tensor

$\Delta\chi$-tensor is anisotropic component of magnetic susceptibility tensor and is responsible for the characterization of anisotropic paramagnetic effects including PCS and RDC. Determination of $\Delta\chi$-tensor is essential for the quantitative use of the anisotropic paramagnetic effects. Given the availability of three-dimensional structure of the protein, $\Delta\chi$-tensor can be determined in principle based only on eight PCS values where the parameters $\Delta\chi_{ax}$, $\Delta\chi_{rh}$, Euler angles (α, β, γ), and metal position (x, y, z) are determined. A larger number of PCSs enable more reliable analysis, but PCSs from the flexible region of the protein (e.g., loop or terminal region) can disturb the fitting. PCSs should be collected from the resonances from the rigid region of the protein. PCS-based tensor fitting is supported by several programs such as Numbat [59], Echidna [60], and Olivia (Yokochi et al. http://fermi.pharm.hokudai.ac.jp/olivia/). Although the details of the operation differ from program to program, the basic procedure is common as described below:

1. Prepare a table of PCSs and 3D coordinates of the protein (.pdb). The coordinate should contain hydrogen if the PCS table contains proton PCSs.
2. Perform tensor fitting. PCSs from flexible region should be avoided. The fitting often depends on starting parameters, thus put approximate values for parameters such as $\Delta\chi_{ax}$, $\Delta\chi_{rh}$, and metal position (x, y, z) (*see* Note 11). Tensor fitting based on the data from multiple lanthanide ions makes the result more reliable since one can reduce variables assuming that all of the lanthanide ions attached to the same tag have shared metal position (x, y, z).
3. Once the reasonable tensor parameters are obtained, back calculate PCS values. Check the correlation between calculated and observed values (Fig. 3b) to see if there is any misassignment.
4. The PCSs from crowded regions can be additionally assigned with the reference of calculated PCS.
5. Fit the tensor based on the updated PCS table. Although γ term in Euler angle tends to vary by lanthanide ion, the lanthanide ions in the same tag generally have similar α and β that represent the angle of χ_{zz} axis (Fig. 3a) (Table 3) (*see* Notes 12 and 13).

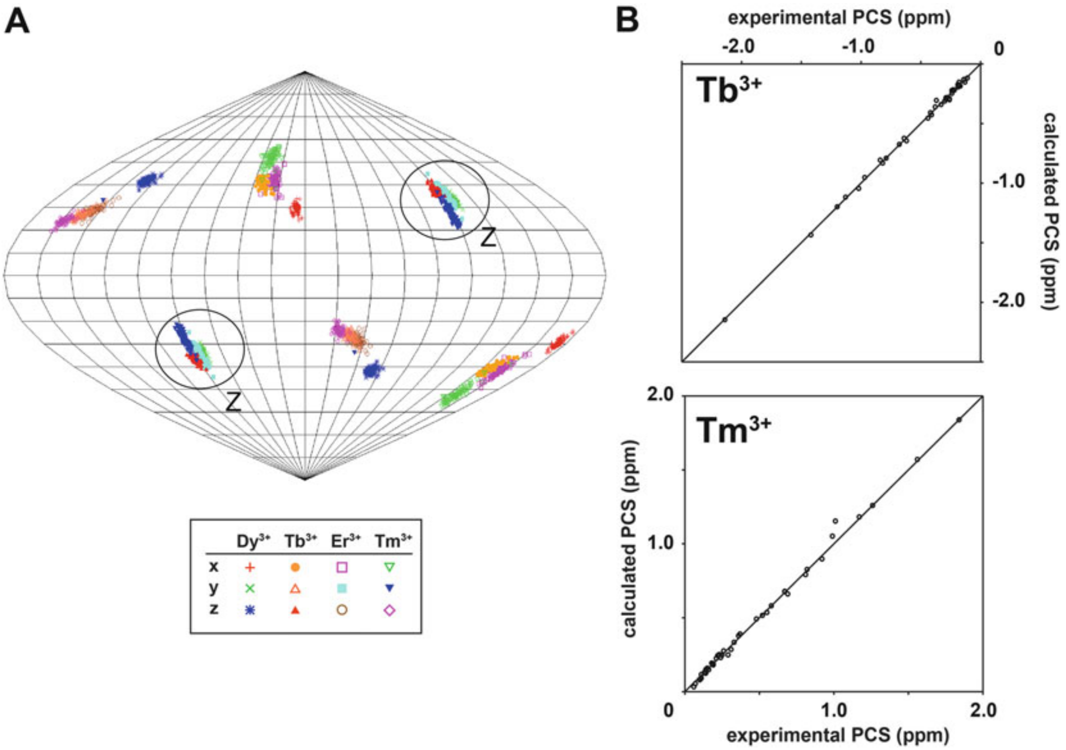


Fig. 3 $\Delta\chi$-tensor determination for LBT-Grb2 [25]. (**a**) Orientation of the principal axes of the $\Delta\chi$-tensors of Dy^{3+}, Tb^{3+}, Er^{3+}, and Tm^{3+} in complex with LBT-Grb2, visualized in Sanson-Flamsteed projection. The plots show the points where the principal axes of the $\Delta\chi$-tensor penetrate the sphere. One hundred sets of plots represent the result of Monte Carlo analysis using the 100 partial PCS data sets in which 30 % of the input data were randomly deleted. (**b**) Comparison between experimental and back-calculated PCS of backbone amide protons observed in LBT-Grb2 in the presence of Tb^{3+} and Tm^{3+}

Table 3
$\Delta\chi$-tensor parameters for lanthanide ions in complex with LBT-Grb2

	Dy^{3+}	Tb^{3+}	Er^{3+}	Tm^{3+}
$\Delta\chi_{ax}$[a]	22.7 ± 1.3	29.2 ± 1.7	−7.7 ± 0.7	−17.5 ± 1.6
$\Delta\chi_{rh}$[a]	17.6 ± 0.7	16.9 ± 0.5	−7.3 ± 0.2	−17.1 ± 0.5
α[b]	106	97	104	99
β[b]	57	52	57	65
γ[b]	53	34	36	27

[a] $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ values are in 10^{-32} [m^3], and error estimates were obtained by Monte Carlo protocol using 100 partial PCS data sets in which 30 % of the input data were randomly deleted

[b] Euler angle rotations in ZYZ convention (degrees)

3.6 Use of the Paramagnetic Information in Structural Analysis

Most of the major structure calculation software can now handle paramagnetic restraints, such as PCS, RDC, and PRE. Patches for paramagnetic restraints in structure calculation have been developed in Bertini's group from the early stage: PARArestraints for Xplor-NIH [67] and paramagnetic DYANA/CYANA [68]. These patches enable the use of paramagnetic restraints along with standard restraints such as NOE distance restraints and dihedral angle restraints. Recently more and more software have been upgraded so that they can incorporate paramagnetic restraints in the calculation. For example, paramagnetic restraints are implemented in CYANA3.0 [69], HADDOCK [70], and PCS-ROSETTA [23]. Here, we will describe, as an example, the details of the structure calculation of protein-protein complex by rigid-body minimization by Xplor-NIH [71] equipped with PARArestraints for Xplor-NIH [67] (http://www.cerm.unifi.it/softwares/para-restraints-for-xplor-nih) that has been frequently used in the structure determination of metalloproteins [36, 72, 73] and proteins attached with lanthanide-binding tags [14, 26, 40, 43, 74]. Examples of the Xplor script are available in the previous reports [14, 26, 43]:

1. Attach two-point anchored LBT to one of the proteins in the complex as described in the Sects. 3.1 and 3.2.
2. Observe PCSs for both of the proteins in the complex as described in the Sect. 3.4.
3. Determine the tensor parameters including the metal position based on the PCSs observed for the protein containing LBT (*see* Note 14), as described in the Sect. 3.5.
4. Set up pseudo-atoms representing tensor axes and an atom for the paramagnetic lanthanide ion. Use of the multiple sets of PCS data from different paramagnetic lanthanide ions requires multiple sets of tensor axes. All of the origins of the tensor axes should match to the position of the lanthanide ion which is determined in the tensor fitting.
5. Randomize the relative orientation of the proteins; the coordinates of the protein having paramagnetic lanthanide are held fixed, while the binding partner is treated as rigid body so that the protein can be freely rotated and translated.
6. Starting from randomized position, dock the proteins using PCSs observed for both of the proteins in the complex (Fig. 4). Due to the symmetric nature of the $\Delta\chi$-tensor (Fig. 1b), the result may have degenerated solutions where the protein is located at multiple positions that equally satisfy PCS restraints (Fig. 5) [14, 43]. Since only one of the solutions generally has physical contact between the proteins, the degeneracy can be overcome by the use of contact-surface restraint derived from chemical shift perturbation mapping (Fig. 4a) [14] or a couple of intermolecular NOEs if available (*see* Note 15). This

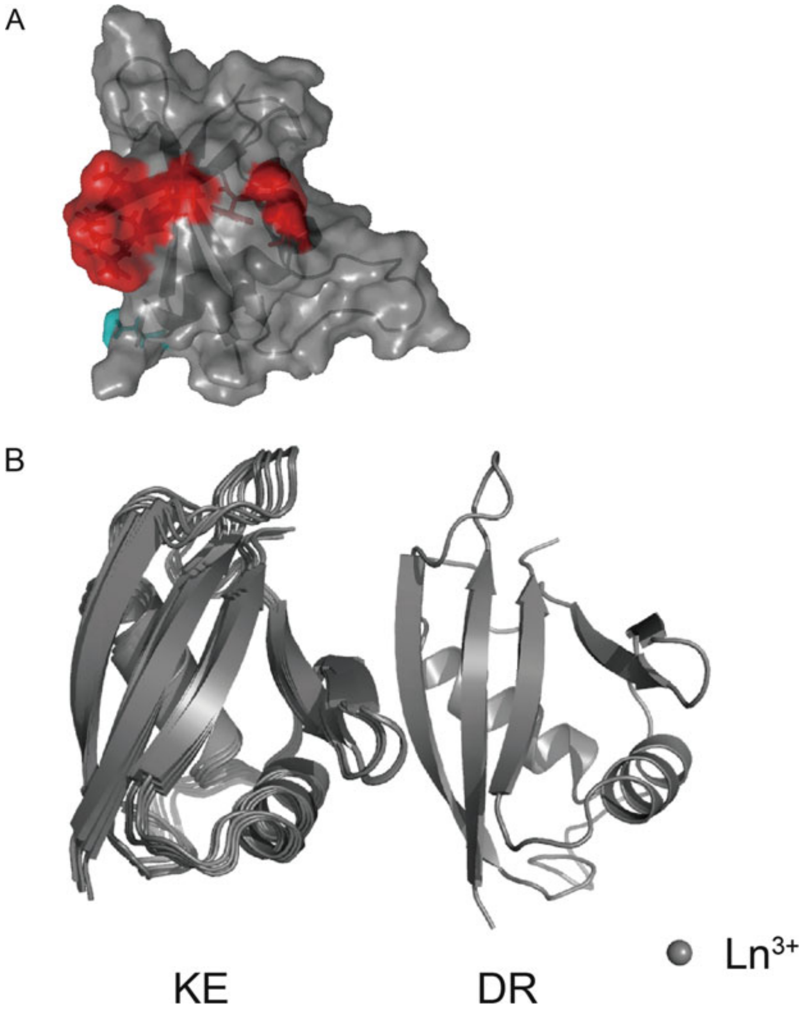


Fig. 4 PCS-driven rigid-body docking of the p62 PB1 D67A/D69R (DR) mutant and K7E/R94A (KE) mutant [14]. (**a**) Chemical shift perturbation of backbone amide groups of KE upon the complex formation with LBT-DR at a ratio of 1:1. The clustered residues indicating large chemical shift perturbation are colored in *red*. The residues colored in *red* are defined as "contact residues" in the calculation. (**b**) The docking structure of the DR/KE complex, calculated based on PCSs derived from Tm^{3+} and Tb^{3+} as well as contact-surface restraints. The ten lowest-energy structures are superimposed. The average backbone rmsd was 0.31 Å. The metal position is represented as a sphere

degeneracy also can be overcome by the use of two or more sets of PCS data with lanthanide-binding tags introduced at different positions or simply by the use of data sets from two-point anchored LBT with different spacer lengths [43] (*see* Note 16).

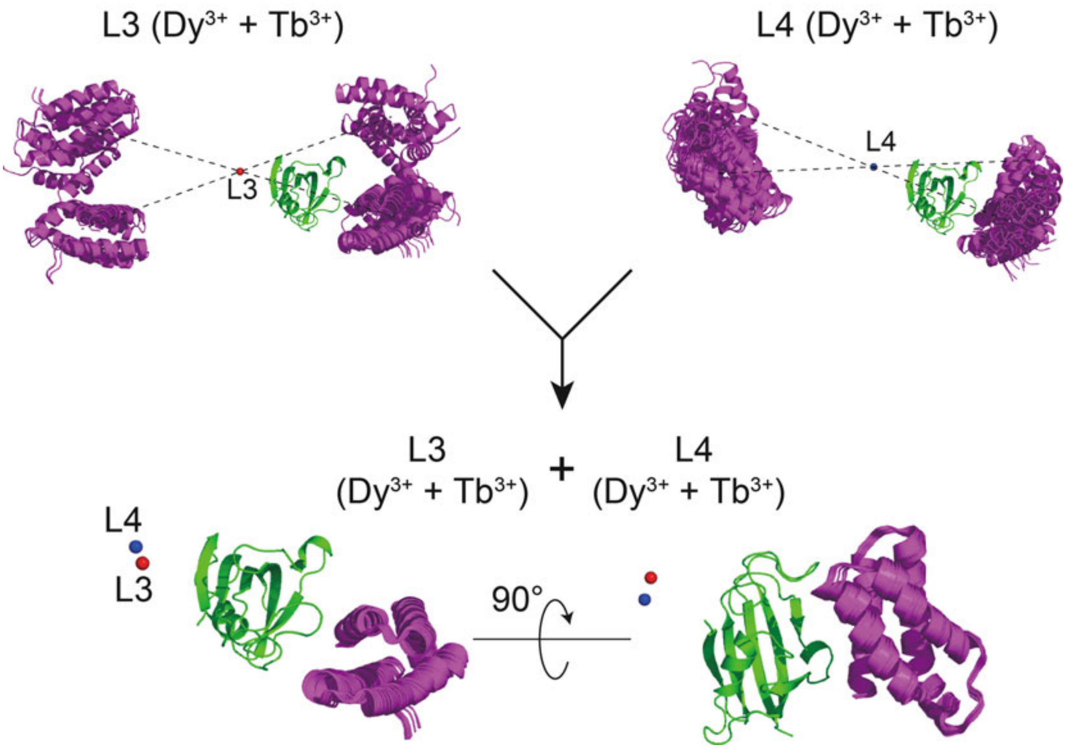


Fig. 5 The PCS-based docking between the FKBP12-rapamycin and FRB domains [43]. The degenerated solutions due to the symmetry of $\Delta\chi$-tensor were resolved by the use of the two sets of PCSs from the two protein samples having different spacer lengths. The calculation based on PCSs from LBT-FKBP12 with three residues (L3) or four residues (L4) as a spacer between LBT and FKBP12 resulted in the four degenerated solutions due to the symmetric nature of $\Delta\chi$-tensor. The degeneracy was resolved by the use of the two sets of PCS data from LBT-FKBP12 with three and four spacer residues. These structures have an average backbone rmsd of 0.2 Å

3.7 Application of the Paramagnetic Lanthanide Probe to Ligand Screening and Drug Design

Lanthanide-induced long-range paramagnetic effects are also useful in ligand screening and drug design, especially in fragment-based drug design (FBDD) where small simple compounds (fragments) are screened for binding to a target protein, and the hit compounds are then optimized to increase their affinity. For efficient FBDD, it is inevitable to obtain structural information on the ligand-protein complex, even for weakly bound ligands. Despite its reliability in the detection and evaluation of the binding, NMR especially with NOE-based conventional approach requires much effort and time to determine the structure of protein-ligand complex. Here paramagnetic lanthanide probe can make it shorter and simpler. Once the paramagnetic center is introduced to the target protein, one can exploit both of PRE and PCS by the use of appropriate lanthanide ions. Saio et al. proposed a hybrid method that screens ligands bound to the target protein by gadolinium (III)-induced PRE, followed by the rapid structure determination of protein-ligand complex based on PCS [25]:

1. *Screening*: Ligands bound to the target protein can be identified from compound mixture based on Gd^{3+}- induced PRE.
   (a) Load Gd^{3+} to LBT attached to the target protein, by the addition of 1 eq. of Gd^{3+} (*see* Note 8). The anchoring points for LBT should be designed so that the lanthanide ion is located close (<25 Å) to the ligand-binding sites. The sample should be prepared in ^2H_2O solution.
   (b) Add ~0.1 eq of the protein containing Gd^{3+} into the mixture of ~10 compounds dissolved in ^2H_2O, and acquire ^1H spinlock 1D NMR spectra [75, 76] with spinlock period of 10 or 200 ms (Fig. 6a). The compound bound to the protein is identified by the signal reduction due to Gd^{3+}-induced PRE.
2. *Structural analysis*: The ligands identified in the screening step are further analyzed, where the structure of the ligand-protein complex can be rapidly determined based on PCSs [25–27, 40]. PCS restraints can be collected by replacing the Gd^{3+} ion with other paramagnetic lanthanide ions, such as Tb^{3+}, Tm^{3+}, and Dy^{3+}, that have anisotropic magnetic susceptibility tensors. Once $\Delta\chi$-tensor parameters are determined for each lanthanide ion based on the PCSs observed from the protein, PCSs from the ligand can be readily translated into quantitative structural information on the complex.
   (a) Prepare two or more protein samples: one containing Lu^{3+} as a diamagnetic reference and the others containing Tm^{3+} or other paramagnetic lanthanide ions having anisotropic magnetic susceptibility tensor. The sample should be prepared in ^2H_2O solution.
   (b) Titrate the protein into the compound dissolved in ^2H_2O step by step and measure ^1H 1D NMR spectra (Fig. 6b).
   (c) In the case of the lower-affinity ligands, which are the major targets in FBDD screening, the observed chemical shift changes are the weighted averages of the free and bound states because of the fast exchange process. The chemical shift differences between the free and bound states ($\Delta\delta^{\text{bound}}{}_{\text{ppm}}$) of the ligand were calculated from the curve fitting to observed chemical shift change ($\Delta\delta_{\text{ppm}}$), using Eq. 3:

$$\Delta\delta_{\text{ppm}} = \Delta\delta_{\text{ppm}}^{\text{bound}} \frac{[L]+[P]+K_{\text{d}} - \sqrt{([L]+[P]+K_{\text{d}})^2 - 4[L][P]}}{2[L]}, \quad (3)$$

where $[L]$ and $[P]$ are the concentrations of the ligand and protein, respectively, and K_{d} is a dissociation constant (*see* Note 17). PCS of the bound state, PCS^{bound}, is calculated by Eq. (4)

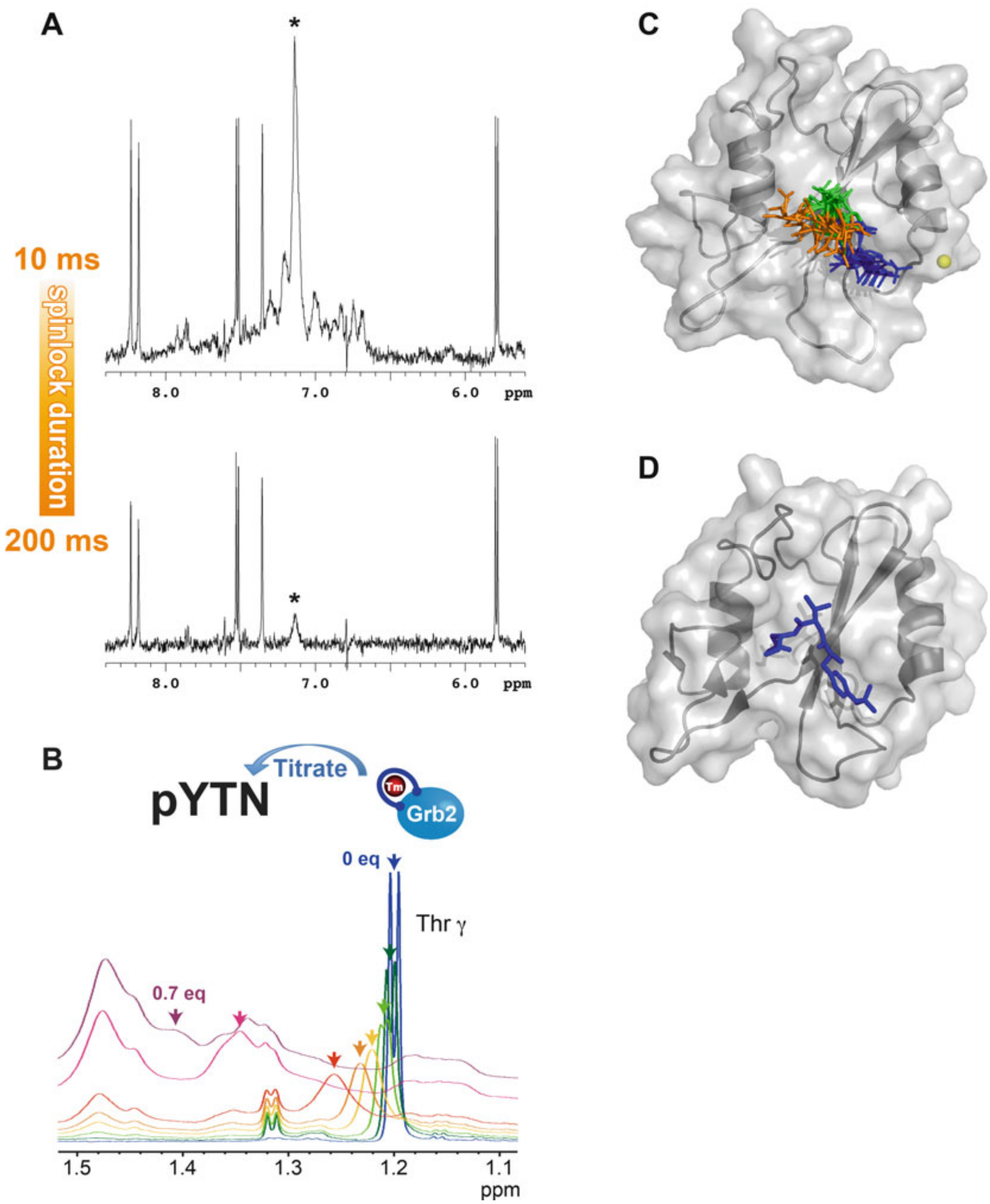


Fig. 6 Application of paramagnetic lanthanide probe to ligand screening [25]. (**a**) ^{1}H spinlock 1D NMR spectra of the pYTN tripeptide in the presence of the other six compounds. The spectra were acquired with a spinlock period of 10 ms or 200 ms in the presence of 0.2 eq. of LBT-Grb2 SH2 with Gd^{3+}. The resonance from pYTN peptide that binds to Grb2 SH2 is indicated by asterisk. (**b**) Selected region of the ^{1}H NMR spectra of the pYTN tripeptide, acquired during titration of LBT-Grb2 containing Tm^{3+}. (**c**) The ten lowest-energy structures of pYTN tripeptide in complex with Grb2 SH2 determined by PCSs. The *yellow* sphere represents the position of the lanthanide ion. Phosphorylated Tyr, Thr, and Asn in the tripeptide are colored in *blue*, *green*, and *orange*, respectively. Despite the moderate convergence of the peptide, the binding surface and the orientation of the peptide well correspond to the X-ray crystal structure (**d**) of Grb2 SH2 in complex with a phosphorylated peptide (PSpYVNVQN) (1jyr.pdb [79])

$$PCS^{\mathrm{bound}} = \Delta\delta_{\mathrm{ppm}}^{\mathrm{bound}}(\mathrm{para}) - \Delta\delta_{\mathrm{ppm}}^{\mathrm{bound}}(\mathrm{dia}), \tag{4}$$

where $\Delta\delta^{\mathrm{bound}}{}_{\mathrm{ppm}}(\mathrm{para})$ and $\Delta\delta^{\mathrm{bound}}{}_{\mathrm{ppm}}(\mathrm{dia})$ are the chemical shift differences of the ligand upon the binding to the proteins loaded with a paramagnetic and diamagnetic lanthanide ion, respectively.

(d) The ligand can be docked onto the protein based on the calculated PCSs, following the procedure described in the Sect. 3.6 (Fig. 6c) (*see* Notes 18 and 19).

4 Notes

1. In the case of N-terminal fusion of LBT, expression tag or affinity tag can be cleaved without any artificial amino acid left by TEV protease where the C-terminal Gly in the recognition sequence (ENLYFQ/G) is substituted by Cys (Fig. 2b).
2. The LBT sequence can be easily cloned into the plasmid coding the target protein, without any restriction enzyme site, by the use of overlap extension PCR using megaprimers [66]. This method consists of two PCR steps. In the first PCR, LBT sequence is amplified by chimeric primers that have 3′ end complementary to LBT sequence and 5' end complementary to the plasmid of the protein. The PCR product is purified by gel recovery and used as primers in the second PCR, where the entire plasmid is amplified. The product of the second PCR contains LBT sequence inserted into the plasmid of the protein. The parental plasmid is digested by DpnI, followed by transformation to *E. coli*.
3. The length of the spacer as well as the position of the Cys mutation can be optimized by NMR spectra in the presence of paramagnetic lanthanide ion such as Tm^{3+} and Yb^{3+}. The optimal constructs give significant PCS without peak doubling. Peak doubling is an indication of shortage of the spacer. Reduced PCS with broad signals means undefined paramagnetic center, indicating too much spacer. The constructs also can be evaluated based on melting temperature that is acquired by differential scanning fluorimetry (DSF), differential scanning calorimetry (DSC), or circular dichroism (CD) spectroscopy [43].
4. Exposed cysteine can be detected by Ellman's reagent (DTNB). In the presence of exposed thiol group, the addition of DTNB results in the dissociation of yellow-colored NTB^{2-} ion that is quantified with extinction coefficient of 14,150 $M^{-1}cm^{-1}$ at 412 nm wavelength.
5. Introduction of LBT and Cys mutation may affect the solubility of the protein. However, the protein expressed in insoluble fraction can be easily recovered by high-pressure refolding

where 250 MPa of the hydrostatic pressure is applied to the pellet resuspended in buffer containing reducing reagent for ~16 h [77]. Unlike traditional refolding method using chaotropic reagents such as urea or guanidine, hydrostatic pressure refolding doesn't fully unfold the protein.

6. Avoid phosphate for NMR buffer, since lanthanide ion binds to phosphate group. MES and Tris buffers are often used.
7. Formation of intramolecular disulfide bridge is confirmed by NMR in the presence of paramagnetic lanthanide ion. If intramolecular bridge is properly formed, the significance of overall PCSs should be reduced by the addition of ~5 mM DTT, since the release of disulfide bridge results in higher mobility of the lanthanide ion that averages out anisotropic paramagnetic effect.
8. Addition of concentrated solution of lanthanide ion sometimes induces protein precipitation. Stock solution at lower (5–10 mM) concentration is preferable.
9. Longer incubation after addition of IPTG may cause scrambling. Shorter incubation time (e.g., 12 h at 18 °C or 6 h at 25 °C) is recommended.
10. Since the ^{1}H and ^{15}N atoms of each amide group are close in space, the PCS has similar ppm values in both ^{1}H and ^{15}N dimensions. This linearity helps to track large PCSs induced by strong paramagnetic lanthanide ions based on smaller PCSs induced by weaker paramagnetic lanthanide ions (Fig. 1a).
11. The paramagnetic lanthanide ion ligated in the same coordination should give similar tensor parameters, given the lanthanide-binding tag is well fixed in the protein frame. In the case of LBT, the initial parameters for $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ can be taken from previous reports [14, 20, 25, 35, 41–43]. Initial metal position can be set at one of the side-chain atoms of the residue mutated to Cys for LBT attachment.
12. Most of the tensor fitting program defines the axes according to $|\chi_{zz}| > |\chi_{yy}| > |\chi_{xx}|$, which sometimes results in swapped axes of χ_{zz} and χ_{yy}, especially for the lanthanide ions with higher rhombicity such as Tm^{3+} and Yb^{3+}. The swapped axes cause apparently very different Euler angles compared to other lanthanide ions, even though the actual tensor axes are similar to each other. The $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ can be recalculated using Eq. (1), by exchanging χ_{zz} and χ_{yy}. Euler angles also can be updated by matrix calculation using MATLAB or FreeMat, where the Euler angles are translated into matrix and then the frame is rotated so that χ_{zz} axis is aligned with χ_{yy}. To avoid complexity in Euler anger representation, the axis orientations can also be visualized in Sanson-Flamsteed projection, in which

points where the principal axes of the $\Delta\chi$-tensor penetrate the sphere are plotted (Fig. 3a).

13. Orientation of the principal axes of $\Delta\chi$-tensor depends on the atomic coordinates of the protein; a different pdb file of the same protein gives different angles.

14. Xplor equipped with PARArestraints for Xplor-NIH uses van Vleck units (vvu; $\text{m}^3/3.77\ 10^{-35}$) for $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$.

15. In the use of chemical shift perturbation for contact-surface restraints, interfacial residues are selected according to the three criteria [14, 78]: (A) significant chemical shift perturbation is observed upon complex formation, (B) at least one or two atoms of the residue are exposed on the surface of the protein, and (C) the selected residue is involved in a cluster of residues on a contiguous, single binding surface. The contact-surface restraints are set up as distance restraints between the atoms of the selected residues of the protein and all atoms of the binding partner using the r^{-6} averaging option [78]. For the r^{-6} averaging option, the distance between selected sets of atoms is averaged according to the equation:

$$d = \left(\sum_{ij} r_{ij}^{-6}\right)^{-1/6} \tag{5}$$

where r_{ij} represents the distance between the atom i in the selected residues of the protein and atom j in all residues in the binding partner. Averaging the minus 6th power of the distance emphasizes the smaller distance values; thus, a restraint is satisfied when at least one pair of the atoms locates close to each other:

16. The degeneracy also can be resolved by the use of multiple PCS data set with different metal positions and different directions of the principal axes of $\Delta\chi$-tensor, which can be obtained from two-point anchored LBT with two different spacer lengths (Figs. 2b and 5) [43]. Two-point anchored LBT allows at least two sets of the spacer lengths: "minimum" and "minimum plus one" (Table 2), resulting in two different sets of PCS that suppress the degenerated solutions (Fig. 5c).

17. In order to obtain better fitting, K_d may need to be fixed. K_d can be obtained by another experiment such as NMR titration where the compound is titrated into the labeled protein, isothermal titration calorimetry, fluorescence, and surface plasmon resonance.

18. Xplor-NIH calculation handling organic compound requires several input files for the compound, such as topology file, parameter file, and PSF file. Topology file (*.top) and parameter file (*.par) can be generated from .pdb file of the compound

on HIC-Up server. .pdb file can be generated from chemical structure of the compound on PRODRG server. PSF file can be generated by VEGA ZZ software or VEGA ZZ server.

19. Although the structure determined based on PCSs is less converged compared to ones determined by standard method using NOEs (Fig. 6c, d), the structure has enough resolution to provide the binding site on the protein and orientation of the ligand. The quality can be improved by the use of multiple sets of PCSs obtained from the lanthanide-binding tags at different positions [26].

Acknowledgments

We would like to thank Dr. Yoshihiro Kobashigawa for providing the data for Fig. 5. A part of this work was supported by the Creation of Innovation Centers for Advanced Interdisciplinary Research Areas Program, Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Kanelis V, Forman Kay JD, Kay LE (2001) Multidimensional NMR methods for protein structure determination. IUBMB Life 52:291–302. doi:10.1080/152165401317291147
2. Pellecchia M, Sem DS, Wüthrich K (2002) NMR in drug discovery. Nat Rev Drug Discov 1:211–219. doi:10.1038/nrd748
3. Gelis I, Bonvin AMJJ, Keramisanou D et al (2007) Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. Cell 131:756–769. doi:10.1016/j.cell.2007.09.039
4. Saio T, Guan X, Rossi P et al (2014) Structural basis for protein antiaggregation activity of the trigger factor chaperone. Science 344:1250494–1250494. doi:10.1126/science.1250494
5. Sprangers R, Kay LE (2007) Quantitative dynamics and binding studies of the 20S proteasome by NMR. Nature 445:618–622. doi:10.1038/nature05512
6. Hiller S, Garces RG, Malia TJ et al (2008) Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. Science 321:1206–1210. doi:10.1126/science.1161302
7. Bokoch MP, Zou Y, Rasmussen SGF et al (2010) Ligand-specific regulation of the extracellular surface of a G-protein-coupled receptor. Nature 463:108–112. doi:10.1038/nature08650
8. Nygaard R, Zou Y, Dror RO et al (2013) The dynamic process of β2-adrenergic receptor activation. Cell 152:532–542. doi:10.1016/j.cell.2013.01.008
9. Kofuku Y, Ueda T, Okude J et al (2012) Efficacy of the β2-adrenergic receptor is determined by conformational equilibrium in the transmembrane region. Nat Commun 3:1045. doi:10.1038/ncomms2046
10. Kofuku Y, Ueda T, Okude J et al (2014) Functional dynamics of deuterated β2-adrenergic receptor in lipid bilayers revealed by NMR spectroscopy. Angew Chem Int Ed 53:13376–13379. doi:10.1002/anie.201406603
11. Tugarinov V, Kanelis V, Kay LE (2006) Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. Nat Protoc 1:749–754. doi:10.1038/nprot.2006.101
12. Pervushin K, Riek R, Wider G, Wüthrich K (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci U S A 94:12366–12371

13. Pintacuda G, Park AY, Keniry MA et al (2006) Lanthanide labeling offers fast NMR approach to 3D structure determinations of protein – protein complexes. J Am Chem Soc 128:3696–3702. doi:10.1021/ja057008z
14. Saio T, Yokochi M, Kumeta H, Inagaki F (2010) PCS-based structure determination of protein–protein complexes. J Biomol NMR 46:271–280. doi:10.1007/s10858-010-9401-4
15. Keizers PHJ, Mersinli B, Reinle W et al (2010) A solution model of the complex formed by adrenodoxin and adrenodoxin reductase determined by paramagnetic NMR spectroscopy. Biochemistry 49:6846–6855. doi:10.1021/bi100598f
16. Detlef B, Ivano B, Cremonini MA et al (1997) Solution structure of the paramagnetic complex of the N-terminal domain of calmodulin with two Ce^{3+} ions by 1H NMR†,‡. Biochemistry 6(39):11605–11618. doi:10.1021/bi971022
17. Allegrozzi M, Bertini I, Janik MBL et al (2000) Lanthanide-induced pseudocontact shifts for solution structure refinements of macromolecules in shells up to 40 Å from the metal ion. J Am Chem Soc 122:4154–4161. doi:10.1021/ja993691b
18. Bertini I, Janik MBL, Liu G et al (2001) Solution structure calculations through self-orientation in a magnetic field of a cerium(III) substituted calcium-binding protein. J Magn Reson 148:23–30. doi:10.1006/jmre.2000.2218
19. Bertini I, Donaire A, Jiménez B et al (2001) Paramagnetism-based versus classical constraints: an analysis of the solution structure of Ca Ln calbindin D9k. J Biomol NMR 21:85–98. doi:10.1023/A:1012422402545
20. Saio T, Ogura K, Yokochi M et al (2009) Two-point anchoring of a lanthanide-binding peptide to a target protein enhances the paramagnetic anisotropic effect. J Biomol NMR 44:157–166. doi:10.1007/s10858-009-9325-z
21. Yagi H, Pilla KB, Maleckis A et al (2013) Three-dimensional protein fold determination from backbone amide pseudocontact shifts generated by lanthanide tags at multiple sites. Structure 21:883–890. doi:10.1016/j.str.2013.04.001
22. Li J, Pilla KB, Li Q et al (2013) Magic angle spinning NMR structure determination of proteins from pseudocontact shifts. J Am Chem Soc 135:8294–8303. doi:10.1021/ja4021149
23. Schmitz C, Vernon R, Otting G et al (2012) Protein structure determination from pseudocontact shifts using ROSETTA. J Mol Biol 416:668–677. doi:10.1016/j.jmb.2011.12.056
24. Bhaumik A, Luchinat C, Parigi G et al (2013) NMR crystallography on paramagnetic systems: solved and open issues. Cryst Eng Commun 15:8639–8656. doi:10.1039/C3CE41485J
25. Saio T, Ogura K, Shimizu K et al (2011) An NMR strategy for fragment-based ligand screening utilizing a paramagnetic lanthanide probe. J Biomol NMR 51:395–408. doi:10.1007/s10858-011-9566-5
26. Guan J-Y, Keizers PHJ, Liu W-M et al (2013) Small-molecule binding sites on proteins established by paramagnetic NMR spectroscopy. J Am Chem Soc 135:5859–5868. doi:10.1021/ja401323m
27. John M, Pintacuda G, Park AY et al (2006) Structure determination of protein – ligand complexes by transferred paramagnetic shifts. J Am Chem Soc 128:12910–12916. doi:10.1021/ja063584z
28. Bertini I, Del Bianco C, Gelis I et al (2004) From the cover: experimentally exploring the conformational space sampled by domain reorientation in calmodulin. Proc Natl Acad Sci U S A 101:6841–6846. doi:10.1073/pnas.0308641101
29. Bertini I, Gupta YK, Luchinat C et al (2007) Paramagnetism-based NMR restraints provide maximum allowed probabilities for the different conformations of partially independent protein domains. J Am Chem Soc 129:12786–12794. doi:10.1021/ja0726613
30. la Cruz de L, Nguyen THD, Ozawa K et al (2011) Binding of low molecular weight inhibitors promotes large conformational changes in the dengue virus NS2B-NS3 protease: fold analysis by pseudocontact shifts. J Am Chem Soc 133:19205–19215. doi:10.1021/ja208435s
31. John M, Schmitz C, Park AY et al (2007) Sequence-specific and stereospecific assignment of methyl groups using paramagnetic lanthanides. J Am Chem Soc 129:13749–13757. doi:10.1021/ja0744753
32. Skinner SP, Moshev M, Hass MAS, Ubbink M (2013) PARAssign—paramagnetic NMR assignments of protein nuclei on the basis of pseudocontact shifts. J Biomol NMR 55:379–389. doi:10.1007/s10858-013-9722-1
33. Otting G (2010) Protein NMR using paramagnetic ions. 39:387–405. doi:10.1146/annurev.biophys.093008.131321. http://dx.doi.org/10.1146/annurevbiophys093008131321

34. Pintacuda G, Keniry MA, Huber T (2004) Fast structure-based assignment of 15N HSQC spectra of selectively 15N-labeled paramagnetic proteins. J Am Chem Soc 126:2963–2970. doi:10.1021/ja039339m

35. Bertini I, Janik MBL, Lee Y-M et al (2001) Magnetic susceptibility tensor anisotropies for a lanthanide ion series in a fixed protein matrix. J Am Chem Soc 123:4181–4188. doi:10.1021/ja0028626

36. Pintacuda G, John M, Su X-C, Otting G (2007) NMR structure determination of protein – ligand complexes by lanthanide labeling. Acc Chem Res 40:206–212. doi:10.1021/ar050087z

37. Wöhnert J, Franz KJ, Nitz M et al (2003) Protein alignment by a coexpressed lanthanide-binding tag for the measurement of residual dipolar couplings. J Am Chem Soc 125:13338–13339. doi:10.1021/ja036022d

38. Martin LJ, Hähnke MJ, Nitz M et al (2007) Double-lanthanide-binding tags: design, photophysical properties, and NMR applications. J Am Chem Soc 129:7106–7113. doi:10.1021/ja070480v

39. Ma C, Opella SJ (2000) Lanthanide ions bind specifically to an added "EF-Hand" and orient a membrane protein in micelles for solution NMR spectroscopy. J Magn Reson 146:381–384. doi:10.1006/jmre.2000.2172

40. Zhuang T, Lee HS, Imperiali B, Prestegard JH (2008) Structure determination of a Galectin-3–carbohydrate complex using paramagnetism-based NMR constraints. Protein Sci 17:1220–1231. doi:10.1110/ps.034561.108

41. Su X-C, Huber T, Dixon NE, Otting G (2006) Site-specific labelling of proteins with a rigid lanthanide-binding tag. Chembiochem 7:1599–1604. doi:10.1002/cbic.200600142

42. Xun-Cheng S, McAndrew K, Thomas Huber A, Otting G (2008) Lanthanide-binding peptides for NMR measurements of residual dipolar couplings and paramagnetic effects from Multiple angles. J Am Chem Soc 130:1681–1687. doi:10.1021/ja076564l

43. Kobashigawa Y, Saio T, Ushio M et al (2012) Convenient method for resolving degeneracies due to symmetry of the magnetic susceptibility tensor and its application to pseudo contact shift-based protein–protein complex structure determination. J Biomol NMR 53:53–63. doi:10.1007/s10858-012-9623-8

44. Su X-C, Man B, Beeren S et al (2008) A dipicolinic acid tag for rigid lanthanide tagging of proteins and paramagnetic NMR spectroscopy. J Am Chem Soc 130:10486–10487. doi:10.1021/ja803741f

45. Dvoretsky A, Gaponenko V, Rosevear PR (2002) Derivation of structural restraints using a thiol-reactive chelator. FEBS Lett 528:189–192. doi:10.1016/S0014-5793(02)03297-0

46. Haberz P, Rodriguez-Castañeda F, Junker J et al (2006) Two new chiral EDTA-based metal chelates for weak alignment of proteins in solution. Org Lett 8:1275–1278. doi:10.1021/ol053049o

47. Pintacuda G, Moshref A, Leonchiks A et al (2004) Site-specific labelling with a metal chelator for protein-structure refinement. J Biomol NMR 29:351–361. doi:10.1023/B:JNMR.0000032610.17058.fe

48. Prudêncio M, Rohovec J, Peters JA et al (2004) A caged lanthanide complex as a paramagnetic shift agent for protein NMR. Chem Eur J 10:3252–3260. doi:10.1002/chem.200306019

49. Ikegami T, Verdier L, Sakhaii P et al (2004) Novel techniques for weak alignment of proteins in solution using chemical tags coordinating lanthanide ions. J Biomol NMR 29:339–349. doi:10.1023/B:JNMR.0000032611.72827.de

50. Leonov A, Voigt B, Rodriguez Castañeda F et al (2005) Convenient synthesis of multifunctional EDTA-based chiral metal chelates substituted with an S-mesylcysteine. Chem Eur J 11:3342–3348. doi:10.1002/chem.200400907

51. Gaponenko V, Altieri AS, Li J, Byrd RA (2002) Breaking symmetry in the structure determination of (large) symmetric protein dimers – Springer. J Biomol NMR 24:143–148. doi:10.1023/A:1020948529076

52. Gaponenko V, Sarma SP, Altieri AS et al (2004) Improving the accuracy of NMR structures of large proteins using pseudocontact shifts as long-range restraints. J Biomol NMR 28:205–212. doi:10.1023/B:JNMR.0000013706.09264.36

53. Vlasie MD, Comuzzi C, van den Nieuwendijk AMCH et al (2007) Long-range-distance NMR effects in a protein labeled with a lanthanide–DOTA chelate. Chem Eur J 13:1715–1723. doi:10.1002/chem.200600916

54. Keizers PHJ, Desreux JF, Overhand M, Ubbink M (2007) Increased paramagnetic effect of a lanthanide protein probe by two-point attachment. J Am Chem Soc 129:9292–9293. doi:10.1021/ja0725201

55. Keizers PHJ, Saragliadis A, Hiruma Y et al (2008) Design, synthesis, and evaluation of a lanthanide chelating protein probe: CLaNP-5

yields predictable paramagnetic effects independent of environment. J Am Chem Soc 130:14802–14812. doi:10.1021/ja8054832

56. Loh CT, Ozawa K, Tuck KL et al (2013) Lanthanide tags for site-specific ligation to an unnatural amino acid and generation of pseudocontact shifts in proteins. Bioconjug Chem 24:260–268. doi:10.1021/bc300631z
57. Nitz M, Franz KJ, Maglathlin RL, Imperiali B (2003) A powerful combinatorial screen to identify high-affinity terbium(III)-binding peptides. Chembiochem 4:272–276. doi:10.1002/cbic.200390047
58. Nitz M, Sherawat M, Franz KJ et al (2004) Structural origin of the high affinity of a chemically evolved lanthanide-binding peptide. Angew Chem Int Ed 43:3682–3685. doi:10.1002/anie.200460028
59. Schmitz C, Stanton-Cook MJ, Su X-C et al (2008) Numbat: an interactive software tool for fitting Δχ-tensors to molecular coordinates using pseudocontact shifts. J Biomol NMR 41:179–189. doi:10.1007/s10858-008-9249-z
60. Schmitz C, John M, Park AY et al (2006) Efficient χ-tensor determination and NH assignment of paramagnetic proteins. J Biomol NMR 35:79–87. doi:10.1007/s10858-006-9002-4
61. Banci L, Bertini I, Bren KL et al (1996) The use of pseudocontact shifts to refine solution structures of paramagnetic metalloproteins: Met80Ala cyano-cytochrome c as an example. JBIC 1:117–126. doi:10.1007/s007750050030
62. Banci L, Bertini I, Savellini GG et al (1997) Pseudocontact shifts as constraints for energy minimization and molecular dynamics calculations on solution structures of paramagnetic metalloproteins. Proteins Struct Funct Bioinforma 29:68–76. doi:10.1002/(SICI)1097-0134(199709)29:1<68::AID-PROT5>3.0.CO;2-B
63. Kleywegt GJ, Jones TA (1998) Databases in protein crystallography. Acta Crystallogr D Biol Crystallogr 54:1119–1131. doi:10.1107/S0907444998007100
64. Schüttelkopf AW, van Aalten DMF (2004) PRODRG: a tool for high-throughput crystallography of protein–ligand complexes. Acta Crystallogr D Biol Crystallogr 60:1355–1363. doi:10.1107/S0907444904011679
65. Pedretti A, Villa L, Vistoli G (2002) VEGA: a versatile program to convert, handle and visualize molecular structure on windows-based PCs. J Mol Graph Model 21:47–49. doi:10.1016/S1093-3263(02)00123-7
66. Anton V, Bryksin IM (2010) Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. Biotechniques 48:463–465. doi:10.2144/000113418
67. Banci L, Bertini I, Cavallaro G et al (2004) Paramagnetism-based restraints for Xplor-NIH. J Biomol NMR 28:249–261. doi:10.1023/B:JNMR.0000013703.30623.f7
68. Banci L, Bertini I, Huber JG et al (1998) Partial orientation of oxidized and reduced cytochrome b5at high magnetic fields: magnetic susceptibility anisotropy contributions and consequences for protein solution structure determination. J Am Chem Soc 120:12903–12909. doi:10.1021/ja981791w
69. Güntert P (2004) Automated NMR structure calculation with CYANA. In: Protein NMR techniques. Humana Press, Totowa, pp 353–378
70. Schmitz C, Bonvin AMJJ (2011) Protein–protein HADDocking using exclusively pseudocontact shifts. J Biomol NMR 50:263–266. doi:10.1007/s10858-011-9514-4
71. Schwieters CD, Kuszewski JJ, Tjandra N, Marius Clore G (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65–73. doi:10.1016/S1090-7807(02)00014-9
72. Bertini I, Kursula P, Luchinat C et al (2009) Accurate solution structures of proteins from X-ray data and a minimal set of NMR data: calmodulin – peptide complexes as examples. J Am Chem Soc 131:5134–5144. doi:10.1021/ja8080764
73. Amero CD, Boomershine WP, Xu Y, Foster M (2008) Solution structure of pyrococcus furiosusRPP21, a component of the archaeal RNase P holoenzyme, and interactions with its RPP29 protein partner †. Biochemistry 47:11704–11710. doi:10.1021/bi8015982
74. Xu X, Keizers PHJ, Reinle W et al (2009) Intermolecular dynamics studied by paramagnetic tagging. J Biomol NMR 43:247–254. doi:10.1007/s10858-009-9308-0
75. Wolfgang J, Perez LB, Paris CG et al (2000) Second-site NMR screening with a spin-labeled first ligand. J Am Chem Soc 122:7394–7395. doi:10.1021/ja001241
76. Jahnke W, Simon Rüdisser A, Zurini M (2001) Spin label enhanced NMR screening. J Am Chem Soc 123:3149–3150. doi:10.1021/ja005836g
77. Ogura K, Kobashigawa Y, Saio T et al (2013) Practical applications of hydrostatic pressure to refold proteins from inclusion bodies for NMR structural studies. Protein Eng Des Sel 26:409–416. doi:10.1093/protein/gzt012

78. Marius G, Clore A, Schwieters CD (2003) Docking of protein – protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from 1HN/15N chemical shift mapping and backbone 15N – 1H residual dipolar couplings using conjoined rigid body/torsion angle dynamics. J Am Chem Soc 125:2902–2912. doi:10.1021/ja028893d

79. Nioche P, Liu W-Q, Broutin I et al (2002) Crystal structures of the SH2 domain of grb2: highlight on the binding of a new high-affinity inhibitor. J Mol Biol 315:1167–1177. doi:10.1006/jmbi.2001.5299