

Methods in
Molecular Biology 2199

Springer Protocols

Yu Wai Chen
Chin-Pang Benu Yiu *Editors*

Structural Genomics

General Applications

Second Edition

MOREMEDIA



Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor

John M. Walker

School of Life and Medical Sciences

University of Hertfordshire

Hatfield, Hertfordshire, UK

For further volumes:

<http://www.springer.com/series/7651>

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

Structural Genomics

General Applications

Second Edition

Edited by

Yu Wai Chen

Department of Applied Biology and Chemical Technology and the State Key Laboratory of Chemical Biology and Drug Discovery, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

Chin-Pang Benu Yiu

BMI Biotechnology Appraisals & Investments Limited, Wan Chai, Hong Kong

 **Humana Press**

Editors

Yu Wai Chen
Department of Applied Biology
and Chemical Technology and
the State Key Laboratory of Chemical
Biology and Drug Discovery
The Hong Kong Polytechnic University
Hung Hom, Hong Kong

Chin-Pang Bennu Yiu
BMI Biotechnology Appraisals & Investments Limited
Wan Chai, Hong Kong

ISSN 1064-3745

ISSN 1940-6029 (electronic)

Methods in Molecular Biology

ISBN 978-1-0716-0891-3

ISBN 978-1-0716-0892-0 (eBook)

<https://doi.org/10.1007/978-1-0716-0892-0>

© Springer Science+Business Media, LLC, part of Springer Nature 2014, 2021

Chapter 11 is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). For further details see license information in the chapter.

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

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

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

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.

The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Preface

Two decades have lapsed since the ambitious Protein Structure Initiative (PSI) was launched by the U.S. National Institute of General Medical Sciences in 2000. With the initial enthusiasm curbed, the outcome of structural genomics (SG) can now be more proportionally assessed within an established context. To the researchers who are not directly involved in SG projects, they have been observing with skeptical eyes, pondering the justification of these expensive endeavors. While the grand objective of populating the “protein structure universe” is yet far from complete, it is undeniable that, alongside the course of pursuing this goal, the field of SG has produced many technological advances that transform and accelerate protein production, structural determination, and analysis.

Like before in the first edition, the second edition of this SG-themed book steers clear of collecting interim reports of SG centers. While staying close to the spirit of SG, this volume uniquely emphasizes the benefits to the wider structural research community. It is meant to strike a balance and fill some gaps; the target reader is an “average” structural biologist in a small or medium-sized laboratory. We carefully sampled a diverse range of methods applicable to SG research.

The topics are grouped under three parts: (I) protein production, (II) structural analyses and data management, and (III) modeling, simulation, and visualization. Half of this book is devoted to the first part, as recombinant protein production remains a major bottleneck in many structural projects. We have extended this section to include new methodologies for membrane and metal-binding proteins, as well as high-throughput protein production and screening. As a result of high-throughput practices, structural data is accumulating at an ever-increasing rate. This calls for improved quality control and management. The experimental structure determination contents in the previous volume have been largely replaced by an extended part on computational tools for molecular simulation and visualization. The power of modern-day computing allows experimental results to be interpreted in the light of structural models at the molecular level. Overall, the spectrum of topics reflects the trend towards tackling more diverse challenges of studying macromolecular machineries and complexes.

The preparation of this book falls into the period when there was, unfortunately, a global outbreak of a new coronavirus (SARS-CoV-2, aka 2019-nCoV). Following the release of its genome sequence, the complete set of viral protein structural models were generated within days, as illustrated on the Global Health Drug Discovery Institute portal (<https://ghddi-aillab.github.io/Targeting2019-nCoV>). Virtual and real screening of drug candidates immediately follow. This is modern structure-based medicine at work.

In compiling this volume, we witnessed the generosity of the SG community to share experiences and methods. The outcome is most satisfactory: it represents a global effort with a shared vision. We would like to thank all the authors for their contributions.

Hung Hom, Hong Kong

*Yu Wai Chen
Chin-Pang Bennis Yiu*

Contents

<i>Preface</i>	<i>v</i>
<i>Contributors</i>	<i>xi</i>

PART I PROTEIN PRODUCTION

1 High-Throughput Protein Engineering by Massively Parallel Combinatorial Mutagenesis	3
<i>Yuk Kei Wan, Gigi C. G. Choi, and Alan S. L. Wong</i>	
2 Rational Design and Construction of Active-Site Labeled Enzymes	13
<i>Man-Wah Tsang, Yun-Chung Leung, and Kwok-Yin Wong</i>	
3 Screening and Production of Recombinant Human Proteins: Ligation-Independent Cloning	23
<i>Claire Strain-Damerell, Pravin Mahajan, Alejandra Fernandez-Cid, Opher Gileadi, and Nicola A. Burgess-Brown</i>	
4 Screening and Production of Recombinant Human Proteins: Protein Production in <i>E. coli</i>	45
<i>Nicola A. Burgess-Brown, Pravin Mahajan, Claire Strain-Damerell, Alejandra Fernandez-Cid, Opher Gileadi, and Susanne Gräslund</i>	
5 Screening and Production of Recombinant Human Proteins: Protein Production in Insect Cells	67
<i>Pravin Mahajan, Claire Strain-Damerell, Shubhashish Mukhopadhyay, Alejandra Fernandez-Cid, Opher Gileadi, and Nicola A. Burgess-Brown</i>	
6 Expression Screening of Human Integral Membrane Proteins Using BacMam	95
<i>Pravin Mahajan, Katherine Ellis, Shubhashish Mukhopadhyay, Alejandra Fernandez-Cid, Gamma Chi, Henry Man, Katharina L. Dürr, and Nicola A. Burgess-Brown</i>	
7 High-Throughput Expression Screening in Mammalian Suspension Cells	117
<i>Susan D. Chapple and Michael R. Dyson</i>	
8 In Vitro Production of Perdeuterated Proteins in H ₂ O for Biomolecular NMR Studies	127
<i>Lionel Imbert, Rachel Lenoir-Capello, Elodie Crublet, Alicia Vallet, Rida Awad, Isabel Ayala, Celine Juillan-Binard, Hubert Mayerhofer, Rime Kerfab, Pierre Gans, Emeric Miclet, and Jerome Boisbouvier</i>	

- 9 Minimizing Heterogeneity of Protein Samples
for Metal Transporter Proteins Using SAXS
and Metal Radioisotopes 151
Shah Kamranur Rahman

PART II STRUCTURAL ANALYSES AND DATA MANAGEMENT

- 10 Hydrogen–Deuterium Exchange Mass Spectrometry
for Probing Changes in Conformation and Dynamics of Proteins 159
Pui-Kin So
- 11 BeStSel: From Secondary Structure Analysis to Protein
Fold Prediction by Circular Dichroism Spectroscopy 175
András Micsónai, Éva Bulyáki, and József Kardos
- 12 Navigating the Global Protein–Protein Interaction Landscape
Using iRefWeb 191
*Andrei L. Turinsky, Sam Dupont, Alexander Botzki,
Sabry Razick, Brian Turner, Ian M. Donaldson,
and Shoshana J. Wodak*
- 13 State-of-the-Art Data Management: Improving the Reproducibility,
Consistency, and Traceability of Structural Biology
and in Vitro Biochemical Experiments 209
*David R. Cooper, Marek Grabowski, Matthew D. Zimmerman,
Przemyslaw J. Porebski, Ivan G. Shabalin, Magdalena Woinska,
Marcin J. Domagalski, Heping Zheng, Piotr Sroka,
Marcin Cymborowski, Mateusz P. Czub, Ewa Niedzialkowska,
Barat S. Venkataramany, Tomasz Osinski, Zbigniew Fraczak,
Jacek Bajor, Juliusz Gonera, Elizabeth MacLean,
Kamila Wojciechowska, Krzysztof Konina, Wojciech Wajerowicz,
Maksymilian Chruszcz, and Wlodek Minor*

PART III MODELING, SIMULATION, AND VISUALIZATION

- 14 Protein Structure Modeling with MODELLER 239
Benjamin Webb and Andrej Sali
- 15 Parameterization of a Dioxygen Binding Metal Site
Using the MCPB.py Program 257
Pengfei Li and Kenneth M. Merz Jr.
- 16 Parameterization of Large Ligands for Gromacs Molecular
Dynamics Simulation with LigParGen 277
*Yu Wai Chen, Yong Wang, Yun-Chung Leung,
and Kwok-Yin Wong*
- 17 Simulation of Proteins Modified with a Fluorescent Label 289
Zoe Chan and Yun-Chung Leung
- 18 Protocol for Simulations of PEGylated Proteins with Martini 3 315
*Fabian Grünewald, Peter C. Kroon, Paulo C. T. Souza,
and Siewert J. Marrink*

19	Molecular Data Visualization on Mobile Devices: A Quick Starter's Guide	337
	<i>Chin-Pang Bennu Yiu and Yu Wai Chen</i>	
20	Molecular Data Visualization with Augmented Reality (AR) on Mobile Devices	347
	<i>Chin-Pang Bennu Yiu and Yu Wai Chen</i>	
	<i>Index</i>	357

Contributors

- RIDA AWAD • CNRS, CEA, *Institut de Biologie Structurale (IBS), University of Grenoble Alpes, Grenoble, France*
- ISABEL AYALA • CNRS, CEA, *Institut de Biologie Structurale (IBS), University of Grenoble Alpes, Grenoble, France*
- JACEK BAJOR • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- JEROME BOISBOUVIER • CNRS, CEA, *Institut de Biologie Structurale (IBS), University of Grenoble Alpes, Grenoble, France*
- ALEXANDER BOTZKI • *VIB Bioinformatics Core, Ghent, Belgium*
- ÉVA BULYÁKI • *ELTE NAP Neuroimmunology Research Group, Department of Biochemistry, Institute of Biology, ELTE Eötvös Loránd University, Budapest, Hungary*
- NICOLA A. BURGESS-BROWN • *Structural Genomics Consortium, University of Oxford, Oxford, UK*
- ZOE CHAN • *Department of Applied Biology and Chemical Technology and the State Key Laboratory of Chemical Biology and Drug Discovery, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*
- SUSAN D. CHAPPLE • *IONTAS Ltd, Unit 2, Pampisford, Cambridge, UK*
- YU WAI CHEN • *Department of Applied Biology and Chemical Technology and the State Key Laboratory of Chemical Biology and Drug Discovery, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*
- GAMMA CHI • *Structural Genomics Consortium, University of Oxford, Oxford, UK*
- GIGI C. G. CHOI • *Laboratory of Combinatorial Genetics and Synthetic Biology, School of Biomedical Sciences, The University of Hong Kong, Hong Kong, SAR, China*
- MAKSYMILIAN CHRUSZCZ • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC, USA*
- DAVID R. COOPER • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA; HKL Research, Inc., Charlottesville, VA, USA*
- ELODIE CRUBLET • *NMR-Bio, Grenoble, France*
- MARCIN CYMBOROWSKI • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA*
- MATEUSZ P. CZUB • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA*
- MARCIN J. DOMAGALSKI • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA*
- IAN M. DONALDSON • *Adaptimmune, Abingdon, UK*
- SAM DUPONT • *VIB Bioinformatics Core, Ghent, Belgium*
- KATHARINA L. DÜRR • *Structural Genomics Consortium, University of Oxford, Oxford, UK*

- MICHAEL R. DYSON • *Department of Antibody Engineering, Ichnos Sciences S.A., Biopôle Lausanne-Epalinges, Epalinges, Switzerland*
- KATHERINE ELLIS • *Jenner Institute, University of Oxford, Oxford, UK*
- ALEJANDRA FERNANDEZ-CID • *Structural Genomics Consortium, University of Oxford, Oxford, UK*
- ZBIGNIEW FRATCZAK • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- PIERRE GANS • *CNRS, CEA, Institut de Biologie Structurale (IBS), University of Grenoble Alpes, Grenoble, France*
- OPHER GILEADI • *Structural Genomics Consortium, University of Oxford, Oxford, UK*
- JULIUSZ GONERA • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- MAREK GRABOWSKI • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA*
- SUSANNE GRÄSLUND • *Structural Genomics Consortium, Department of Medicine, Solna, Karolinska Institutet, Solna, Sweden*
- FABIAN GRÜNEWALD • *Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands; Zernike Institute for Advanced Materials, University of Groningen, Groningen, The Netherlands*
- LIONEL IMBERT • *CNRS, CEA, Institut de Biologie Structurale (IBS), University of Grenoble Alpes, Grenoble, France; CNRS, CEA, EMBL, Integrated Structural Biology Grenoble (ISBG), University of Grenoble Alpes, Grenoble, France*
- CELINE JUILLAN-BINARD • *CNRS, CEA, Institut de Biologie Structurale (IBS), University of Grenoble Alpes, Grenoble, France; CNRS, CEA, EMBL, Integrated Structural Biology Grenoble (ISBG), University of Grenoble Alpes, Grenoble, France*
- JÓZSEF KARDOS • *ELTE NAP Neuroimmunology Research Group, Department of Biochemistry, Institute of Biology, ELTE Eötvös Loránd University, Budapest, Hungary*
- RIME KERFAH • *NMR-Bio, Grenoble, France*
- KRZYSZTOF KONINA • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- PETER C. KROON • *Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands; Zernike Institute for Advanced Materials, University of Groningen, Groningen, The Netherlands*
- RACHEL LENOIR-CAPELLO • *CNRS, Laboratoire des biomolécules, LBM, Sorbonne Université, École normale supérieure, PSL University, Paris, France*
- YUN-CHUNG LEUNG • *Department of Applied Biology and Chemical Technology and the State Key Laboratory of Chemical Biology and Drug Discovery, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*
- PENGFEI LI • *Department of Chemistry, Yale University, New Haven, CT, USA*
- ELIZABETH MACLEAN • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- PRAVIN MAHAJAN • *Astex Pharmaceuticals, Cambridge, UK*
- HENRY MAN • *Structural Genomics Consortium, University of Oxford, Oxford, UK*
- SIEWERT J. MARRINK • *Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands; Zernike Institute for Advanced Materials, University of Groningen, Groningen, The Netherlands*

- HUBERT MAYERHOFER • CNRS, CEA, *Institut de Biologie Structurale (IBS), University of Grenoble Alpes, Grenoble, France*
- KENNETH M. MERZ JR. • *Department of Chemistry, Michigan State University, East Lansing, MI, USA; Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA; Institute of Cyber-Enabled Research, Michigan State University, East Lansing, MI, USA*
- EMERIC MICLET • CNRS, *Laboratoire des biomolécules, LBM, Sorbonne Université, École normale supérieure, PSL University, Paris, France*
- ANDRÁS MICSONAI • *ELTE NAP Neuroimmunology Research Group, Department of Biochemistry, Institute of Biology, ELTE Eötvös Loránd University, Budapest, Hungary*
- WLADEK MINOR • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA*
- SHUBHASHISH MUKHOPADHYAY • *Structural Genomics Consortium, University of Oxford, Oxford, UK*
- EWA NIEDZIAŁKOWSKA • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA*
- TOMASZ OSINSKI • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- PRZEMYSŁAW J. POREBSKI • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- SHAH KAMRANUR RAHMAN • *Department of Infection Biology, London School of Hygiene & Tropical Medicine, London, UK*
- SABRY RAZICK • *Infrastructure Services, University of Oslo, Oslo, Norway*
- ANDREJ SALI • *Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA; Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, USA; California Institute for Quantitative Biosciences (QB3), University of California San Francisco, San Francisco, CA, USA*
- IVAN G. SHABALIN • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA*
- PUI-KIN SO • *University Research Facility in Life Sciences, The Hong Kong Polytechnic University, Kowloon, Hong Kong*
- PAULO C. T. SOUZA • *Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands; Zernike Institute for Advanced Materials, University of Groningen, Groningen, The Netherlands*
- PIOTR SROKA • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA*
- CLAIRE STRAIN-DAMERELL • *Diamond Light Source Ltd., Didcot, Oxfordshire, UK*
- MAN-WAH TSANG • *Department of Applied Biology and Chemical Technology and the State Key Laboratory of Chemical Biology and Drug Discovery, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*
- ANDREI L. TURINSKY • *Centre for Computational Medicine, Hospital for Sick Children, Toronto, ON, Canada*

- BRIAN TURNER • *Centre for Computational Medicine, Hospital for Sick Children, Toronto, ON, Canada*
- ALICIA VALLET • *CNRS, CEA, Institut de Biologie Structurale (IBS), University of Grenoble Alpes, Grenoble, France*
- BARAT S. VENKATARAMANY • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- WOJCIECH WAJEROWICZ • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- YONG WANG • *Department of Applied Biology and Chemical Technology and the State Key Laboratory of Chemical Biology and Drug Discovery, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*
- YUK KEI WAN • *Laboratory of Combinatorial Genetics and Synthetic Biology, School of Biomedical Sciences, The University of Hong Kong, Hong Kong, SAR, China*
- BENJAMIN WEBB • *Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA; Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA, USA; California Institute for Quantitative Biosciences (QB3), University of California San Francisco, San Francisco, CA, USA*
- SHOSHANA J. WODAK • *VIB-VUB Center for Structural Biology, Brussels, Belgium*
- MAGDALENA WOJNSKA • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases, University of Virginia, Charlottesville, VA, USA*
- KAMILA WOJCIECHOWSKA • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- ALAN S. L. WONG • *Laboratory of Combinatorial Genetics and Synthetic Biology, School of Biomedical Sciences, The University of Hong Kong, Hong Kong, SAR, China; Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, SAR, China*
- KWOK-YIN WONG • *Department of Applied Biology and Chemical Technology and the State Key Laboratory of Chemical Biology and Drug Discovery, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*
- CHIN-PANG BENNU YIU • *BMI Biotechnology Appraisals & Investments Limited, Wan Chai, Hong Kong*
- HEPING ZHENG • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*
- MATTHEW D. ZIMMERMAN • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA*

Part I

Protein Production



Chapter 1

High-Throughput Protein Engineering by Massively Parallel Combinatorial Mutagenesis

Yuk Kei Wan, Gigi C. G. Choi, and Alan S. L. Wong

Abstract

Exploring how combinatorial mutations can be combined to optimize protein functions is important to guide protein engineering. Given the vast combinatorial space of changing multiple amino acids, identifying the top-performing variants from a large number of mutants might not be possible without a high-throughput gene assembly and screening strategy. Here we describe the CombiSEAL platform, a strategy that allows for modularization of any protein sequence into multiple segments for mutagenesis and barcoding, and seamless single-pot ligations of different segments to generate a library of combination mutants linked with concatenated barcodes at one end. By reading the barcodes using next-generation sequencing, activities of each protein variant during the protein selection process can be easily tracked in a high-throughput manner. CombiSEAL not only allows the identification of better protein variants but also enables the systematic analyses to distinguish the beneficial, deleterious, and neutral effects of combining different mutations on protein functions.

Key words CombiSEAL, Protein engineering, Combinatorial mutagenesis, Combinatorial genetics en masse, High-throughput screening, Protein variant characterization, Next-generation sequencing

1 Introduction

Protein engineering, involving the modification of the original protein sequence, has shown success in enhancing protein functions of antibodies, enzymes, and gene-editing proteins [1]. Knowledge on the rule of how changes of multiple amino acid residues can be combined to improve protein function is enormously useful yet challenging to acquire given the unknown epistasis among different residues [2]. It is impossible to test every combination as the number of mutants increases dramatically with every additional residue to be changed. Prior knowledge and structural information can constrain the number of residues to be changed, allowing for an efficient screening of protein variants.

A protein variant library can be built by multiple sites-directed mutagenesis, where short oligonucleotides can encode the

mutations if the mutations lie close to each other. Otherwise, the library of protein variants can be assembled either using longer fragments synthesized with higher cost and error rate or seamlessly ligating mutagenized fragments using methods such as Golden Gate assembly [3] and Gibson assembly [4], which require costly long-read sequencing to track and screen a large number of mutants. Furthermore, long-read sequencing makes accurately identifying the mutations among highly similar sequences in the library difficult due to its high error rate. The CombiSEAL platform can overcome this limitation by introducing short barcode to every mutagenized part, which can be concatenated into a unique combination of barcodes at one end of the genetic construct for short-read sequencing [5]. This method can be applied to assemble any protein variants by modularizing the sequence into multiple segments. Seamless assembly is achieved by flanking the barcoded segments with any Type IIS restriction enzyme sites to give digested overhangs originating from the protein-coding sequence. Including another set of Type IIS restriction enzyme sites in the linker connecting the mutagenized part and the short unique barcode specifying predetermined mutations allows for seamless integration of successive rounds of pool ligation of succeeding parts of the protein.

As barcoded combination mutants can be easily tracked by high-throughput short-read sequencing, CombiSEAL offers an efficient and cost-effective way to scale up the experimentation of a massive number of combination mutants to study epistasis and decipher how changes in sequences leads to corresponding changes in protein activities. This method circumvents the need to characterize a large number of clonal isolates. By knowing the rule of the sequence-to-activity relationship, it could accelerate the protein optimization process and lead to a better design in protein engineering.

2 Materials

2.1 For Assembling a Barcoded Combinatorial Library by CombiSEAL

1. Any protein-coding DNA sequence.
2. PCR primers for amplifying and introducing site-directed mutations of gene fragments for direct cloning into expression or storage vectors.
3. High-fidelity DNA polymerase.
4. Prebarcoded storage vectors (*see Note 1*).
5. Expression vector.
6. Any two Type IIS restriction enzymes compatible with the protein-coding and vector sequences (*see Note 2*).
7. Ligase.
8. Competent cells with high transformation efficiency.

2.2 Sample Preparation for Barcode Sequencing

1. Plasmid purification kit.
2. Genomic DNA extraction kit (e.g., Qiagen DNeasy Blood and Tissue Kit for mammalian cells).
3. High-fidelity DNA polymerase.
4. PCR primers with Illumina adapter sequences.
5. DNA quantification kit (e.g., Quant-iT PicoGreen dsDNA Assay Kit from Life Technologies).
6. SPRI (Solid Phase Reversible Immobilisation) paramagnetic beads for size selection and PCR purification (e.g., Agencourt AMPure XP beads from Beckman Coulter Genomics).
7. Illumina Library Quantification from Kapa Biosystems or NEBNext Library Quant Kit for Illumina from NEB.
8. SYBR Green PCR Mix for real-time PCR.
9. Agilent 2100 Bioanalyzer with the high-sensitivity DNA chip from Agilent.
10. Illumina HiSeq or other next-generation sequencing (NGS) platforms.
11. BD Influx or other fluorescence activated cell sorters.

3 Methods

3.1 Creating Mutagenized Parts

3.1.1 Creating Mutagenized Parts Using a Library of Prebarcoded Storage Vectors

1. Create barcoded storage vectors that can be linearized by restriction enzyme(s) for cloning in the insert via Gibson assembly. The insertion site should be flanked by two Type IIS restriction enzyme sites in a specific orientation as described in Fig. 1. Synthesize random oligo sequences (e.g., 8-base-pair NNNNNNNN), and clone them into the storage vector to serve as a randomized barcode sequence as described in Fig. 1. Make sure that the barcoded vector library has sufficient barcode diversity so that each clone is represented by a unique barcode.
2. Select the protein residues to be mutagenized. Modularize the protein-coding sequence into multiple parts (Fig. 2). Use high-fidelity DNA polymerase to amplify and/or mutate the sequence by PCR. To create site-directed mutations at specific amino acid sites, design the primers with specific codon changes in those sites. Include overlapping end of the linearized storage vector into the primer sequences for cloning in the PCR fragments via Gibson assembly.
3. Isolate single clone after transforming the Gibson assembly product into bacterial competent cells. Verify the sequence of the insert and barcode of each clone by Sanger sequencing.

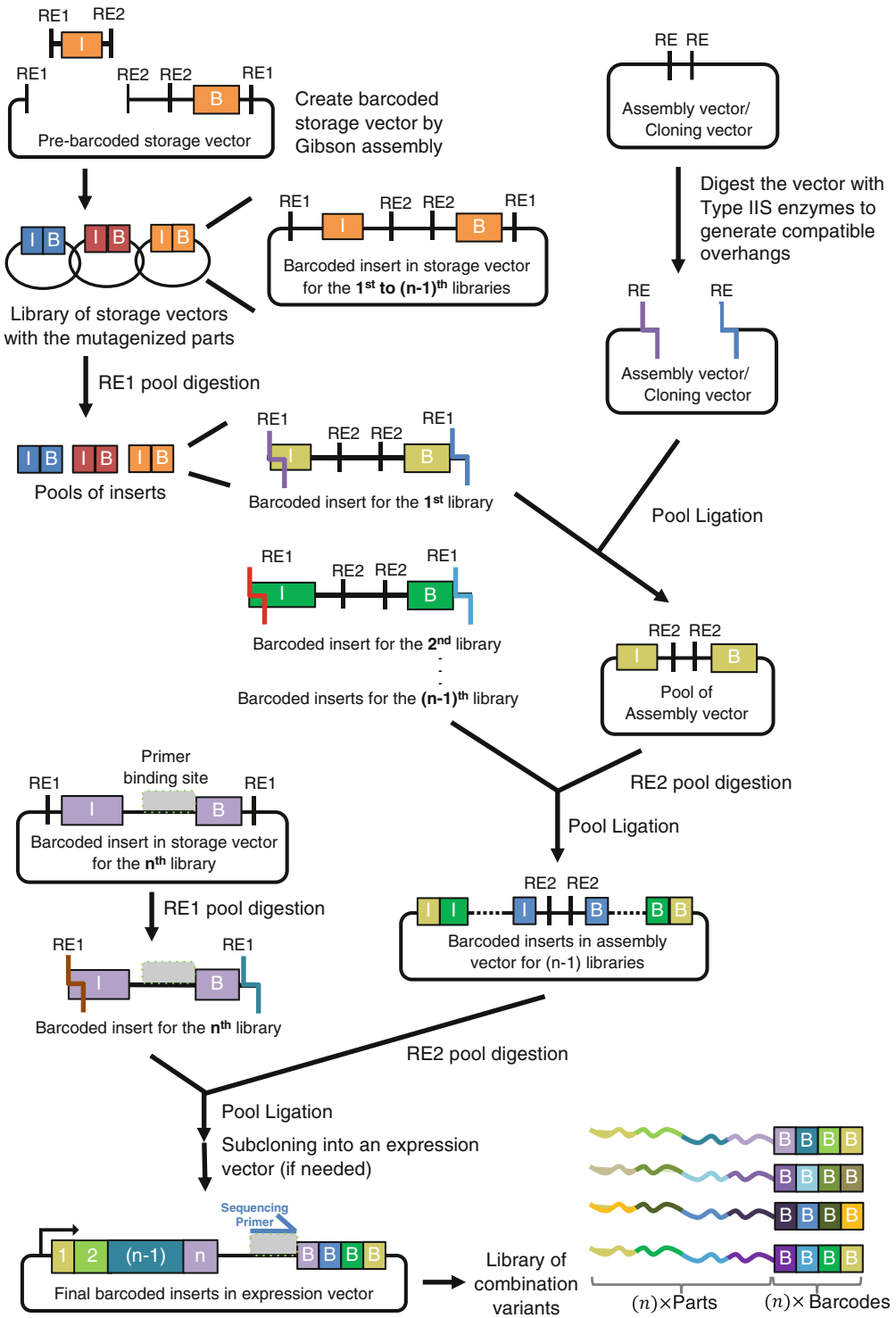


Fig. 1 Strategy for seamless assembly of the barcoded combination mutant library pool. To build a barcoded combination mutant library using storage vectors with a random barcode, the storage vector containing two

4. Mix the modularized insert storage vectors at equal molar ratio. Digest the pooled vectors with the first Type IIS restriction enzyme (i.e., RE1) to generate the pooled inserts. It is important to minimize variation in representation across the inserts with different combinatorial mutations.
5. The inserts will be used in Subheading 3.2.

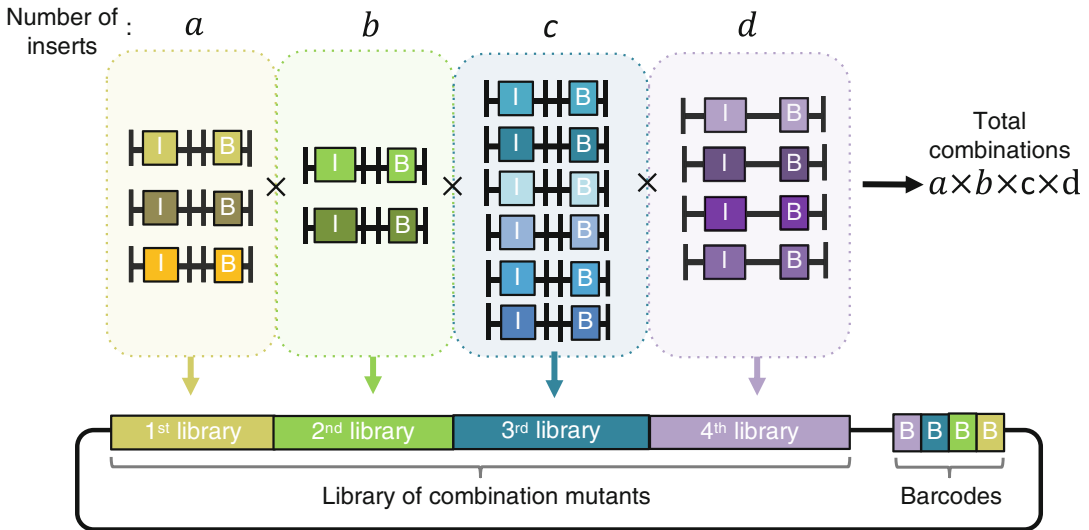


Fig. 2 Modularization of a protein-coding sequence. Protein-coding sequence can be modularized into multiple parts (four segments in this example). The number of mutagenized inserts in each library is flexible, where a , b , c , and d denote the number of inserts in each of the four libraries respectively. Assembly of a protein variant library from each part will give a total number of $a \times b \times c \times d$ combinations (or multiplication of the number of inserts in each insert library)

Fig. 1 (continued) sets of Type IIS restriction enzyme sites (RE1 and RE2) is first linearized for the insertion of DNA parts via Gibson assembly. After the libraries of storage vector with the mutagenized parts are built, these libraries are pool-digested with RE1 to generate barcoded fragments with overhangs complementary to that of the digested assembly vector for the pool ligation afterward. After the insertion of the first DNA parts, the libraries of assembly vector are digested with RE2 to generate overhangs compatible with the digested inserts from subsequent libraries (until the $(n-1)$ th library) of storage vector digested by RE1 for pool ligation. The storage vector of the n th insert library is different from that of the other libraries as it does not contain RE2 sites between the insert and the barcode but instead has a primer binding site for NGS. The final assembly vector library contains the full protein-coding sequence composed of n modules that are linked to barcodes concatenated at one end. If the assembly vector is not an expression vector, then the barcoded protein-coding sequences can be subcloned into the expression vector for protein expression in the host cells for selection. “I” and “B” denote the insert and the barcode, respectively

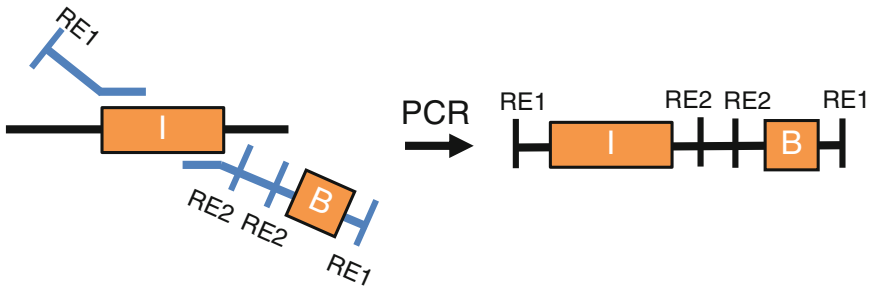


Fig. 3 An alternative PCR-based method to generate barcoded mutagenized inserts. Primers are synthesized for incorporating mutations, type IIS restriction enzyme sites (RE1 and RE2), and barcodes into the protein-encoding parts by PCR. Digestion of the PCR products with RE1 enzymes generates barcoded fragments with overhangs complementary to that of the digested assembly vector for pool ligation, as described in Fig. 1. “I” and “B” denote the insert and the barcode, respectively

3.1.2 Creating Mutagenized Parts with Barcodes by PCR (an Alternative Strategy to Subheading 3.1.1 That Does Not Require the Storage Vectors)

1. Design primers with restriction sites according to Fig. 3.
2. Use high-fidelity DNA polymerase to amplify and/or mutate the sequence by PCR.
3. Keep a portion of the PCR products for Sanger sequencing.
4. Quantify and mix the PCR products at equal molar ratio. Digest the pooled PCR products with RE1 to generate the pooled inserts.
5. Use PCR purification kit or Agencourt AMPure XP beads to purify the digested PCR products, which will be used in Subheading 3.2.

3.2 Creating a Barcoded Combinatorial Protein Library

1. Select a suitable cloning vector for assembling the protein-coding sequence.
2. Modify or construct a cloning vector by introducing restriction enzyme sites that generate compatible overhangs upon cutting for the insertion of barcoded parts from Subheading 3.1.1 or 3.1.2.
3. Ligate the cloning vector and the inserts of the first modularized part of the protein.
4. Transform the library of vectors into competent cells for plasmid preparation (Refer to Subheading 3.4).
5. Digest the library of cloning vectors with the second Type IIS restriction enzyme (i.e., RE2) and ligate it with the second modularized parts. Then, the cloning vector will contain libraries of the first and second modularized parts.
6. Repeat **step 4** until the cloning vectors incorporate the libraries for all modularized parts. The final library of constructs should have the full-length protein-coding sequence, where the modularized segments are seamlessly linked together while the barcodes are concatenated at one end.

7. Subclone the final assembled library of protein-coding sequences with barcodes into an expression vector for screening in host cells if the cloning vector being used is not an expression vector (e.g., Lentiviral vector for delivering constructs to mammalian cells).

3.3 Amplification of the Plasmid Library

1. Since each plasmid library is a mixture of different DNA constructs, maintaining its diversity by amplifying the library is important. Use highly efficient competent cells to transform the plasmid library.
2. To avoid the loss of representation, plate several dilution plates for colony counting and calculating the transformation efficiency. A coverage of 100-fold or more colonies per construct in the plasmid library is recommended.
3. To prevent recombination between the lentiviral long-terminal repeats, either use bacterial strain that reduces the frequency of homologous recombination or incubate the bacterial culture at a lower temperature is recommended.
4. Purify plasmid library with a midi- or maxi-scale plasmid purification kit.
5. Sequence the plasmid library to verify the representation of variants and the diversity of the pool.

3.4 Expressing the Protein Library in Cell Culture for Screening

1. Plasmid library of the expression vectors can be directly transformed into an *E. coli* expression strain if the screen is to be performed in *E. coli*. For screening in mammalian cells, a lentiviral vector library with packaging and envelope plasmids can be transfected into HEK293T cells for lentivirus production. Infect mammalian cells at a low multiplicity of infection (MOI) to ensure that most of the cells in the population carries only one copy of construct.
2. Choose an appropriate selection scheme for picking the clones of interest as different protein may require different strategy for various desired property. In general, if the desired protein confers a selective growth advantage or a better survival fitness in a particular environment, clones with desirable properties will be enriched after the selection process, and this will be reflected by their representation in the pool. If the desirable phenotype can be translated into a gain or a loss of fluorescent signal by using a fluorescent reporter system in the host cells, fluorescence activated cell sorting (FACS) can be utilized to select desirable variants from the pool.
3. Perform the screen at a 300-fold or more representation with at least 2 biological replicates (*see* **Note 3**).

3.5 Preparing Samples for Barcode Sequencing

1. Harvest both the selected and the total population of cells.
2. Extract the genomic DNA from the cells (e.g., use DNeasy Blood and Tissue kit for mammalian cells).
3. Quantify the DNA concentration using Quant-iT PicoGreen dsDNA Assay kit.
4. Calculate the amount of genomic DNA required to recover the barcodes from the cells by PCR amplification to achieve desired coverage of the library (*see Note 4*).
5. Amplify the barcoded region using PCR primers containing the P5/P7 flowcell attachment sequence, Illumina primer binding sequence, vector primer binding sequence, and the indexing barcode for multiplexed sequencing (*see Note 5*).
6. Purify the PCR amplicons of correct fragment size using Agencourt AMPure XP beads.
7. Quantify the purified PCR amplicons by real-time PCR using Illumina Library Quantification (Kapa Biosystems) or NEB-Next Library Quant Kit for Illumina (NEB).
8. Assess the quality of the quantified samples using Agilent 2100 Bioanalyzer with the high-sensitivity DNA chip. About 5–10 ng/ μ l DNA sample is needed.
9. Pool the quantified samples at a ratio based on their desired share of sequencing reads, and sequence the pooled sample with the Illumina HiSeq sequencing system.

3.6 Barcode Sequencing Data Analysis

1. Categorize the barcode reads from the sequencing data based on the indexing barcodes.
2. Normalize the barcode reads for each combination into per million reads.
3. Calculate the frequency of each combination mutant between selected population and the total population without selection. Changes in the relative abundance of the combination in a pool suggest a selective advantage or disadvantage during the selection process (*see Note 6*).
4. Calculate the enrichment ratio (E) by comparing the frequency of each mutant in the selected population (N_{selected}) to that in the total population (N_{total}) relative to the rest of the population, where $E = (N_{\text{selected}}/N_{\text{total}})/((1 - N_{\text{selected}})/(1 - N_{\text{total}}))$.
5. Calculate the mean of enrichment ratio from multiple biological replicates to compute the \log_2 -transformed mean ($\log_2(E)$) for ranking the mutants (*see Note 7* and Fig. 4).

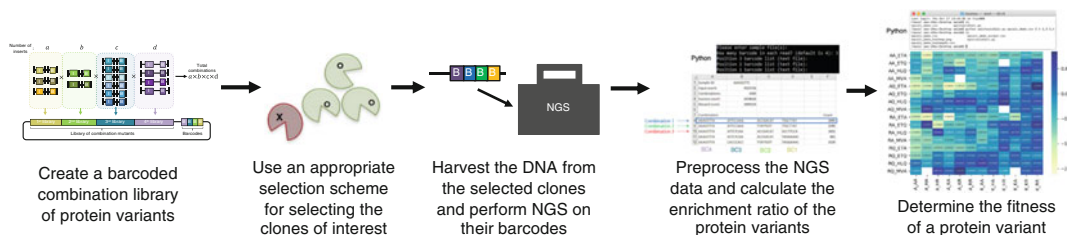


Fig. 4 The workflow of the CombiSEAL screening platform. After building a barcoded combination library of protein variants and selecting the desired clones by an appropriate selection scheme, genomic DNA purified from the selected population can be used for NGS to retrieve the barcodes for analysis. Using the NGS data, read count for each variants can be generated using the BC-analyzer program available on Github (<https://github.com/AWHKU/BC-analyzer>) for calculating the enrichment ratio of each variant. Further analyses including mutational epistasis can be done by using the Epistasis Calculator program available on Github (<https://github.com/AWHKU/epistasisCalculator>) for calculating the epistasis score of each variant

3.7 Epistasis Analysis

1. Calculate the expected fitness [6, 7] for a combination variant $[X,Y]$ by $\log_2(E_{[X]}) + \log_2(E_{[Y]})$.
2. Calculate the epistasis scores by (observed fitness) – (expected fitness). Combinations showing better observed fitness than expected fitness have a positive epistasis while those with worse observed fitness than expected fitness have a negative epistasis (*see* **Note 8** and Fig. 4).
3. Exclude the $\log_2(E)$ values of combination mutants causing a lethal phenotype to avoid false detection of predicted fitness.

4 Notes

1. Inserts can be barcoded using a prebarcoded storage vector or by PCR methods [5, 8, 9]. In the CombiSEAL study, the mutant insert and the barcode sequence were cloned into a storage vector in the configuration of BsaI–insert–BbsI–BbsI–barcode–BsaI. However, it should be noted that the configuration of the last insert was different from the first to $(n-1)$ libraries as described in Fig. 1. The last insert should include a primer binding site between the insert and the barcode in the configuration of BsaI–insert–primer binding site–barcode–BsaI.
2. Avoid using Type IIS restriction enzyme sites contained within the protein-coding sequence or remove those Type IIS restriction enzyme sites by changing a nucleotide to introduce a silent mutation.
3. Reduce the experimental noise by increasing the fold representation of cells per combination in the pooled screen.

4. PCR condition should be optimized to ensure that the number of PCR cycles falls within the linear phase of amplification to avoid bias in PCR.
5. To give sufficient sequence diversity of amplicon across flow cells, use a mixture of primers with stagger regions of different length to balance the base composition during Illumina sequencing.
6. Remove the barcodes with too little reads (e.g., less than 100 reads) in the total populations (or the control group) to remove noise resulting from low-representation of variants.
7. We have generated a Python program (bc_analyzer_v4), which is available on Github (<https://github.com/AWHKU/BC-analyzer>), for calculating the enrichment ratio of each variant.
8. Mutational epistasis can be determined by using a program we generated and posted on Github (<https://github.com/AWHKU/epistasisCalculator>) (last updated on Oct 22, 2019) for calculating the epistasis score of each variant.

Acknowledgments

This work was supported by the Croucher Foundation Start-up Allowance and Hong Kong Research Grants Council (GRF-17104619) to A.S.L.W.

References

1. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K (2012) Engineering the third wave of biocatalysis. *Nature* 485 (7397):185–194. <https://doi.org/10.1038/nature11117>
2. Fowler DM, Fields S (2014) Deep mutational scanning: a new style of protein science. *Nat Methods* 11(8):801–807. <https://doi.org/10.1038/nmeth.3027>
3. Engler C, Kandzia R, Marillonnet S (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS One* 3 (11):e3647. <https://doi.org/10.1371/journal.pone.0003647>
4. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA 3rd, Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* 6(5):343–345. <https://doi.org/10.1038/nmeth.1318>
5. Choi GCG, Zhou P, Yuen CTL, Chan BKC, Xu F, Bao S, Chu HY, Thean D, Tan K, Wong KH, Zheng Z, Wong ASL (2019) Combinatorial mutagenesis en masse optimizes the genome editing activities of SpCas9. *Nat Methods* 16 (8):722–730. <https://doi.org/10.1038/s41592-019-0473-0>
6. Aakre CD, Herrou J, Phung TN, Perchuk BS, Crosson S, Laub MT (2015) Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* 163 (3):594–606. <https://doi.org/10.1016/j.cell.2015.09.055>
7. Olson CA, Wu NC, Sun R (2014) A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol* 24(22):2643–2651. <https://doi.org/10.1016/j.cub.2014.09.072>
8. Wong AS, Choi GC, Cheng AA, Purcell O, Lu TK (2015) Massively parallel high-order combinatorial genetics in human cells. *Nat Biotechnol* 33(9):952–961. <https://doi.org/10.1038/nbt.3326>
9. Wong AS, Choi GC, Cui CH, Pregernig G, Milani P, Adam M, Perli SD, Kazer SW, Gaillard A, Hermann M, Shalek AK, Fraenkel E, Lu TK (2016) Multiplexed bar-coded CRISPR-Cas9 screening enabled by CombiGEM. *Proc Natl Acad Sci U S A* 113 (9):2544–2549. <https://doi.org/10.1073/pnas.1517883113>



Rational Design and Construction of Active-Site Labeled Enzymes

Man-Wah Tsang, Yun-Chung Leung, and Kwok-Yin Wong

Abstract

With a growing amount of structural information of proteins, deciphering the linkage between the structure and function of these proteins is the next important task in structural genomics. To characterize the function of an enzyme at molecular level, placing a reporter on the active site of an enzyme can be a strategy to examine the dynamics of the interaction between enzyme and its substrate/inhibitor. In this chapter, we describe an approach of active-site labeling of enzyme for this purpose. Provided with the fabrication of a fluorescein-labeled AmpC β -lactamase as an example, we herein depict the methodology of a structure-based selection of the location in an enzyme's active site for bioconjugation and the preparation of the active-site labeled enzyme.

Key words Active site, Bioconjugation, Chemical modification, Enzyme, Fluorescence labeling, Protein engineering

1 Introduction

Enzymes are biocatalysts involved in diverse biochemical processes in all living cells [1]. Besides, they have a wide variety of industrial and biomedical applications [1–5]. Understanding their molecular functions are important for both fundamental and applied science. In this sense, numerous biochemical and biophysical methods have been devised to achieve this purpose [6–9].

In this chapter, we present a method of active-site labeling of enzyme for the characterization of enzyme–substrate/inhibitor interaction. In this approach, a small molecular probe is covalently conjugated onto the active site pocket of an enzyme. With the objective that the conjugated probe can sense and report the local environmental changes occurring inside the active site upon the substrate/inhibitor binding, the attachment site is rationally chosen based on the already-known three-dimensional structure of the enzyme. This method has been previously adopted to construct a series of fluorescently labeled β -lactamases that enable a real-time

monitoring of the interaction between these enzymes and their β -lactam substrates/inhibitors [10–14]. These fluorescent enzymes were β -lactamases covalently attached with a single fluorescein onto their active site's entrance via a maleimide linker. Here, the fluorescein was strategically placed onto a location without perturbing the binding capability of the enzymes. The preparation of these labeled enzymes involves two steps: (1) a cysteine point mutation at the attachment site of the enzyme to generate a thiol reactive group for the labeling reaction; and (2) a fluorophore attachment via a cross-linking reaction between the sulfhydryl group of the introduced cysteine and the maleimide group of a fluorescein-5-maleimide (Fig. 1). The attached fluorescein enables a track of the dynamic change taken place in the active site pocket [10, 11, 14]. When the active site is empty, it partially occupies the pocket (Fig. 2a) and gives out a low fluorescence intensity. With a binding of substrate/inhibitor to the active site, it is displaced by this molecule and becomes more exposed to the solvent environment (Fig. 2b). This subtle change in the microenvironment of the fluorescein leads to an increase in the fluorescence intensity of this fluorescent probe (Fig. 3). The fluorescence keeps turning-on as long as the active-site is being occupied whereas it returns to its basal level when the active site is free (Fig. 3). Based on this binding event-modulated fluorescence switching system, the pattern of the fluorescence signal varies in accordance to the mode of interaction between the enzyme and the substrate/inhibitor [10, 11, 14]. This provides a real-time monitoring of the trajectory of these various interactions. For further study of the enzyme's structure–function relationship, mutation of interest can be introduced into the labeled enzyme and its influence on the enzyme's interaction with its substrates/inhibitors can be assessed by fluorescence measurements [13].

In this article, we utilize the construction of a fluorescein-labeled AmpC β -lactamase [11] as an example to demonstrate the selection criteria of the attachment site of a fluorescein (reporter) on the enzyme's active site and the bioconjugation procedure.

2 Materials

Reagents for site-directed mutagenesis, expression, and purification of a specific protein are required but not listed here. Freshly prepare a 20 mM stock of fluorescein-5-maleimide (F5M stock) by dissolving the dye in dimethylformamide (DMF) or dimethyl sulfoxide (DMSO). Wrap the fluorescent dye with aluminum foil to protect it from light. Use 15 mL ultracentrifugal device (e.g., Amicon[®] Ultra filter device (Millipore) or Vivaspinn[®] Turbo ultrafiltration centrifugal concentrator (Sartorius)) for buffer exchange and concentration of the protein sample. Select the device with Molecular Weight Cutoff (MWCO) that is at least twofold smaller than the molecular

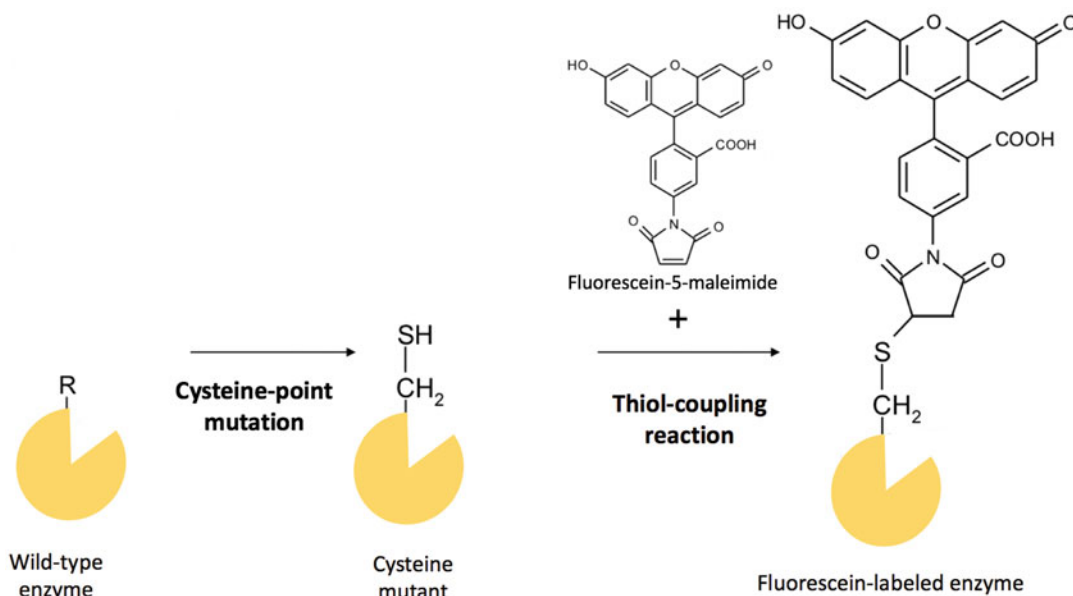


Fig. 1 Generation of a fluorescein-labeled β -lactamase. The desired fluorophore attachment site of a wild-type enzyme is first substituted by a cysteine via site-directed mutagenesis. Then the cysteine mutant is chemically modified with a fluorescein-5-maleimide via a thiol reaction to give the fluorescent β -lactamase

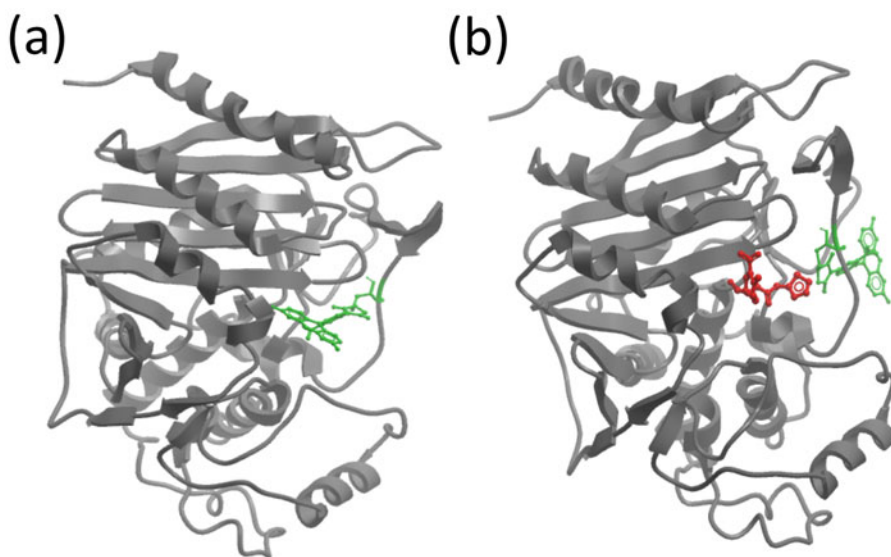


Fig. 2 A fluorescein-labeled AmpC β -lactamase in its (a) *apo* form and (b) substrate-bound form. Green: Fluorescein; Red: β -lactam substrate. (Reprinted figure with permission from ref. 11. Copyright 2011 American Chemical Society)

weight of the protein sample. Rinse and prewet the membrane in the sample reservoir of the device with Milli-Q water. Keep fluid (Milli-Q water or buffer) inside the reservoir to avoid drying out of the wet membrane prior use.

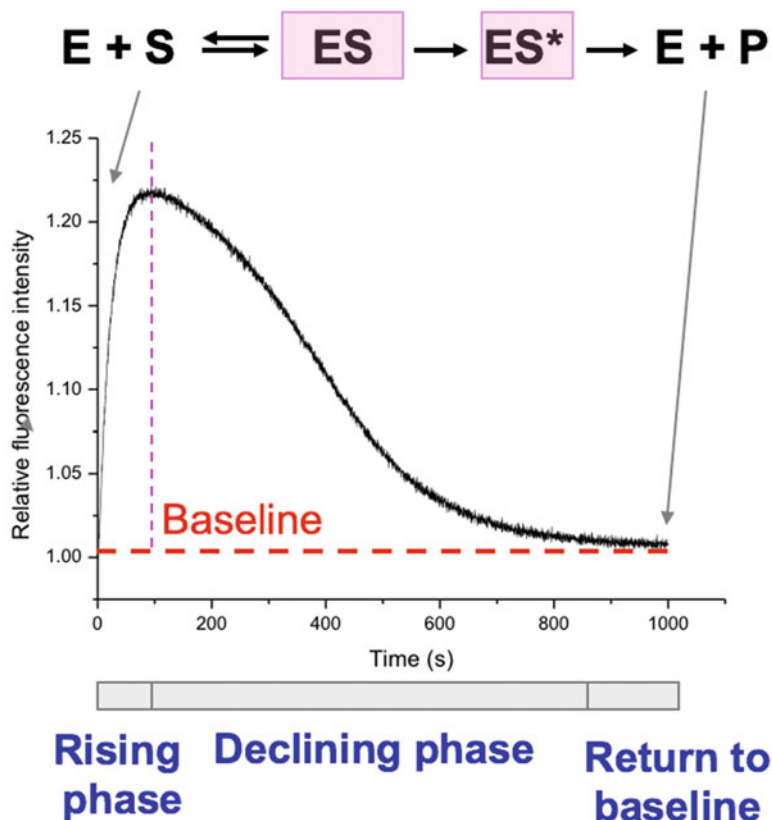


Fig. 3 Fluorescence trace reflects the progress of the hydrolysis of a β -lactam antibiotic (penicillin G) by a fluorescein-labeled AmpC β -lactamase. The rising phase indicates the fluorescence enhancement upon the substrate binding to the active site; the declining phase reveals the reduction of the amount of fluorescent species, ES and ES*, due to the turnover of substrate; and the return of the fluorescence signal to the basal level (relative fluorescence intensity = 1) is resulted from the regeneration of the active site due to the leave of the product (P). *E* free enzyme, *S* substrate, *ES* pre-covalent substrate–enzyme complex, *ES** acyl–enzyme complex, *P* hydrolyzed product

2.1 Acquisition and visualization of the Enzyme's Structural Information

1. Access to Protein Data Bank via <https://www.rcsb.org> [15] and retrieve the structural data of the protein of interest. Our example is *Enterobacter cloacae* P99 AmpC β -lactamase and its PDB ID is 1XX2 [16]. To search for the protein structural information, (1) input the name (*Enterobacter cloacae* P99 AmpC β -lactamase) or PDB ID (1XX2) into the red-circled column and (2) click the “Go” button (Fig. 4).
2. After the page regarding the information of the target protein has been launched, (1) go to “Download Files” and (2) select “PDB Format” to download the PDB file of the protein (Fig. 5).

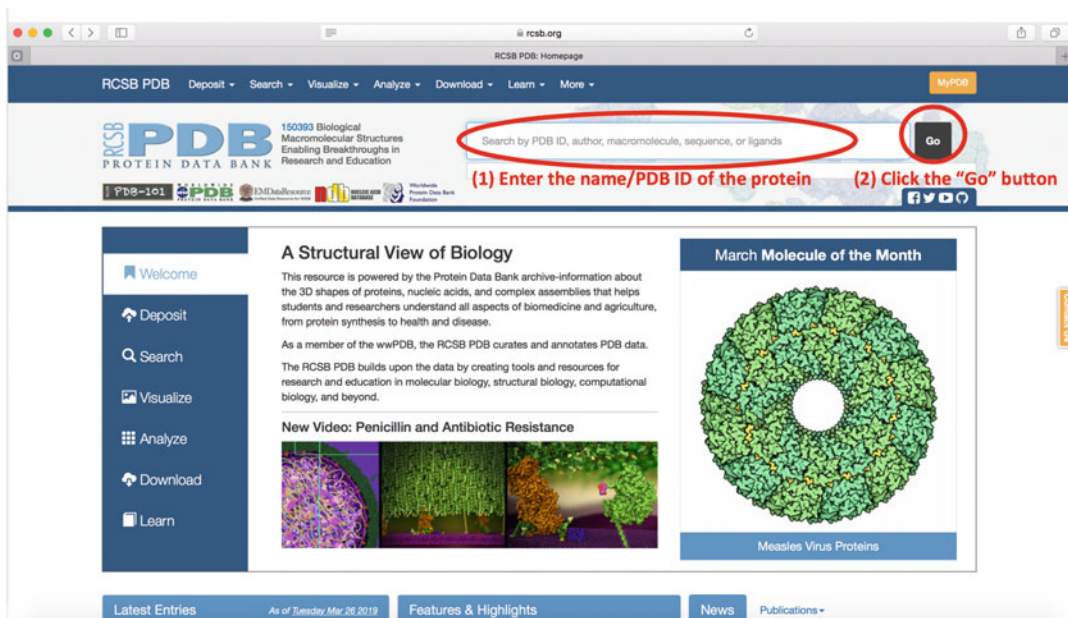


Fig. 4 Procedure for searching structural data of the protein in Protein Data Bank [15]

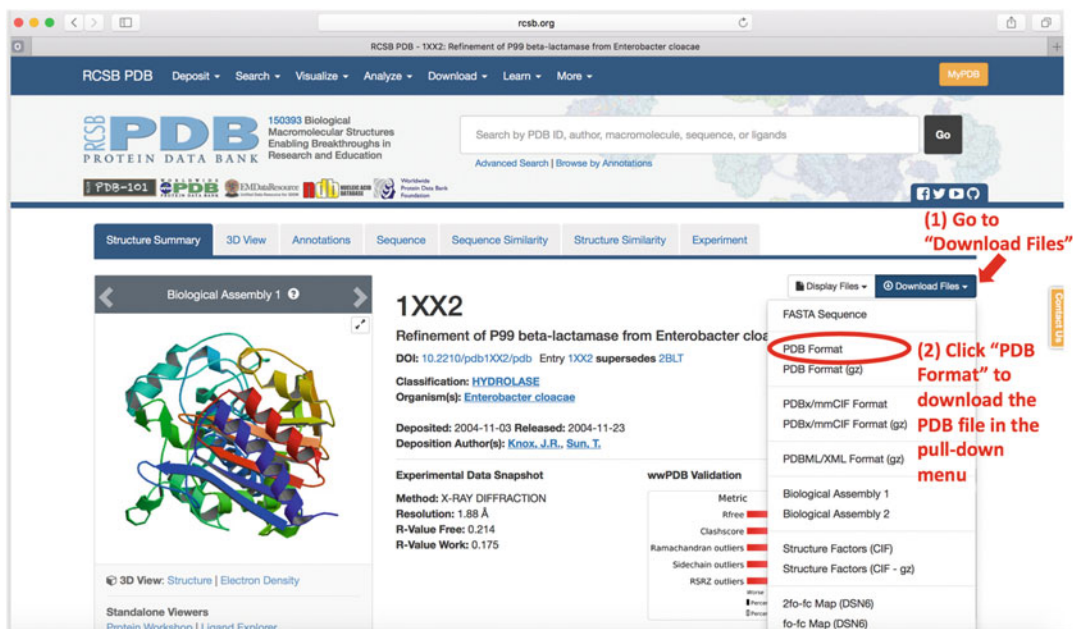


Fig. 5 Procedure for obtaining the PDB file of the protein in Protein Data Bank [15]

3. Load the PDB file with a molecular visualization program (e.g., PyMOL [17] and Swiss-Pdb Viewer [18]) to view the three-dimensional structure of the protein (*see Note 1*).

2.2 Selection of Site for Fluorophore Attachment

From the construction of a series of fluorescein-labeled β -lactamases [10–14], we have generalized a rationale in deciding the fluorophore attachment site, which may be applicable to other enzymes. The followings suggest the selection criteria of our model:

- (a) The fluorophore has to be placed in close proximity to the active site residue so that it can have a higher chance to experience the changes in microenvironment of the active site pocket.
- (b) The fluorophore has to be introduced at the entrance of the active site so as to avoid perturbing the binding and catalytic capabilities of the enzyme.
- (c) The fluorophore has to be attached onto a flexible region of the enzyme to confer the flexibility of conjugated fluorophore to move in response to the dynamic changes taken place in the active site pocket.

To illustrate the above selection criteria, a fluorescent AmpC β -lactamase, V211Cf, is used as an example [11]. V211Cf was generated from a cysteine-free AmpC β -lactamase from *Enterobacter cloacae*. Viewing the structure of the wild-type AmpC β -lactamase (PDB ID: 1XX2), position 211, at which a Valine is situated, is a suitable site for the fluorophore attachment (Fig. 6).

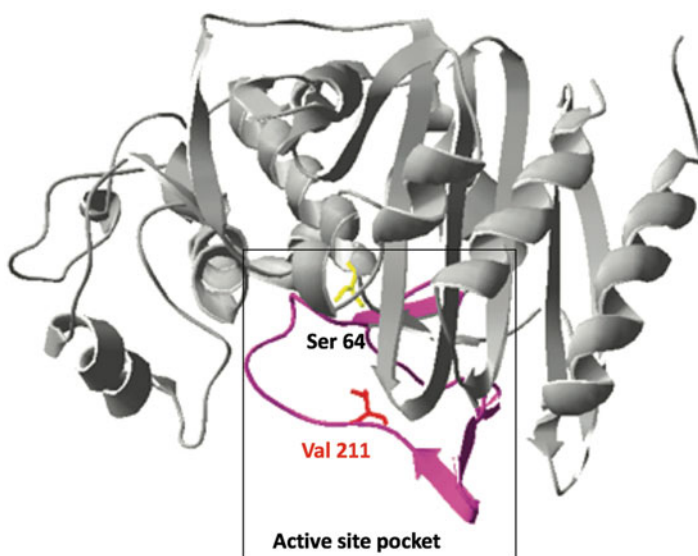


Fig. 6 Valine 211 of AmpC β -lactamase as a site chosen for cysteine substitution and fluorophore attachment. Red: Valine 211; Yellow: Serine 64 (active site serine); Magenta: flexible Omega loop; Black line box: active site pocket. PDB ID: 1XX2 (ref. 16). (Reprinted figure with permission from ref. 11. Copyright 2011 American Chemical Society)

First, Valine 211 is close to the active site (Serine 64). Second, it resides on entrance of the active site pocket. Lastly, it is located on a long flexible loop of the enzyme.

2.3 Preparation of Cysteine Mutant

1. Mutate the selected site for the fluorophore attachment into cysteine by site-directed mutagenesis (*see Note 2*).
2. Overexpress the cysteine mutant and purify it to >90% homogeneity.
3. Lyophilize the purified protein and store the lyophilized protein at $-20\text{ }^{\circ}\text{C}$ or $-80\text{ }^{\circ}\text{C}$ until the labeling reaction.

3 Methods

3.1 Protein Labeling with Fluorescein-5-Maleimide

1. Reconstitute 2 mg of the lyophilized protein into 4 mL of a suitable buffer (*see Note 3*).
2. Adjust the pH of the protein solution to 7.0 (*see Note 4*).
3. Add a five- to ten-fold molar excess of the fluorescein-5-maleimide from the F5M stock to the protein solution. Wrap the reaction mixture with aluminum foil to protect it from light.
4. Incubate the reaction mixture for 2–4 h with stirring in the dark.
5. Add the reaction mixture to the sample reservoir of a prerinsed 15 mL ultracentrifugal filter device for the removal of excess dye, buffer exchange, and sample concentration (*see Note 5*).
6. For the case of using centrifuge with swing-bucket rotor, fill the reservoir with buffer up to 15 mL whereas for the case of centrifuge with fixed angle rotor, add buffer to final volume of 11 mL.
7. Cap the filter device and centrifuge it at $4000 \times g$ for 15–30 min (*see Note 6*).
8. Keep the retentate in the sample reservoir as the labeled protein is supposed to retain in the reservoir, and discard the flowthrough.
9. Repeat **steps 6–8** for approximately 6–7 times until the flowthrough becomes clear in color.
10. Collect the labeled protein and check the efficiency of labeling (*see Subheading 3.2*).
11. Store the labeled sample at a suitable condition until further experiment (*see Note 7*).

3.2 Validation of the Labeling Reaction

3.2.1 Qualitative Analysis by SDS-PAGE

1. Load the protein marker, unlabeled protein (negative control) and labeled protein onto three separate lanes of a SDS-polyacrylamide gel.
2. Run the SDS-PAGE electrophoresis.
3. Illuminate the gel with UV light and visualize if there is any fluorescent band on the gel (*see Note 8*). Save the image of the UV-illuminated gel.
4. Stain the protein with Coomassie blue. Compare the Coomassie blue-stained gel with the image of the UV-illuminated gel to check whether the fluorescent band corresponds to that of the target labeled protein.

3.2.2 Estimation of Degree of Labeling

1. Assess the protein concentration (in mg/mL) by Bradford assay or measurement of absorbance at 280 nm.
2. Calculate the moles of the labeled protein by dividing the protein concentration by the molecular weight of the labeled protein (*see Note 9*).
3. Measure the absorbance of the labeled protein at 495 nm (A_{495}), which is the absorption maximum wavelength of fluorescein-5-maleimide.
4. Calculate the moles of the incorporated fluorescein by dividing the A_{495} value of the labeled protein by $68,000 \text{ M}^{-1} \text{ cm}^{-1}$, which is the molar extinction coefficient of fluorescein-5-maleimide.
5. Determine the degree of labeling by dividing the moles of the incorporated dye by the moles of the labeled protein.

4 Notes

1. PyMOL is a popularly used user-sponsored system accessible at <https://pymol.org/2/> [17] whereas Swiss-Pdb Viewer is a free protein visualization software available at <https://spdbv.vital-it.ch> [18]. In addition, there is a collection of other softwares for protein structure visualization suggested in the Protein Data Bank's website: https://www.rcsb.org/pages/thirdparty/molecular_graphics [15].
2. Endogenous cysteines in the enzyme have to be mutated into other noncharged small-sized amino acids (e.g., alanine or glycine or serine) to ensure that there is only one unique cysteine for the thiol modification. This can avoid the nonspecific coupling of the thiol-reactive dye to the undesired sites of the enzyme. However, if the endogenous cysteines have involved in the disulfide bridge formation and/or deeply embedded inside the protein, they can be retained as the thiol reactive dye is inaccessible to them for the coupling reaction.

3. 10–100 mM HEPES, phosphate or Tris buffer with pH 7.0–7.5 are suitable buffers for the labeling reaction. Reducing agents such as dithiothreitol (DTT) and β -mercaptoethanol should not be added to the protein solution because they can interfere with the thiol coupling reaction [19].
4. For fluorescein-5-maleimide, the optimum pH for the thiol coupling reaction is 7.0. At this neutral pH, maleimide is highly reactive with the thiol groups. Higher pH (>8.0) should be avoided because this leads to the undesired nonselective coupling reaction of maleimide with the amine groups [19, 20].
5. Buffers used in the subsequent downstream experiment are appropriate for the buffer-exchange. Extensive dialysis, desalting chromatography, or gel filtration can also be employed for removing excess dye, but these methods dilute the concentration of the sample. In these cases, the labeled sample has to be concentrated after these processes.
6. To have an efficient buffer exchange, concentrate the sample to a smaller volume in each spin by centrifugation.
7. For a long-term and better storage, the labeled protein is suggested to be frozen by snap freezing with liquid nitrogen and stored at -80°C or lyophilized and stored at -20°C or -80°C .
8. The gel should be illuminated immediately after running the SDS-PAGE electrophoresis. Labeled protein does not show a fluorescent band under UV illumination if the gel has been stained by Coomassie blue.
9. The estimated molecular mass of the labeled protein is the sum of the molecular mass of the unlabeled protein and that of the fluorescent dye. The calculated mass of the unlabeled protein can be determined by inputting the protein sequence to the protein molecular weight calculator (e.g., ProtParam tool which is available at <https://web.expasy.org/protparam/> [21]).

References

1. Robinson PK (2015) Enzymes: principles and biotechnological applications. *Essays Biochem* 59:1–41
2. Vellard M (2003) The enzyme as drug: application of enzymes as pharmaceuticals. *Curr Opin Biotechnol* 14:444–450
3. Fernandes P (2010) Enzymes in food processing: a condensed overview on strategies for better biocatalysts. *Enzyme Res* 2010:862537
4. Adrio JL, Demain AL (2014) Microbial enzymes: tools for biotechnological processes. *Biomolecules* 4:117–139
5. Choi JM, Han SS, Kim HS (2015) Industrial applications of enzyme biocatalysis: current status and future aspects. *Biotechnol Adv* 33:1443–1454
6. Phizicky EM, Fields S (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* 59:94–123

7. Vedadi M, Arrowsmith CH, Allali-Hassani A, Senisterra G, Wasney GA (2010) Biophysical characterization of recombinant proteins: a key to higher structural genomics success. *J Struct Biol* 172:107–119
8. Leake MC (2016) Biophysics: tools and techniques. CRC Press, Taylor and Francis Group, Boca Raton
9. Zhou M, Li Q, Wang R (2016) Current experimental methods for characterizing protein-protein interactions. *ChemMedChem* 11:738–756
10. Chan PH, Liu HB, Chen YW, Chan KC, Tsang CW, Leung YC, Wong KY (2004) Rational design of a novel fluorescent biosensor for a class A β -lactamase. *J Am Chem Soc* 126:4074–4075
11. Tsang MW, Chan PH, So PK, Ma DL, Tsang CW, Wong KY, Leung YC (2011) Engineered AmpC β -lactamase as a fluorescent screening tool for class C β -lactamase inhibitors. *Anal Chem* 83:1996–2004
12. Cheong WL, Tsang MS, So PK, Chung WH, Leung YC, Chan PH (2014) Fluorescent TEM-1 β -lactamase with wild-type activity as a rapid drug sensor for in vitro drug screening. *Biosci Rep* 34:e00136
13. Tsang MW, So PK, Liu SY, Tsang CW, Chan PH, Wong KY, Leung YC (2015) Catalytically impaired fluorescent class C β -lactamase enables rapid and sensitive cephalosporin detection by stabilizing fluorescence signals: implications for biosensor design. *Biotechnol J* 10:126–135
14. Tsang MW, Chan PH, Liu SY, Wong KY, Leung YC (2016) A fluorescein-labeled AmpC β -lactamase allows rapid characterization of β -lactamase inhibitors by real-time fluorescence monitoring of the β -lactamase-inhibitor interactions. *Biotechnol J* 11:257–265
15. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242
16. PDB ID: 1XX2. Knox JR, Sun T. Refinement of P99 β -lactamase from *Enterobacter cloacae*
17. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC
18. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18:2714–2723
19. Nanda JS, Lorsch JR (2014) Labeling of a protein with fluorophores using maleimide derivitization. *Methods Enzymol* 536:79–86
20. Hermanson GT (2013) Bioconjugation techniques, 3rd edn. Academic Press, London
21. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy Server. In: Walker JM (ed) *The proteomics protocols handbook*. Humana Press, Totowa, NJ, pp 571–607



Screening and Production of Recombinant Human Proteins: Ligation-Independent Cloning

Claire Strain-Damerell, Pravin Mahajan, Alejandra Fernandez-Cid, Opher Gileadi, and Nicola A. Burgess-Brown

Abstract

Structural genomics groups have identified the need to generate multiple truncated versions of each target to improve their success in producing a well-expressed, soluble, and stable protein and one that crystallizes and diffracts to a sufficient resolution for structural determination. At the Structural Genomics Consortium, we opted for the ligation-independent cloning (LIC) method which provides the throughput we desire to produce and screen many proteins in a parallel process. Here, we describe our LIC protocol for generating constructs in 96-well format and provide a choice of vectors suitable for expressing proteins in both *E. coli* and the baculovirus expression vector system (BEVS).

Key words PCR, Gene, Ligation-independent cloning (LIC), Construct, Protein, Crystallography

1 Introduction

The knowledge base resulting from sequencing of the human genome has provided a strong foundation for identifying and understanding the role of genes encoding various proteins involved in health and disease as well as in physiological processes. Determining three-dimensional (3D) structures of the proteins is important to understand the biochemical reactions they catalyze at the molecular level. According to the latest estimate by the International Human Genome Sequencing Consortium, the human genome seems to encode 20,000–25,000 proteins [1]. However, there is a major gap between the number of protein sequences and experimentally determined 3D protein structures. The Structural Genomics Consortium (SGC) is a not-for-profit organization that is addressing this gap by solving the structures of medically relevant proteins and placing them into the public domain without restriction (<http://www.thesgc.org/>).

Determining protein structures by X-ray crystallography or cryo-electron microscopy (cryo-EM) on the genome scale creates a number of bottlenecks, the first being expression and purification of the large number of soluble, homogeneous and stable proteins in heterologous systems. We have developed robust protocols for medium-throughput cloning, expression testing and protein production in *E. coli* and in insect cells which have resulted in a portfolio of hundreds of protein domains. We have used *E. coli* as the primary expression system for producing our soluble target proteins; however, for expression of more challenging proteins such as kinases and integral membrane proteins (IMPs), the baculovirus expression system is our first choice. The recombinant proteins expressed globally at SGC have yielded more than 2,000 protein structures, but in addition, these proteins have provided a rich resource for functional genomics, small molecule inhibitor screens, and generation of antibodies.

Ligation-independent cloning [2] was our method of choice as it provided a simple and cost-effective tool for producing many constructs of a single target or multiple targets in parallel without the need to select specific restriction enzymes for each gene. Briefly, the process involves T4 DNA polymerase treatment of linearized vectors in the presence of a single deoxynucleotide (dNTP). PCR fragments of the gene of interest (GOI) with complementary overhangs are generated by adding appropriate 5' extensions into the primers (LIC sequences) and treating the fragments with T4 DNA polymerase in the presence of the paired dNTP (*see* Fig. 1). At the SGC we have engineered many of our vectors to share the same LIC site which allows one LIC-prepared PCR fragment to be cloned into a range of vectors within the same and across different expression systems. Alternative efficient cloning methods are available including Golden Gate Assembly [3, 4], Gateway[®] [5–7], MAGIC [8], and In-Fusion[®] [9], the latter being the method preferred by our SGC node in Toronto. More recently, the LIC method has evolved to SLIC [10, 11] which removes sequence constraints.

In this chapter, we begin the process of medium-throughput screening by describing in detail our methods for (a) identifying domain boundaries to increase the likelihood of producing a stable and correctly folded protein, (b) primer design, PCR and vector preparation, (c) annealing and transformation into *E. coli*, and (d) confirmation of cloning success by colony PCR screening. In the following three chapters, we provide detailed protocols for expression testing using *E. coli*, baculovirus/insect cells and BacMam and producing milligram quantities of protein of sufficient quality and purity for structural studies (e.g., crystallization or cryo-EM) and functional screening. Although our cloning and expression testing protocols are described for 96-well format, the whole process can easily be applied to generate and screen a smaller

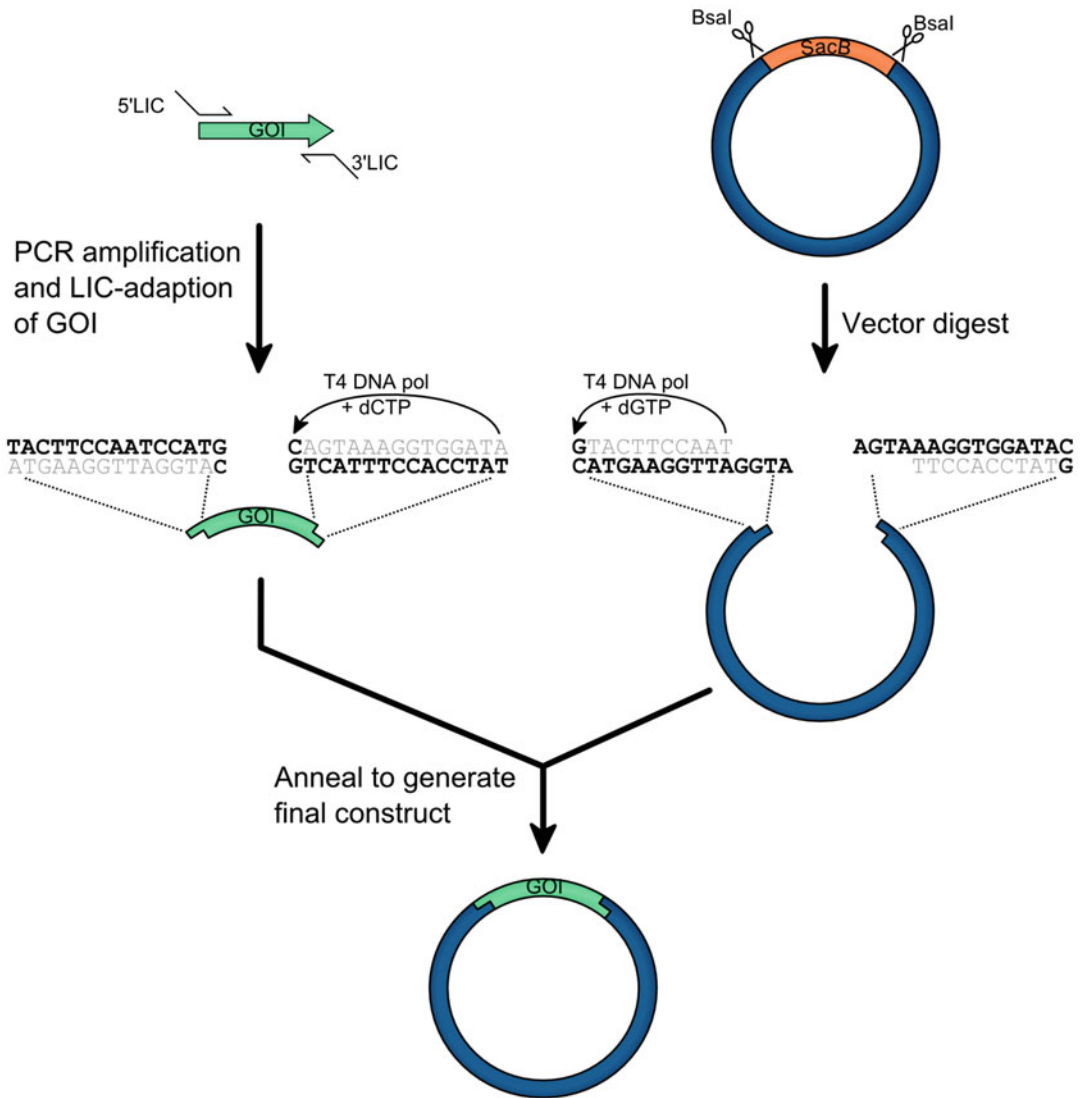


Fig. 1 Overview of the LIC process. The gene of interest (GOI) is amplified with primers that include the LIC sequence specific to the target vector. The vector is linearized by restriction digest, removing the *sacB* gene. Both insert and vector are then T4 DNA polymerase treated to resect 3' ends, creating large overhangs, promoting efficient circularization without the need for T4 DNA ligase

number of proteins. Handling 24 or more samples should be performed in block format rather than in individual tubes as described in the methods.

2 Materials

Unless otherwise stated, molecular biology grade water is used for all dilutions and reactions set out below. Where ultrapure water is

instead specified, it is prepared by purifying deionized water to reach a resistivity of 18.2 M Ω cm at 25 °C. All reagents should be of analytical grade or higher and all plasticware should be DNase-free.

2.1 PCR

1. Primers: Primers are HPSF-purified at 0.01 or 0.05 μ mol scale. Primer stocks are either supplied at or diluted (in 10 mM Tris-HCl buffer, pH 8.0) to 100 μ M and stored at -20 °C.
2. Template library: Human cDNA clones were obtained from the IMAGE cDNA collection (currently distributed by Source BioScience, UK), from other commercial providers (OriGene, Invitrogen, FivePrime), or isolated in-house by PCR from human cDNA. Synthetic DNA clones, including either the natural cDNA or codon-optimized sequences, were synthesized to order by GenScript, BioBasic, or Twist Bioscience.
3. Enzymes: Herculase II Fusion DNA Polymerase (Agilent Technologies), alternatively Q5[®] High-Fidelity DNA Polymerase (New England BioLabs, NEB) for difficult-to-amplify targets; MyTaq[™] Red DNA Polymerase (5 unit/ μ L, Bioline) for colony PCR screening; DpnI (20 units/ μ L).
4. Dimethyl sulfoxide (DMSO), molecular biology grade (DNase/RNase free).
5. 10 mM dNTP solution: Mix 2.5 mM dATP, 2.5 mM dTTP, 2.5 mM dGTP, and 2.5 mM dCTP and store at -20 °C.
6. TE Buffer: 10 mM Tris-HCl and 1 mM EDTA, pH 8.0, filter through a 0.2 μ m syringe filter and store at room temperature (RT).
7. 50 \times TAE buffer (1 L): Dissolve 242 g of Tris base, 57.1 mL of glacial acetic acid, and 100 mL of 0.5 M EDTA, pH 8.0 in water and adjust pH to 8.5. Filter through a 0.2 μ m membrane filter and use as a 1 \times solution.
8. 96-Well 1.5% TAE-agarose gels: Dissolve 3 g of agarose powder in 200 mL of 1 \times TAE buffer using a microwave. Once cooled to hand-hot, add 8 μ L of SYBR-safe DNA gel stain (Invitrogen), mix by swirling and cast in a Sub-cell Model 96 (Bio-Rad or similar) gel cast.
9. DNA ladders: For the E-Gel[®] system, the Low Range Quantitative DNA Ladder (Invitrogen), and for the colony PCR screen, the 1 kb Plus DNA Ladder (Invitrogen) prepared in 1 \times BlueJuice[™] (Invitrogen) are used.
10. PureLink[®] PCR Purification Kit (Invitrogen).
11. MultiScreen PCR₉₆ filter plate (Merck).
12. 96-well PCR plates.
13. Adhesive PCR seals.

14. Adhesive tape pads.
15. V-bottomed microtiter plates.
16. Minisart syringe filters, 0.2 μm .
17. Express™ PLUS filter unit, 0.22 μm (Merck).
18. Supor® PES Membrane Disc Filters, 0.2 μm and unit (Pall).
19. Reagent reservoirs for multichannel pipetting.
20. Multichannel pipettes and repeat pipettors are used to dispense reagents into a 96-well format.
21. 96-well PCR thermocycler with heated lid.
22. E-Gel® 96 Mother base and E-Gel® 96 1% Agarose Gels (Invitrogen).
23. 96-well gel cast and tank (Subcell Model 96 Bio-Rad or similar).
24. Centrifuge suitable for 96-well PCR plates (150 x g).
25. Microcentrifuge.
26. MultiScreen_{HTS} Vacuum Manifold (Merck).
27. Gel Doc™ XR+ (Bio-Rad).
28. Water bath set at 37 °C and 42 °C.

2.2 Cloning

The following reagents, consumables, and equipment are required in addition to those listed above:

1. Competent cells: All cloning is performed in Mach1™ cells (originally purchased from Invitrogen), with chemically competent cells produced in-house using the RbCl method [12]. Alternatively, for large plasmids, NEB® 10-beta cells are used. Other cell lines are suitable for cloning, but we recommend using a *recA*⁻ phage resistant strain, to promote plasmid stability and to reduce the risk of bacteriophage infection during *E. coli* expression, respectively.
2. Vectors: LIC-adapted vectors listed in Table 1 for expression of proteins in Chapters 4, 5, and 6 were generated in-house from commercially available sources. The BacMam vector backbone (pHTBV1.1) was kindly provided by Professor Frederick Boyce (Massachusetts General Hospital, Cambridge, MA) (*see Note 1*).
3. All enzymes and their associated buffers are supplied by NEB; including T4 DNA Polymerase (3 units/ μL), BsaI-HF[®]v2 (20 units/ μL), BfuAI (5 units/ μL), and BseRI (4 units/ μL).
4. 25 mM dGTP: Prepare from 100 mM dNTP set and store at -20 °C.
5. 25 mM dCTP: Prepare from 100 mM dNTP set and store at -20 °C.

Table 1
LIC-adapted vectors for bacterial, baculovirus, and mammalian expression which are available from the SGC on request

Vector name	Antibiotic resistance marker	Tags for purification	Protease site	Restriction site for LIC	dNTP for vector	dNTP for insert	5' LIC primer extension	3' LIC primer extension	Screening primers	bp added during PCR
<i>Bacterial expression vectors</i>										
pNIC28-Bsa4	Kanamycin	N-terminal His ₆	TEV	BsaI	dGTP	dCTP	TACTTCCAA TCCAIG	TATCCACC TTTACTG TCA	pLIC-F+R	~200
pGTVL2	Kanamycin	N-terminal His ₆ +GST	TEV	BsaI	dGTP	dCTP	TACTTCCAA TCCAIG	TATCCACC TTTACTG TCA	pLIC-F+R	~890
pNH-TrxT	Kanamycin	N-terminal His ₆ +Trx	TEV	BsaI	dGTP	dCTP	TACTTCCAA TCCAIG	TATCCACC TTTACTG TCA	pLIC-F+R	~540
pNIC-CTH0	Kanamycin	C-terminal His ₆	TEV	BfuAI	dCTP	dGTP	TTAAGAAGGAGA TATACTAIG	GATTGGAAG TAGAGG TTCTCTGC	pLIC-F+R	~250
pNIC-CTHF	Kanamycin	C-terminal His ₆ +Flag	TEV	BfuAI	dCTP	dGTP	TTAAGAAGGAGA TATACTAIG	GATTGGAAG TAGAGG TTCTCTGC	pLIC-F+R	~230
pNIC-CTI0HF	Kanamycin	C-terminal His ₁₀ +Flag	TEV	BfuAI	dCTP	dGTP	TTAAGAAGGAGA TATACTAIG	GATTGGAAG TAGAGG TTCTCTGC	pLIC-F+R	~240
<i>Baculovirus transfer vectors</i>										
pFB-LIC-Bse	Carbenicillin	N-terminal His ₆	TEV	BseRI	dGTP	dCTP	TACTTCCAA TCCAIG	TATCCACC TTTACTG TCA	FBac-1+2	~290

pFB-HGT-LIC	Carbenicillin	N-terminal His ₆ +GST	TEV	BseRI	dGTP	dCTP	TACTTCCAA <u>TCCAATG</u>	TATCCACC TTTACTG TCA	GST-fwd +FBac-2	~320
pFB-CT6H-LIC	Carbenicillin	C-terminal His ₆	TEV	BfuAI	dCTP	dGTP	TTAAGAAGGAGA TATACTAATG	GATTGGAAG TAGAGG TTCTCTGC	FBac-1+2	~200
pFB-CT6HF-LIC	Carbenicillin	C-terminal His ₆ +Flag	TEV	BfuAI	dCTP	dGTP	TTAAGAAGGAGA TATACTAATG	GATTGGAAG TAGAGG TTCTCTGC	FBac-1+2	~225
pFB-CT10HF-LIC	Carbenicillin	C-terminal His ₁₀ +Flag	TEV	BfuAI	dCTP	dGTP	TTAAGAAGGAGA TATACTAATG	GATTGGAAG TAGAGG TTCTCTGC	FBac-1+2	~240
<i>BacMam vectors</i>										
pHTBV1.1-LIC	Carbenicillin	C-terminal His ₁₀ +Flag	TEV	BfuAI	dCTP	dGTP	TTAAGAAGGAGA TATACTAATG	GATTGGAAG TAGAGG TTCTCTGC	pFBM-fwd +rev	~647
pHTBV1.1-NTSIII-10H-GFP	Carbenicillin	N-terminal Twin-Strep +His ₁₀	TEV	BfuAI	dGTP	dCTP	TACTTCCAA <u>TCCAATG</u>	TATCCACC TTTACTG TCA	pFBM-fwd +rev	~740
pHTBV1.1-NTSIII-10H-GFP	Carbenicillin	N-terminal Twin-Strep +His ₁₀ +GFP	TEV	BfuAI	dGTP	dCTP	TACTTCCAA <u>TCCAATG</u>	TATCCACC TTTACTG TCA	GFP-fwd +pFBM-rev	~400
pHTBV1.1-NTGST-10H-LIC	Carbenicillin	N-terminal GST+His ₁₀	TEV	BfuAI	dGTP	dCTP	TACTTCCAA <u>TCCAATG</u>	TATCCACC TTTACTG TCA	GST-fwd +pFBM-rev	~530
pHTBV1.1-CT10H-SIII-LIC	Carbenicillin	C-terminal His ₁₀ +Twin-Strep	TEV	BfuAI	dCTP	dGTP	TTAAGAAGGAGA TATACTAATG	GATTGGAAG TAGAGG TTCTCTGC	pFBM-fwd +rev	~720

(continued)

Table 1
(continued)

Vector name	Antibiotic resistance marker	Tags for purification	Protease site	Restriction site for LIC	dNTP for vector	dNTP for insert	5' LIC primer extension	3' LIC primer extension	Screening primers	bp added during PCR
pHTBV1.1-CTGFP-SIIL-10H	Carbenicillin	C-terminal GFP +Twin-Strep +His ₁₀	TEV	BfuAI	dCTP	dGTP	TTAAGAAGGAGA <u>TATACTATG</u>	GATTGGAAG TAGAGG TTCTCTGC	pFBM-fwd +GFP- rev	~590
pHTBV1.1-C3C-SIIL-10H	Carbenicillin	C-terminal Twin-Strep +His ₁₀	3C	BfuAI	dGTP	dCTP	TACTTCCAA <u>TCCATG</u>	AAACAACACC TCCAG	pFBM-fwd +rev	~740
pHTBV1.1-C3CGFP-SIIL-10H	Carbenicillin	C-terminal GFP +Twin-Strep +His ₁₀	3C	BfuAI	dGTP	dCTP	TACTTCCAA <u>TCCATG</u>	AAACAACACC TCCAG	GFP-fwd +pFBM- rev	~560

The antibiotic resistance cassette and purification tags are indicated, along with the details required for LIC: forward and reverse primer extension required to LIC-adapt the GOI, restriction enzyme to cut the vector with, and the dNTP required for the T4-treatment step of either vector or PCR product. Note that the start codon included by the 5' LIC sequence is underlined and the stop codon included by the 3' LIC sequence is italicized

6. 100 mM and 1 M DTT: Make up with water, filter through a 0.20 μm syringe filter and store as 1 mL aliquots at $-20\text{ }^{\circ}\text{C}$.
7. 20 mg/mL bovine serum albumin (BSA).
8. 25% (w/v) sucrose: Dissolve 250 g sucrose in 1 L of ultrapure water and filter through a 0.22 μm filter unit.
9. 60% (v/v) glycerol: Autoclave to sterilize.
10. 50 mg/mL Carbenicillin: Prepare in water, filter through a 0.20 μm syringe filter and store at $-20\text{ }^{\circ}\text{C}$.
11. 50 mg/mL Kanamycin: Prepare in water, filter through a 0.20 μm syringe filter and store at $-20\text{ }^{\circ}\text{C}$.
12. LB-agar: Dissolve 22.5 g of premixed LB-broth and 13.5 g of agar in 800 mL of ultrapure water. Adjust volume to 900 mL and autoclave on the same day.
13. LB-agar plates: Melt LB-agar slowly in a microwave and add 5% (w/v) sucrose (*see Note 2*). Once cooled to hand-hot, add the appropriate antibiotic (*see Table 1*) and swirl vigorously to mix. Pour 10 mL of the molten agar into each 50 mm petri dish and once set, upturn and leave open to dry. These can be prepared ahead of time and stored for up to a month at $4\text{ }^{\circ}\text{C}$, sealed in a plastic bag to prevent over-drying.
14. $1\times$ LB: Dissolve 22.5 g of premixed LB-broth in 800 mL of ultrapure water. Adjust volume to 900 mL and autoclave on the same day.
15. SOC medium: Dissolve 18 g of tryptone (or peptone from casein), 4.5 g of yeast extract, 0.45 g of NaCl, and 2.25 mL of 1 M KCl in 800 mL of ultrapure water. Adjust volume to 900 mL and autoclave on the same day. Once cooled, add 9 mL of 2 M MgCl_2 hexahydrate and 18 mL of 1 M (18%) glucose. Both solutions are filtered through a 0.20 μm syringe filter prior to use (*see Note 3*).
16. Virkon.
17. Montage Plasmid Miniprep_{HTS} 96 Kit (Merck).
18. 50 mm petri dishes.
19. Disposable sterile spreaders or 2 mm autoclaved glass balls (2.5–3.5 mm) for spreading as these are reusable and allow for faster plating for the medium-throughput scale.
20. Disposable sterile inoculation loops (1 μL).
21. 96-deep-well blocks.
22. AirOtop porous seals (Thomson or VWR).
23. Centrifuge suitable for 96-deep-well blocks ($3,000\times g$).

24. Micro-Express Glas-Col shaker (Glas-Col, Indiana, USA) or similar set to 37 °C.
25. Incubator set at 37 °C.
26. Heated block set at 50 °C.

3 Methods

3.1 Construct Design

In order to give the best possible chance of producing soluble protein with a high propensity for crystallization we opt for a multiconstruct design approach [13–15]. While we do include the full-length protein in the initial target screen, only 8.6% of our solved structures have arisen from such constructs. By repositioning the start and stop boundaries of our constructs by only 5 amino acids either side, our success increases to 13.3% (unpublished data). By expanding the design out to include only certain domains of the protein, our success rate improves further meaning that structures that would have otherwise been missed make it through to Protein Data Bank (PDB) submission using the multiconstruct approach. Constructs are therefore designed based on available protein domain information, secondary structure predictions and sequence alignments, as well as taking account of disordered regions to try to produce more stable proteins at the expression stage. Due to uncertainty in predictive methods and in our understanding of factors affecting protein behavior, we test a number of construct endpoints (2–5 on either end) closely spaced around the predicted domain boundaries.

3.2 Primer and Plate Design

1. Having identified appropriate construct boundaries in the previous step, design primers for PCR amplification of the desired DNA segments. The primer sequences themselves typically include the appropriate LIC sequence (*see* Table 1) followed by ~20 bp from the construct sequence. In each case, the ATG underlined in Table 1 should be in-frame with the target sequence.
2. Where the construct includes an N-terminal purification tag, the stop codon is incorporated by the 3' LIC sequence marked in italics (*see* Table 1).
3. For C-terminally tagged constructs, the reverse primer must not include a stop codon but must be in-frame with the 3' LIC sequence i.e., do not include additional nucleotides between the 3' of the reverse LIC site and the codon encoding the C-terminal amino acid (*see* Note 4).
4. The arrangement of constructs in a 96-well format is done with the following constraints for ease of cloning: (a) constructs from the same entry clone are kept together; (b) constructs

are arranged in order of size; (c) where possible, only one vector and or T4-treatment condition is used per plate; and (d) if the plate is mixed then like-vectors and T4-treatment conditions are kept together on the plate. Arrangement in this manner enables easy identification of correctly sized products and limits mistakes caused by erroneous pipetting.

5. Once you have designed the plate format keep a record of what primers, template, and vector will be associated with each well and use this for all subsequent steps.

3.3 PCR

1. Using a multichannel pipette and reagent reservoir, add 90 μL of water to each well of a 96-well PCR plate. To this, add 5 μL each of the 100 μM forward and reverse primers and mix well.
2. For each template, prepare a 2.5 ng/ μL dilution in a 1.5 mL Eppendorf tube, mix well and aliquot 20 μL of this into the appropriate wells of a second 96-well PCR plate.
3. Prepare a PCR master mix as follows: 500 μL of 5 \times Herculase II buffer, 75 μL of 10 mM dNTP mixture, 25 μL of Herculase II Fusion DNA Polymerase, and 1.5 mL of water. Mix the solution well. Using a multichannel pipette or repeat pipettor, aliquot 21 μL into each well of a third 96-well PCR plate (*see Note 5*).
4. Using a multichannel pipette, transfer 1.5 μL of the diluted primers, followed by 2.5 μL of diluted template DNA, into the corresponding wells of the reaction plate. Mix well then seal the plate using an adhesive PCR seal, making sure to press down well in order to limit evaporation (*see Note 6*).
5. Place the reaction plate into the thermocycler and cycle with the following conditions—touchdown PCR (*see Note 7*):
 - 95 $^{\circ}\text{C}$, 10 min.
 - (95 $^{\circ}\text{C}$, 30 s; 68 $^{\circ}\text{C}$, 30 s; 68 $^{\circ}\text{C}$, 1–3 min*) \times 5 cycles.
 - (95 $^{\circ}\text{C}$, 30 s; 60 $^{\circ}\text{C}$, 30 s; 68 $^{\circ}\text{C}$, 1–3 min*) \times 5 cycles.
 - (95 $^{\circ}\text{C}$, 30 s; 55 $^{\circ}\text{C}$, 30 s; 68 $^{\circ}\text{C}$, 1–3 min*) \times 5 cycles.
 - (95 $^{\circ}\text{C}$, 30 s; 50 $^{\circ}\text{C}$, 30 s; 68 $^{\circ}\text{C}$, 1–3 min*) \times 20 cycles.
 - 68 $^{\circ}\text{C}$, 10 min.
 - 15 $^{\circ}\text{C}$ hold.

*Extension time dependent on length of PCR product (e.g., 30 s per 1 kb).
6. Remove 3 μL of each reaction and dilute with 12 μL of water. Run on an E-Gel[®] against 20 μL of the diluted Low Range Quantitative DNA Ladder (Fig. 2).
7. Transfer the successful reactions into the corresponding wells of a fresh PCR plate and repeat any failed reactions using

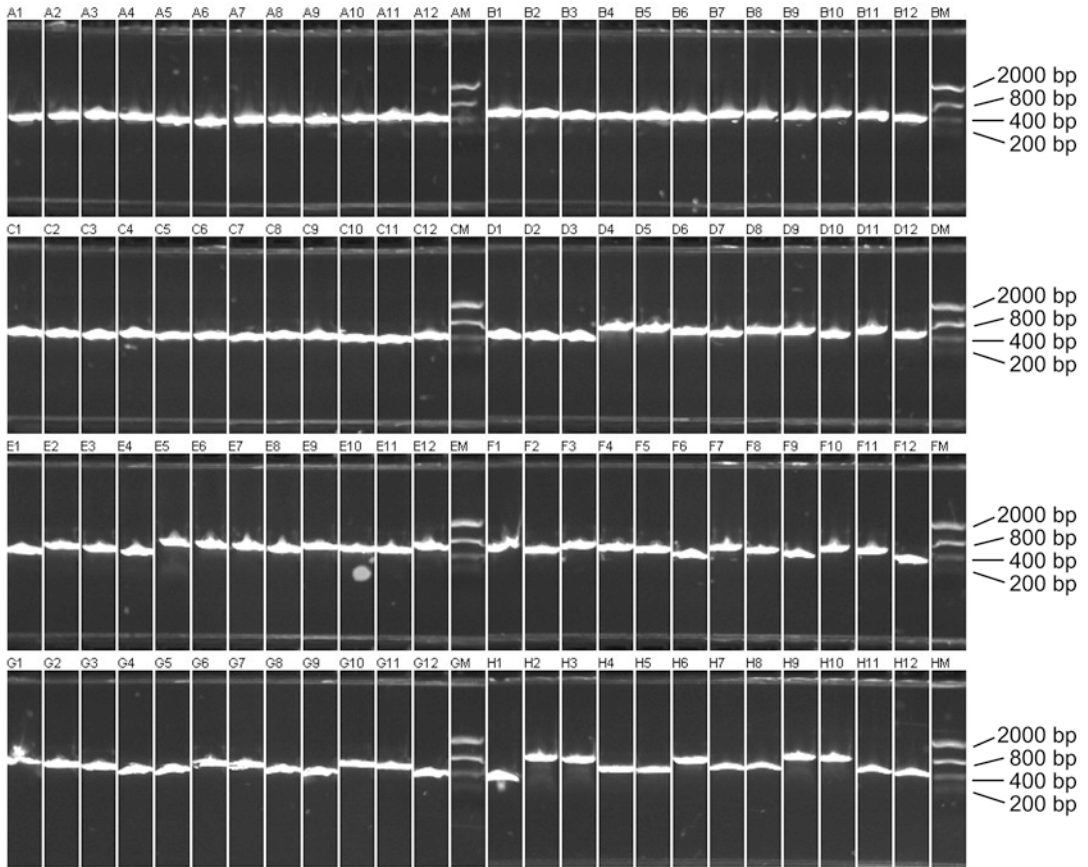


Fig. 2 Image of an initial PCR performed in 96-well format, analyzed using the E-Gel[®] system and Low Range Quantitative DNA Ladder. The sizes of the ladder are indicated. Due to the low resolution of these gels the products are judged based on the sizing of neighboring bands; for example, the products of E5 to E8 should be in decreasing size order, which can be observed on the gel

different cycling conditions or with additives such as the DMSO or GC-enhancer (*see Note 8*).

8. Any products amplified from templates containing the same antibiotic resistance cassette as the target vector, require DpnI treatment to limit template carryover (*see Note 9*). Prepare a 1 in 20 dilution of DpnI (20 units/ μ L) in NEB buffer 2 or CutSmart and aliquot 1 μ L into the appropriate wells of the PCR reaction plate. Incubate the plate in a 37 °C incubator for 1 h.
9. Purify the products (*see Note 10*) using a MultiScreen PCR₉₆ purification plate following the manufacturer's instructions. Recover the DNA from the plate in 50 μ L of TE buffer, transferring into a V-bottomed microtiter plate and store at -20 °C.

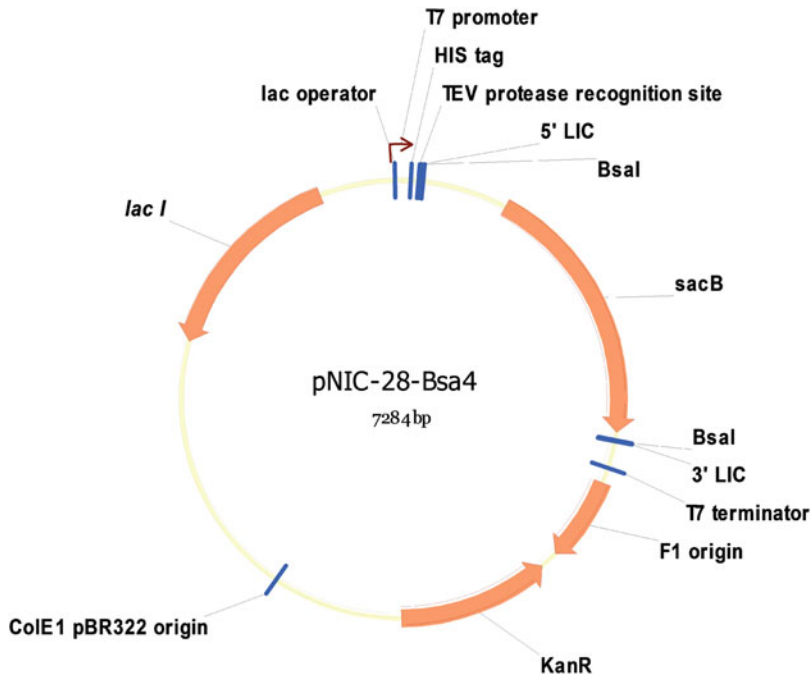


Fig. 3 Vector map of standard bacterial expression vector pNIC28-Bsa4. Digestion with BsaI excises the *sacB* gene and T4-treatment resects the 3' ends of the LIC sites to provide complementary cohesive ends to the PCR products. The vector incorporates a His₆ tag at the N-terminus followed by TEV cleavage site in frame with the PCR product. This vector also includes the T7 promoter and terminator sequences for expression in the BL21(DE3) strain and is under the control of the lac repressor for induction with IPTG during the expression stage (see Chapter 2)

3.4 Vector Preparation

1. Digest the target vector using the restriction enzyme indicated in Table 1 (see Note 11 for alternative restriction enzymes), for example for BsaI vectors (see Fig. 3 for example vector) prepare the digest as follows: 5 µg vector, 10 µL of 10× NEB CutSmart, 1.5 µL of BsaI-HF[®]v2 (20 units/µL), make up to 100 µL with water and incubate at 37 °C for 2 h.
2. Mix 3 µL of the digested vector with 3 µL of 2× BlueJuice[™] and analyze on a 1.5% TAE-agarose gel to confirm complete digestion (see Note 12).
3. Purify the digested vector using a PureLink[®] purification spin column, following the manufacturer's instructions, and elute in 50 µL.

3.5 T4 DNA Polymerase Treatment

1. To the purified vector (50 µL) add 21.5 µL of water, 10 µL of 10× NEB buffer 2.1, 10 µL of 25 mM dCTP or dGTP (see Table 1), 5 µL of 100 mM DTT, 1 µL of 2.5 mg/mL BSA, and 2.5 µL of T4 DNA Polymerase (3 units/µL). Place in a thermocycler with the following conditions: 22 °C for 30 min, 75 °C for 20 min, 15 °C hold (see Note 13).

2. For T4-treatment of the PCR products prepare a master mix as follows: 215 μL of water, 100 μL of 10 \times NEB buffer 2.1, 100 μL of 25 mM dCTP or dGTP (*see* Table 1), 50 μL of 100 mM DTT, 10 μL of 20 mg/mL BSA, and 25 μL of T4 DNA Polymerase (3 units/ μL , NEB). Using a repeat pipettor aliquot 5 μL into each well of a PCR plate. Using a multichannel pipette, transfer 5 μL of the purified PCR product into the corresponding wells of the T4 reaction mix, mixing as you dispense. Place in a thermocycler with the following conditions: 22 $^{\circ}\text{C}$ for 30 min, 75 $^{\circ}\text{C}$ for 20 min, 15 $^{\circ}\text{C}$ hold.

3.6 Annealing and Transformation

1. Using a repeat pipettor, aliquot 1 μL of the T4-treated vector into each well of a 96-well PCR plate and centrifuge briefly at 150 $\times g$. Confirm that there is liquid in each well before progressing to **step 2**.
2. Using a multichannel pipette, transfer 2 μL of T4-treated insert into the corresponding wells of the plate from **step 1** (*see* Note 14). Spin briefly and incubate the reaction at RT for at least 20 min before placing on ice (*see* Note 15).
3. Take two 1.5 mL Eppendorf tubes, label one with “vector-only control” and the other with “insert-only control” (*see* Note 16). To the first add 1 μL of the T4-treated vector and to the other 2 μL of T4-treated insert from a well that has undergone DpnI treatment, then place both tubes on ice.
4. Using a repeat pipettor, aliquot 50 μL of chemically competent sub-cloning efficiency cells (*see* Notes 17 and 18) into each well of the plate from **steps 1** and **2** and into the two tubes from **step 3**. Incubate on ice for 30 min.
5. Heat-shock the cells at 42 $^{\circ}\text{C}$ for 45 s, then return to ice briefly.
6. Using a multichannel, pipette 100 μL of SOC medium (*see* Notes 19 and 20) into each well, seal with a porous seal and incubate at 37 $^{\circ}\text{C}$ for 1.5 h in a stationary incubator.
7. Plate 100 μL of the transformation mixture onto LB-agar plates containing 5% sucrose (*see* Note 1) supplemented with either 50 $\mu\text{g}/\text{mL}$ kanamycin or carbenicillin (*see* Table 1). Spread the sample across the plate using either sterile spreaders or glass beads (*see* Note 21).
8. Incubate the plates at 37 $^{\circ}\text{C}$ for ~16 h, then store at 4 $^{\circ}\text{C}$ until the colony PCR screening step is complete.

3.7 Colony PCR Screening

1. Prepare a 96-deep-well block containing 1 mL of LB and the appropriate antibiotic selection (*see* Table 1).
2. Set up a PCR master mix as follows: 400 μL of 5 \times MyTaqTM Reaction Buffer Red, 1.49 mL of water, 100 μL of 10 μM screening primers (*see* Tables 1 and 2), and 10 μL of MyTaqTM DNA Polymerase (5 unit/ μL). Using a repeat

Table 2
Colony PCR screening primers for SGC vectors

Primer name	Primer sequence
pLIC-F	TGTGAGCGGATAACAATTCC
pLIC-R	AGCAGCCAACTCAGCTTCC
FBac-1	TATTCATACCGTCCCACCA
FBac-2	GGGAGGTTTTTTTTAAAGCAAGTAAA
FBac-3	TTAAAATGATAACCATCTCG
pFBM-fwd	CAAAATGTCGTAACAACCTCCGC
pFBM-rev	TAGTTAAGAATACCAGTCAATCTTTCAC
GFP-fwd	TAACCACTACTTGTCGACGCAGTC
GFP-rev	CTGTCGTACAGATGAACCTTCAAGGTC
GST-fwd	CAATGTGCCTGGATGCGTTCC

The screening primers are situated upstream of the LIC sites, allowing full sequencing of the purification tags incorporated by the vector sequence. The pLIC primers are for the bacterial vectors, the FBac primers are for the baculovirus vectors (BEVs) and the pFBM primers are for the BacMam vectors. Note that FBac-1 and -2 may be used to screen all BEVs but that FBac-1 is located too close to the start codon of the C-terminally tagged vectors to allow complete coverage during sequencing; FBac-3 is recommended for this purpose. The GFP and GST primers are used instead to reduce the size of the PCR fragment

pipettor or a multichannel pipette, aliquot 20 μ L into each well of a 96-well PCR plate.

- Using a 1 μ L sterile loop, pick one colony from each transformation plate and inoculate into the corresponding well of the PCR reaction plate (**step 2**) followed by the corresponding well of the deep well block (**step 1**) (*see Note 22*).
- Once all of the wells have been inoculated, seal the deep well block with a porous seal and incubate at 37 °C overnight in a Glas-Col with shaking at 700 rpm, then store at 4 °C.
- Seal the PCR reaction plate with a thermal resistant adhesive seal and set a thermocycler with the following conditions, making sure that the block is up to temperature before placing your sample plate in the instrument (*see Note 23*):
 - 95 °C, 10 min
 - (95 °C, 30 s; 50 °C, 30 s; 72 °C, 1–3 min*) \times 25 cycles
 - 72 °C, 5 min
 - 15 °C hold

*Extension time dependent on length of PCR product—e.g., 30 s per 1 kb. Please note that additional bp will be added to your products due to the positioning of the screening primers (*see Table 1*).
- While the cycle is running, prepare a 96-well 1.5% TAE-agarose gel.

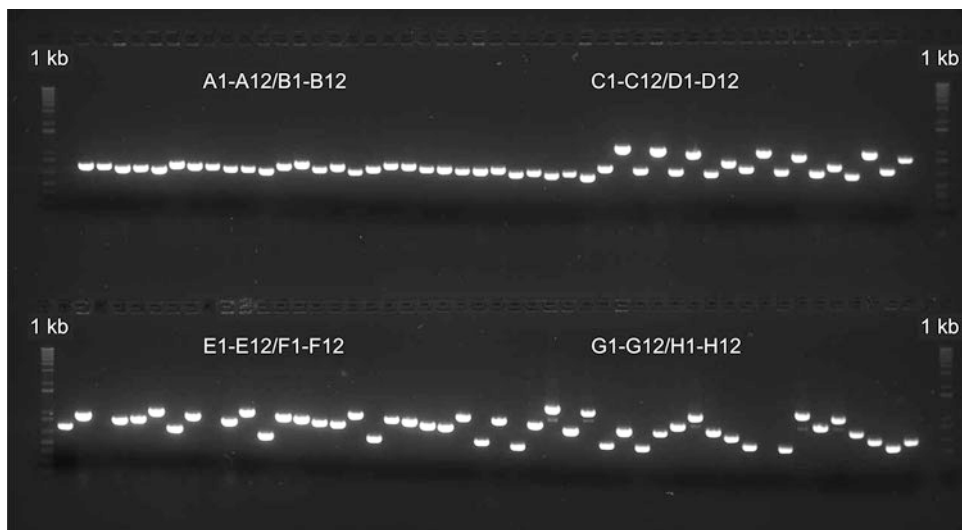


Fig. 4 Image of a colony PCR screen performed in a 96-well format, analyzed on a 1.5% TAE agarose gel. The samples are interleaved (e.g., A1, B1, A2, B2, etc.). Note that the products are larger (~200 bp) at the colony screening stage due to the positioning of the screening primers

7. Using a multichannel pipette, load 10 μL of the PCR reaction mixtures directly onto the gel. Note that the spacing of the wells means that samples will be interleaved (Fig. 4). Load 6 μL of 1 kb DNA Ladder and run the gel at 150 V for 1 h.
8. Confirm the sizing of the products and repeat the screen for additional clones if necessary (*see Note 24*).

3.8 Preparation of Glycerol Stocks and 96-Well Miniprep

1. Combine the correct clones into a single block by inoculating 20 μL of each culture into 1 mL of LB (*see Note 25*) in a new 96-deep-well block, containing the same antibiotic selection as above. Grow overnight at 37 $^{\circ}\text{C}$ in a Glas-Col with shaking at 700 rpm.
2. To each well of a V-bottomed microtiter plate, add 30 μL of 60% (v/v) glycerol, followed by 120 μL of the overnight culture. Mix well as you add the culture (*see Note 26*). Seal with an aluminum foil pad and store at -80°C .
3. Centrifuge the remaining culture at $3,000 \times g$ for 20 min.
4. Discard the supernatant into a waste pot containing 1% Virkon and blot the excess liquid onto a clean paper towel.
5. Use a 96-well plasmid purification kit to purify the plasmids from these cell pellets following the manufacturer's instructions, with a few modifications (*see Note 27*).
6. Recover the DNA in 50 μL of TE and transfer into a V-bottomed microtiter plate. Seal with an adhesive tape pad and store at -20°C .

4 Notes

1. The vectors listed in Table 1 are available from Addgene or Source BioScience. However, if you would like to receive the pHTBV1.1 vectors, please contact the SGC (contact@sgc.ox.ac.uk).
2. The *sacB* gene product, expressed from our LIC-adapted vectors (Fig. 3), is capable of converting sucrose to a toxic by-product [16]. By adding sucrose to the LB-agar plates we select for recombinant plasmids only, as these will lack the *sacB* gene, having been replaced by our GOI (Fig. 1).
3. It is advisable to prepare small volumes of SOC medium at a time as it is prone to contamination.
4. As the primer sequences are dictated by the desired boundaries in the protein sequence, the corresponding DNA sequences may have properties (e.g., repetitions or biased nucleotide composition) that make it difficult to design optimal primers. Primers are thus designed with care to avoid mispriming or primer-dimers and to ensure compatible T_m values, determining the lengths and base composition accordingly.
5. Note that if using the repeat pipettor to aliquot the PCR master mix, the volume added will actually be 20 μL (not 21 μL) but this will not affect the reaction.
6. Spend plenty of time sealing your PCR plate, applying a lot of pressure around the wells to ensure efficient adherence to prevent evaporation. It is important that your thermal cycler has a heated lid as this will again limit the amount of evaporation.
7. When dealing with a mixture of targets and primers on one 96-well plate, it is not always possible to optimize each reaction, therefore the best approach is to perform touchdown PCR as a first pass and then use a more tailored cycle for any missing products. As touchdown cycles through a range of annealing temperatures, it will cover the differences in melting temperatures of your primers across the plate.
8. If you get multiple bands from your PCR, try using a fixed annealing temperature instead which should be $\sim 5^\circ\text{C}$ lower than the melting temperature of your primers. If you get no bands, try using additives such as the GC-enhancer supplied with Q5® High-Fidelity Polymerase or DMSO at a final concentration of 3%, or test higher concentrations of MgSO_4 (1.5–3 mM). However, you may also want to sequence your template to check that it is correct.
9. DpnI is a restriction endonuclease that can only cleave at its recognition sites when they have been methylated. Standard

strains of *E. coli* (including Mach1™) methylate their DNA, thus any entry clones propagated in them will be methylated. By DpnI-treating a PCR product, we specifically cleave the template DNA leaving only the product intact. This limits the chance of template carry-over when the entry clone carries the same antibiotic resistance marker as the cloning vector.

10. It is important to purify the PCR products away from any unincorporated dNTPs in the reaction mixture as these will inhibit resection of the 3' ends during the T4 DNA polymerase step. Note that when DMSO is present in the PCR reaction, the clean-up of the samples is delayed.
11. Alternative restriction digest conditions are as follows: BfuAI vectors: 5 µg vector, 5 µL 10× NEB buffer 3.1, 1 µL BfuAI (5 units/µL), make up to 50 µL with water and incubate at 50 °C for 2 h. Alternatively, if BfuAI digestion is problematic or inefficient, you can try using BveI: 5 µg vector, 10 µL 10× FastDigest, 2.5 µL 20× Oligonucleotide (0.01 mM), 2 µL FastDigest BveI (5 units/µL), made up to 100 µL with water and incubated at 37 °C for 2 h. BseRI vectors: 5 µg vector, 10 µL 10× NEB buffer 2, 1.5 µL BseRI (4 units/µL), made up to 100 µL with water and incubated at 37 °C for 2 h.
12. Check by agarose gel analysis that your vector has two clearly distinct bands; the top one is the vector backbone that you will ligate your fragment into and the lower band is the *sacB* fragment (~2 kb). You do not need to purify the lower fragment away from the top fragment as self-ligation is selected against by using sucrose in the medium (*see Note 1*).
13. It is important that you only add one dNTP to your reaction as this will determine the stop position of the 3' resection (Fig. 1). For this reason, it is also important that your dNTP stock is stored at -20 °C when not in use to ensure that it remains fresh. The same rule applies to DTT.
14. The longer you give the annealing the more successful your transformation will be. Give your samples no less than 20 min but give them longer time whenever possible.
15. If the annealing reaction is not successful and transformation fails, repeat using 1.5 µL of the T4-treated vector and 3.5 µL of the T4-treated insert, and/or increase the volume of competent cells to 100 µL. We recommend using these alternative conditions with BacMam vectors, since we noticed a decrease in annealing efficiency when using them.
16. It is important to include a vector-only control during the transformation to check that the rate of insert-independent colonies is low. The sucrose will select against reinsertion of the *sacB* fragment and uncut vector but the vector backbone can occasionally close on itself. If there are many colonies on

this plate then there may be an issue with your sucrose selection or with your T4-treatment step as self-ligation should be rare. Note that these will be distinguishable at the PCR screen step as they will produce a ~200-bp product. You should also include an insert-only control at the transformation step when your PCR products have required DpnI treatment as this will indicate any template carryover from insufficient DpnI treatment.

17. When using a repeat pipettor to aliquot your cells, care should be taken to prevent cross-contamination between wells caused by splash-back.
18. It is important to use high-cloning efficiency cells for the transformation and if you fail to get colonies this is normally the reason why. If you prepare your cells in-house, then check the efficiency is on the order of 1×10^6 CFU per μg by transforming 0.5 ng of vector. This test should be done every time new competent cells are prepared. You should also test for contamination by plating 50 μL of untransformed cells on plates containing either carbenicillin or kanamycin. This test should be performed using aseptic techniques to ensure that the cells are the only potential source of contamination.
19. Other media can be used during this step (e.g., $1 \times$ or $2 \times$ LB); however, SOC gives a higher transformation efficiency when dealing with the low DNA concentrations that are used in this protocol.
20. We recommend using 2-mL blocks and 500 μL of SOC media when working with BacMam vectors. Blocks should be incubated in a shaker (Glas-col or similar) at 700 rpm and 37 °C for 1.5–2 h.
21. To plate using sterile glass beads: stack the plates, agar-side down, in order of row (e.g., A1 to A12) and add ~5 beads per plate. Working from one side of the transformation plate to the other transfer 100 μL of the culture to the relevant agar plate. When each row is completed, split the stack into two blocks of six and shake the plates from side to side to spread the culture. Once all wells have been plated, shake the plates once more and upturn to move the beads onto the lid. The beads can then be transferred into a beaker containing 1% (w/v) Virkon to be cleaned, autoclaved, and reused.
22. Give the inoculation loop a twirl in both the PCR mixture and the LB to transfer more material for the PCR and growth, respectively.
23. We have found MyTaq™ Red DNA Polymerase reactions to be more successful when the samples are placed in a thermocycler preheated to 95 °C, rather than allowing the enzyme to heat up to 95 °C. If using an alternative screening polymerase, check

the conditions specified by the manufacturer, however note that Bioline do not specify preheating with their product.

24. If your colony screen is not working there may be several reasons why: If you get a smear on your agarose gels then it can often be remedied by cleaning your pipettes and gel tank before starting the screen. If you get no product, then check that your reagents and cycling conditions are working by including a small sample of your uncut vector (use 2 μL of a 2.5 ng/ μL dilution for a 20 μL reaction) to act as a positive control. If this works but your screen does not, then there may be an issue with your cells (*see* **Note 15**). If the positive control fails, then you may want to try alternative reagents and/or cycling conditions and if the initial PCR requires specific conditions, then try these for the screen as well. If you are using ampicillin (as opposed to carbenicillin) as the selectable marker, we have found that colonies with lots of satellites surrounding them tend not to yield products during the PCR screen. If this is the case, try retransforming and always store the plates at 4 °C when you are not screening them.
25. For the plasmid miniprep, we have found that any medium richer than LB yields pellets too large for efficient clearing during the miniprep process.
26. It is important to mix your cells when preparing glycerol stocks to ensure the viability of stock—should you need to go back to it.
27. The volume of each buffer used to isolate the plasmid DNA is 100 μL instead of 150 μL which is recommended in the manufacturer's instruction booklet. In addition, we assemble our clearing plate above the manifold, with the plasmid plate inside the manifold, and apply ~300 mbar pressure. This is contrary to the manufacturer's instructions due to risk of cross-contamination; however, we find this to be more effective and have had no issue with samples missing wells when this level of pressure is applied.

Acknowledgments

We would like to thank all the SGC scientists (past and present) who contributed toward the development of the method. The SGC is a registered charity (number 1097737) that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genome Canada, Innovative Medicines Initiative (EU/EFPIA), Janssen, Merck KGaA, MSD, Novartis Pharma AG, Ontario Ministry of Economic Development and Innovation, Pfizer, São Paulo

Research Foundation-FAPESP, Takeda, and Wellcome. The BacMam vector backbone (pHTBV1.1) was kindly provided by Professor Frederick Boyce (Massachusetts General Hospital, Cambridge, MA).

References

1. Zea A (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431 (7011):931–945
2. Haun RS, Serventi IM, Moss J (1992) Rapid, reliable ligation-independent cloning of PCR products using modified plasmid vectors. *Bio-techniques* 13(4):515–518
3. Sanjana NE, Cong L, Zhou Y et al (2012) A transcription activator-like effector toolbox for genome engineering. *Nat Protoc* 7 (1):171–192
4. Potapov V, Ong JL, Kucera RB et al (2018) Comprehensive profiling of four base overhang ligation fidelity by T4 DNA ligase and application to DNA assembly. *ACS Synth Biol* 7 (11):2665–2674
5. Hartley JL, Temple GF, Brasch MA (2000) DNA cloning using in vitro site-specific recombination. *Genome Res* 10(11):1788–1795
6. Invitrogen (2010) Gateway[®] Technology: a universal technology to clone DNA sequences for functional analysis and expression in multiple systems. Invitrogen Life Technologies, Carlsbad
7. Walhout AJ, Temple GF, Brasch MA et al (2000) GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol* 328:575–592
8. Li MZ, Elledge SJ (2005) MAGIC, an *in vivo* genetic method for the rapid construction of recombinant DNA molecules. *Nat Genet* 37 (3):311–319
9. Clontech (2012) In-Fusion[®] HD cloning kit user manual. Clontech, Mountain View
10. Li MZ, Elledge SJ (2007) Harnessing homologous recombination *in vitro* to generate recombinant DNA via SLIC. *Nat Methods* 4 (3):251–256
11. Li MZ, Elledge SJ (2012) SLIC: a method for sequence- and ligation-independent cloning. *Methods Mol Biol* 852:51–59
12. Hanahan D, Jessee J, Bloom FR (1991) Plasmid transformation of *Escherichia coli* and other bacteria. *Methods Enzymol* 204:63–113
13. Graslund S, Sagemark J, Berglund H et al (2008) The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. *Protein Expr Purif* 58(2):210–221
14. Bray JE (2012) Target selection for structural genomics based on combining fold recognition and crystallisation prediction methods: application to the human proteome. *J Struct Funct Genomics* 13(1):37–46
15. Savitsky P, Bray J, Cooper CD et al (2010) High-throughput production of human proteins for crystallization: the SGC experience. *J Struct Biol* 172(1):3–13
16. Gay P, Le Coq D, Steinmetz M et al (1983) Cloning structural gene *sacB*, which codes for exoenzyme levansucrase of *Bacillus subtilis*: expression of the gene in *Escherichia coli*. *J Bacteriol* 153(3):1424–1431



Screening and Production of Recombinant Human Proteins: Protein Production in *E. coli*

Nicola A. Burgess-Brown, Pravin Mahajan, Claire Strain-Damerell, Alejandra Fernandez-Cid, Opher Gileadi, and Susanne Gräslund

Abstract

In Chapter 3, we described the Structural Genomics Consortium (SGC) process for generating multiple constructs of truncated versions of each protein using LIC. In this chapter we provide a step-by-step procedure of our *E. coli* system for test expressing intracellular (soluble) proteins in a 96-well format that enables us to identify which proteins or truncated versions are expressed in a soluble and stable form suitable for structural studies. In addition, we detail the process for scaling up cultures for large-scale protein purification. This level of production is required to obtain sufficient quantities (i.e., milligram amounts) of protein for further characterization and/or structural studies (e.g., crystallization or cryo-EM experiments). Our standard process is purification by immobilized metal affinity chromatography (IMAC) using nickel resin followed by size exclusion chromatography (SEC), with additional procedures arising from the complexity of the protein itself.

Key words *E. coli*, Bacteria, Expression, Recombinant, Protein, Purification, Immobilized metal affinity chromatography (IMAC), Size exclusion chromatography (SEC), Gel filtration

1 Introduction

Choosing from which expression system to produce your protein can depend on many different factors such as its size, location within the cell, and the requirement for posttranslational modifications (PTMs) [1]. To provide a starting point for researchers, structural genomics groups collectively identified trends and common strategies for producing proteins for structural determination [2]. At the SGC, we preferentially start with *E. coli* for testing and producing human intracellular (soluble) proteins, specifically a tRNA-enhanced strain of BL21(DE3) which often compensates for codon bias [3, 4]. This low cost prokaryotic expression system is easy to use, suitable for increasing throughput and has a high success rate for many targets, particularly when truncated or mutated versions of the protein are screened [5, 6]. In 2010, we

showed that 48% of the human proteins attempted in *E. coli* were successfully purified, and of those, the structures of approximately 40% were solved by X-ray crystallography [7]. Protein crystallization demands availability of soluble, pure, monodisperse, and homogeneous proteins in sufficient quantities (usually in milligram quantities). Nevertheless, the limited amount of protein obtained from initial small-scale expression testing can provide valuable information on protein solubility, expression level, molecular weight and PTMs of target proteins. In addition to our standard histidine (his)-tagged vectors, we have engineered a number of other vectors harboring different tags and/or fusion partners (some of which are listed in Chapter 3, Table 1) and a variety of *E. coli* host strains [7]. All of these vectors also contain a six or ten his-tag enabling the use of IMAC purification for fast and efficient capture of recombinant proteins from cell lysates.

A version of the bacterial methods from expression testing to large-scale protein production was published previously [8]. The method presented here has been modified, in particular, the changes in the method used to test protein expression in small scale (1 mL) cultures has provided better correlation with the results of large-scale expression. We found that using n-Dodecyl beta-D-maltoside (DDM) to lyse the bacterial membranes gave hits most comparable to those from large-scale cultures lysed by sonication or homogenization. The previous method we employed, extracting the protein with BugBuster[®], produced many false negative results (unpublished data) and often required purification from a 50 mL culture to distinguish the true positives. Since we implemented this change in procedure, our false negative rate has declined substantially. Although we screen for expression in a 96-well format, the methods do not require expensive or specialized equipment and are easily adaptable to lower throughput in individual tubes and flasks. As a consequence, they can be performed in any lab, with minimal equipment, at whatever scale is required.

The methods for large-scale protein expression and purification are also described in this chapter to provide the researcher with a complete process for obtaining quality protein in quantities sufficient for crystallization and/or cryo-EM experiments or developing assays for functional screening. The generic methods described here are routinely used in our laboratory for expression and purification of a large number of proteins. Following the standard IMAC purification, many highly expressed proteins only require one additional step of SEC to yield pure protein, but for difficult-to-purify proteins, additional steps such as his-tag cleavage using TEV protease and rebinding to nickel resin or ion exchange chromatography are often required. Moreover, occasionally variations in the methodology are incorporated to address the need arising from complexity of the proteins, by introducing changes such as buffer type,



Fig. 1 Overview of the bacterial expression pipeline. The process takes ~3–4 weeks from LIC to scale-up

pH, ionic strength, and use of additives to the buffer in order to stabilize the proteins. The pipeline from cloning to expression testing through to large-scale protein expression and purification is outlined in Fig. 1. The processes that we use for screening and producing proteins in the baculovirus expression vector system (BEVS) and mammalian (BacMam) system are described in the subsequent chapters.

2 Materials

Unless otherwise stated, all solutions are prepared using ultrapure water (prepared by purifying deionized water to reach a resistivity of 18 M Ω cm at 25 °C) and analytical grade reagents.

2.1 Transformation and Test Expression

1. BL21(DE3)-R3-pRARE2 *E. coli* strain: phage-resistant derivative of BL21(DE3) isolated in-house containing the pRARE2 plasmid which was extracted from the strain Rosetta2 from Novagen. This strain supplies tRNAs for 7 rare codons (AGA, AGG, AUA, CUA, GGA, CCC, and CGG) on a compatible chloramphenicol-resistant plasmid. Chemically competent bacterial cells are prepared in-house as described [9].
2. 60% (v/v) glycerol: Autoclave to sterilize.
3. 50 mg/mL kanamycin: Prepare in water, filter through a 0.20 μ m syringe filter, and store at –20 °C.
4. 34 mg/mL chloramphenicol: Prepare in ethanol and store at –20 °C.

5. 1 M IPTG: Prepare in water, filter through a 0.20 μm syringe filter, and store at $-20\text{ }^{\circ}\text{C}$.
6. LB agar: Dissolve 22.5 g of premixed LB broth and 13.5 g agar in 800 mL of water. Adjust volume to 900 mL and autoclave on the same day.
7. LB agar plates: Melt LB agar slowly in a microwave and add 5% (w/v) sucrose. Once cooled to hand-hot, add the appropriate antibiotic and swirl vigorously to mix. Pour 10 mL of the molten agar into each 50 mm petri dish and once set, upturn and leave open to dry. These can be prepared ahead of time and stored for up to a month at $4\text{ }^{\circ}\text{C}$, sealed in a plastic bag to prevent overdrying.
8. $1\times$ LB: Dissolve 22.5 g of premixed LB broth in 800 mL of water. Adjust volume to 900 mL and autoclave on the same day.
9. $2\times$ LB: Dissolve 45 g of premixed LB broth in 800 mL of water. Adjust volume to 900 mL and autoclave on the same day.
10. SOC medium: Dissolve 18 g of tryptone (or peptone from casein), 4.5 g of yeast extract, 0.45 g of NaCl, and 2.25 mL of 1 M KCl in 800 mL of water. Adjust volume to 900 mL and autoclave on the same day. Once cooled, add 9 mL of 2 M MgCl_2 hexahydrate and 18 mL of 1 M (18%) glucose. Filter both solutions through a 0.20 μm syringe filter prior to use (*see Note 1*).
11. TB medium: Dissolve 12 g of Bacto tryptone, 24 g of yeast extract, and 4 mL of glycerol in 800 mL of water. Adjust volume to 900 mL and autoclave on the same day. Once cooled to room temperature (RT), adjust volume to 1 L with 100 mL of a separately autoclaved solution of 0.17 M KH_2PO_4 and 0.72 M K_2HPO_4 .
12. Virkon tablets.
13. 24-well cell culture plates.
14. 96-well PCR plates.
15. 96-well microtiter plates.
16. 96-deep-well blocks.
17. Disposable sterile spreaders or glass balls (2.5–3.5 mm; VWR 33212 4G).
18. Disposable sterile inoculation loops, 1 μL .
19. AirOtop porous seals (Thomson or VWR).
20. Adhesive tape pads.
21. Adhesive foil for microplates.
22. Disposable cuvettes.

23. Reagent reservoir for multichannel pipetting.
24. Multichannel pipettes and repeat pipettors are used to dispense reagents into a 96-well format.
25. Micro-Express Glas-Col shaker (Glas-Col, Indiana, USA) or alternative that ranges in temperature from 18 °C to 37 °C and shakes up to 800 rpm.
26. Water bath set at 42 °C.
27. Incubator set at 37 °C.
28. 96-well block mixer (Eppendorf MixMate or similar).
29. A visible light spectrophotometer for measuring OD_{600nm} (optical density) of bacterial cultures (for individual cuvettes).
30. 96-well plate reader is also useful but not essential.

2.2 Test Purification

The following reagents, consumables, and equipment are required in addition to those listed above:

1. Benzonase (Novagen, HC, 250 units/μL).
2. Protease Inhibitor Cocktail Set VII (Calbiochem).
3. 10 mg/mL Lysozyme: Prepared freshly in water.
4. 10% (w/v) DDM (n-Dodecyl beta-D-maltoside), Sol-grade (Anatrace or Glykon): Prepare in water, filter through a 0.20 μm syringe filter, and store at -20 °C.
5. 0.5 M Tris(2-carboxyethyl)phosphine (TCEP): Prepare in water, filter through a 0.20 μm syringe filter, and store at -20 °C.
6. 1 M dithiothreitol (DTT): Prepare in water, filter through a 0.20 μm syringe filter, and store in 1 mL aliquots at -20 °C.
7. SeeBlue[®] Plus2 Pre-Stained Standard (Invitrogen).
8. InstantBlue[™] (Expedeon Protein Solutions).
9. 20× NuPAGE[™] MES SDS Running Buffer (Invitrogen).
10. 1 M HEPES, pH 7.5: Prepare in water, filter through a 0.2 μm membrane filter, and store at RT.
11. 5 M NaCl: Prepare in water, filter through a 0.2 μm membrane filter, and store at RT.
12. 3 M imidazole, pH 8.0: Prepare in water, filter through a 0.2 μm membrane filter, and store at RT.
13. 200 mM MgSO₄: Prepare in water, filter through a 0.2 μm membrane filter, and store at RT.
14. 50% (v/v) glycerol: Autoclave and store at RT.
15. Lysis buffer (1 L): 100 mM HEPES, pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, and 10 mM imidazole prepared in advance, filtered through a 0.2 μm membrane filter, and stored at 4 °C.

On the day of purification, add 50 $\mu\text{L}/\text{mL}$ lysozyme, 0.2 $\mu\text{L}/\text{mL}$ Benzonase, 1 $\mu\text{L}/\text{mL}$ protease inhibitor cocktail, 10 $\mu\text{L}/\text{mL}$ DDM, 5 $\mu\text{L}/\text{mL}$ MgSO_4 , and 1 $\mu\text{L}/\text{mL}$ TCEP from stock solutions (*see Note 2*).

16. Wash buffer (1 L): 20 mM HEPES, pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, and 25 mM imidazole prepared in advance, filtered through a 0.2 μm membrane filter, and stored at 4 °C. Add 0.5 mM TCEP on the day of purification.
17. Elution buffer (0.1 L): 20 mM HEPES, pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, and 500 mM imidazole prepared in advance, filtered through a 0.2 μm membrane filter, and stored at 4 °C. Add 0.5 mM TCEP on the day of purification.
18. Affinity buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, 10% glycerol, and 10 mM imidazole prepared in advance, filtered through a 0.2 μm membrane filter, and stored at 4 °C.
19. 50% (w/v) Ni-IDA Metal Chelate Resin (Genxon) or Ni-NTA-agarose (Qiagen): The IMAC resins are generally supplied in 20% ethanol. To equilibrate, wash the resin twice in water and then three times in Affinity buffer in a 50 mL tube, by inverting to resuspend the resin and centrifuging at $500 \times g$ for 1 min. After the final wash, resuspend the resin in Affinity buffer as 50% (w/v) slurry and store at 4 °C when not in use.
20. SB: Prepare a stock of NuPAGE LDS sample buffer (Invitrogen) containing DTT (1:4 dilution of 1 M DTT in NuPAGE LDS sample buffer) and store at -20 °C.
21. MultiScreen[®] Filter Plates, 1.2 μm (Merck).
22. MultiScreen_{HTS} Vacuum Manifold (Merck).
23. Precast 26-Lane SDS-PAGE gradient gels (4–12% Bis-Tris) (Invitrogen).
24. Protein gel electrophoresis apparatus (Invitrogen).
25. 96-well thermocycler with heated lid.
26. All gels are imaged on a Gel Doc[™] XR+ (Bio-Rad).
27. Centrifuge suitable for 96-deep-well blocks ($3,000 \times g$).

2.3 Large-Scale Expression

The following reagents, consumables, and equipment are required in addition to those listed above:

1. Glycerol stocks of transformed expression strain.
2. 2.5 L Ultra Yield baffled flasks (Thomson) or glass flasks.
3. Shaker-incubators with cooling capacity: Innova 44R (New Brunswick) or Multitron (Infors HT).
4. Avanti J-20XP or Avanti J-26XP centrifuge or similar (Beckman Coulter) with a JLA 8.1000 rotor for harvesting large volumes of cells.

2.4 Protein Extraction and Large-Scale Purification

The following reagents, consumables, and equipment are required in addition to those listed above:

1. 5% (w/v) Polyethyleneimine (PEI): Dilute a 50% solution tenfold then adjust to pH 7.5 with HCl.
2. 2× Lysis buffer: 100 mM HEPES, pH 7.5, 1 M NaCl, 20% (v/v) glycerol, and 20 mM imidazole. Filter through a 0.2 µm membrane filter and store at 4 °C. On the day of purification, add Benzonase (0.2 µL/mL of cell lysate), Protease inhibitor cocktail (2 µL/mL of cell lysate), and 1 mM TCEP.
3. Lysis buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, and 10 mM imidazole. Filter through a 0.2 µm membrane filter and store at 4 °C. On the day of purification, add Benzonase (0.1 µL/mL of cell lysate), Protease inhibitor cocktail (1 µL/mL cell lysate) or cOmplete EDTA-free protease inhibitor cocktail (1 tablet/25 mL cell lysate), and 0.5 mM TCEP.
4. Affinity buffer: 50 mM HEPES buffer, pH 7.5, 500 mM NaCl, 10% glycerol, and 10 mM imidazole. Add 0.5 mM TCEP on the day of purification.
5. Wash buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, 10% glycerol, and 30 mM imidazole. Add 0.5 mM TCEP on the day of purification.
6. Elution buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, 10% glycerol, and 500 mM imidazole. Add 0.5 mM TCEP on the day of purification.
7. Size Exclusion Chromatography (SEC) buffer: 20 mM HEPES, pH 7.5, 500 mM NaCl, and 5% glycerol. Filter through a 0.2 µm membrane filter and store at 4 °C. Add 0.5 mM TCEP on the day of purification.
8. Minisart syringe filters, 0.20 µm, 0.45 µm, and 0.80 µm.
9. Millex[®]-GV Low Protein Binding Filter, 0.22 µm (Merck).
10. Amicon Ultra protein concentrators.
11. Sonicator (Sonics Vibra-Cell, VCX 750, Sonics & Materials INC) or basic Z model cell disruptor (Constant Systems Ltd.) or EmulsiFlex-C5 high-pressure homogenizer (Avestin).
12. Econo-Columns (Bio-Rad or similar).
13. ÄKTA-Xpress or ÄKTA-Pure liquid chromatography system.
14. HiTrap 5 mL FF columns for his-tagged protein purification.
15. Ion exchange chromatography columns such as HiTrap 5 mL Q FF and SP FF.
16. HiLoad Superdex columns (GE) for preparative size exclusion chromatography such as HiLoad 16/600 Superdex[™] S75 pg, S200 pg, or Superose[™] 6 Increase 10/300 GL.

17. UV spectrophotometer for measuring DNA and protein concentration (e.g., The NanoDrop™ spectrophotometer allows for measurements from as low as 1.5 µL volumes).
18. General-purpose benchtop centrifuge.
19. JA-17 rotor for centrifugation of cell lysates.
20. Microcentrifuge:
21. Supor® PES Membrane Disc Filters, 0.2 µm and unit (Pall).

3 Methods

3.1 Transformation into *E. coli* BL21(DE3)-R3-pRARE2

1. Prepare four 24-well tissue culture plates containing 1 mL of LB agar, supplemented with 50 µg/mL kanamycin and 34 µg/mL chloramphenicol and once set allow to dry, inverted at RT.
2. Using a multichannel pipette, add 3 µL of recombinant DNA to a 96-well PCR plate. Place on ice and add 30 µL of chemically competent *E. coli* BL21(DE3)-R3-pRARE2 cells using a repeat pipettor (*see Note 3*). Cover with an adhesive tape pad and incubate for 30 min on ice. It is advisable to include a positive control protein (i.e., a protein that has previously shown soluble expression in *E. coli*) in position H12 of the 96-well plate.
3. Heat-shock in a water bath at 42 °C for 45 s, return to ice briefly then add 100 µL of SOC (or 2× LB) medium (*see Note 4*). Cover with a porous seal and incubate for 1 h at 37 °C.
4. Pipette 30 µL of the transformation mixture onto the agar in the 24-well plates according to the format presented (*see Fig. 2*). Gently rock the plates to cover the surface and allow them to dry before incubating at 37 °C inverted overnight (*see Note 5*).
5. Inoculate three colonies or a streak of colonies from each well (*see Note 6*) into the corresponding well of a 96-deep-well block containing 1 mL of LB (or 2× LB) medium supplemented with 50 µg/mL kanamycin and 34 µg/mL chloramphenicol.
6. Cover the block with porous film and place in the Glas-Col shaker in the afternoon at 37 °C, with shaking at 700 rpm.
7. The following morning, prepare four replicate glycerol stocks. Dispense 30 µL of 60% (v/v) glycerol into 96-well microtiter plates. Transfer 120 µL of each culture into the corresponding wells of the microtiter plates and mix by pipetting. Seal the plate with an aluminum foil seal and store at −80 °C. Keep the remainder of the overnight culture for setting up the test expression.

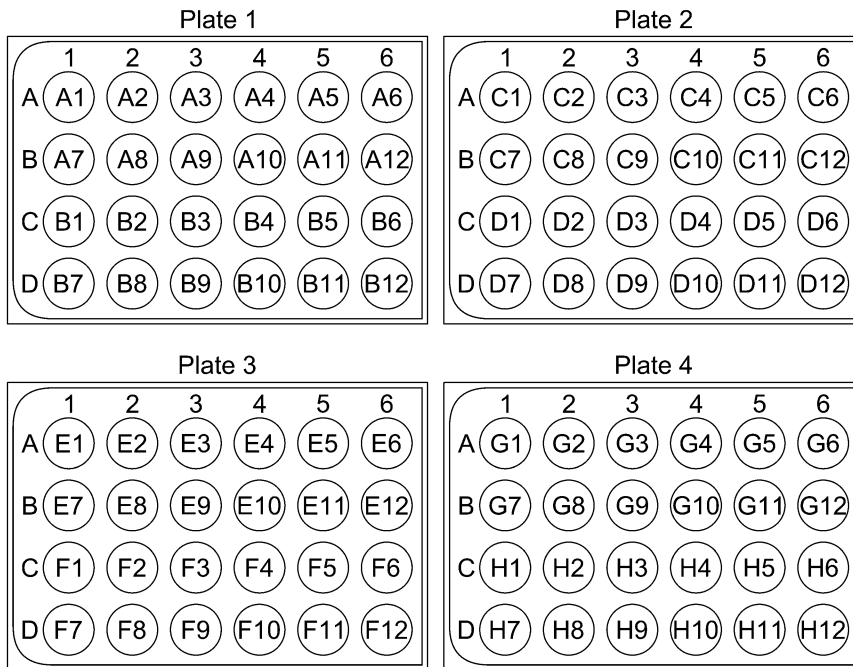


Fig. 2 Format for plating cultures grown in a 96-well block onto four 24-well agar plates. This template can be printed out to scale and placed underneath the 24-well plates when plating

3.2 Test Expression

1. Inoculate 20 μL of the overnight culture (or thawed glycerol stock) into each well of two 96-deep-well blocks containing 1 mL of fresh TB medium, supplemented with 50 $\mu\text{g}/\text{mL}$ kanamycin (*see Note 7*) and grow to an $\text{OD}_{600\text{nm}}$ of 2–3 (approximately 5 h) in a Glas-Col shaker set at 37 $^{\circ}\text{C}$ and 700 rpm. Label one block as “OD measurement block” and the other as “test block.”
2. Determine the OD measurement for a few wells (at least one appearing visually to have low density and one high density) by diluting 1 in 4 in TB medium and using a visible light spectrophotometer. If the $\text{OD}_{600\text{nm}}$ is between 2 and 3, and you have available a 96-well plate reader, dilute aliquots of the test block 1 in 4 (in a total of 200 μL) in a flat-bottomed clear microtiter plate for OD measurement using the plate reader (*see Note 8*).
3. Leave the cultures to cool down at RT for 30 min or place them in the cold room for 15 min and change the temperature setting of the shaker to 18 $^{\circ}\text{C}$.
4. Induce expression by adding 0.1 mM IPTG (10 mM stock prepared in TB medium and 10 μL added to the block) and incubating in the Glas-Col shaker overnight at 18 $^{\circ}\text{C}$ and 700 rpm.

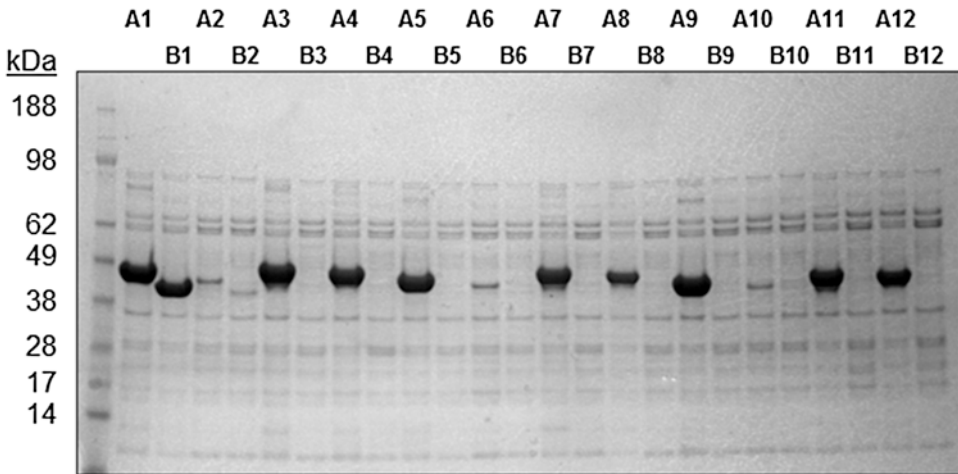


Fig. 3 Image showing the Coomassie SDS-PAGE result of a test purification from *E. coli*. The gel shows a range of high, medium, and low expressing proteins of different molecular weights. Note that samples loaded using a multichannel pipette are interleaved

3.3 Test Purification

1. Centrifuge the 96-deep-well block at $3,000 \times g$ for 20 min, pour off the medium into a waste pot containing 1% Virkon and tap the block on absorbent paper (*see* **Notes 9** and **10**).
2. Add 200 μL of Lysis buffer (*see* **Note 2**) and resuspend the pellets either using the Glas-Col shaker at 18 $^{\circ}\text{C}$ and 700 rpm or a 96-well block mixer (Eppendorf) at 1,000–2,000 rpm. Use a multichannel pipette to resuspend any remaining solid pellets and store the block at -80°C for at least 20 min, until all pellets are completely frozen.
3. Thaw the pellets in a shallow water bath (at RT) for approximately 15 min and resuspend in the Glas-Col shaker for 10 min. Remove 3 μL (Total fraction) and pipette into a PCR plate containing 37 μL of water and 20 μL of SB for storage at 4 $^{\circ}\text{C}$ until required (Total fraction).
4. Centrifuge the block at $3,000 \times g$ for 10 min.
5. Meanwhile, add 50 μL of a previously washed and equilibrated 50% slurry (Ni-IDA or Ni-NTA) to each well of a Multi-Screen® Filter Plate, 1.2 μm .
6. Transfer the clarified lysate (*see* **Note 11**) using a 1 mL capacity multichannel pipette, to the filter plate containing the resin, taking care to avoid transferring any pelleted material (*see* **Note 12**).
7. Place an adhesive tape pad on top and incubate the plate in the Glas-Col shaker at 18 $^{\circ}\text{C}$ for 1 h at 400 rpm (*see* **Note 13**).

8. Assemble the vacuum manifold according to the manufacturer's instructions and then filter the contents through the plate into waste for approximately 20 s, taking care not to dry out the resin (*see* **Note 14**). Turn off the vacuum.
9. Add 200 μL of Wash buffer and filter quickly. Repeat this step three more times, turning the vacuum off after each step to prevent overdrying, and then place the filter plate on top of a waste block and centrifuge for 2 min at $300 \times g$ to remove all trace of the Wash buffer (*see* **Note 15**).
10. Place a fresh 96-well microtiter plate under the filter plate, add 40 μL of Elution buffer and seal the plate with an adhesive tape pad.
11. Incubate the plate in the Glas-Col shaker for 10–20 min at 400 rpm and 18 °C (or at RT on a shaking platform).
12. Elute the protein by centrifugation at $300 \times g$ for 3 min.
13. Store the eluent (Purified fraction) at 4 °C until required (or –20 °C for long term storage).
14. Dispense 5 μL of SB in all wells of a 96-well PCR plate, add 15 μL of each Purified fraction, denature by heating at 80 °C for 10 min and load 15 μL samples into each lane of the SDS-PAGE gels using a multichannel pipette, by alternating rows (e.g., A1, B1, A2, B2, etc.; *see* **Note 16**). Include a protein marker in the first lane (e.g., SeeBlue[®] Plus2 Pre-Stained Standard).
15. Run the gels at 150 V for approximately 1 h or until the first dye front has reached the bottom of the gel, then stain with InstantBlue[™] to identify which constructs are positive for soluble expression (*see* Fig. 3).

3.4 Large-Scale Expression

1. After identifying the positive constructs from the test expression and purification, prepare a starter culture by inoculating a loop of the glycerol stock into 10 mL of TB medium containing 50 $\mu\text{g}/\text{mL}$ kanamycin and 34 $\mu\text{g}/\text{mL}$ chloramphenicol in a 50 mL tube (*see* **Notes 17** and **18**). Grow the starter culture overnight at 37 °C in a shaker-incubator.
2. The next morning, inoculate 10 mL of the starter culture into a 2.5 L Ultra Yield or baffled glass flask containing 1 L of sterile TB medium, freshly supplemented with 50 $\mu\text{g}/\text{mL}$ kanamycin only (*see* **Note 7**). Cover the flask with a porous seal and incubate at 37 °C, with shaking at 200 rpm (*see* **Note 19**).
3. Monitor $\text{OD}_{600\text{nm}}$ by taking 1 mL of the sample every hour and continue the incubation at 37 °C until the $\text{OD}_{600\text{nm}}$ reaches 2.00 ± 1 (*see* **Note 20**).
4. Move the cultures to the cold room and reduce the temperature of the incubator to 18 °C and after approximately 30 min,

induce protein expression by adding IPTG (from a 1 M stock solution) to a final concentration of 0.1 mM (*see Note 21*), then continue the incubation overnight at 18 °C.

5. If needed, measure OD_{600nm} by diluting 25 µL of the culture into 1 mL of the TB medium (*see Note 22*) and harvest the remaining cells by centrifugation at 9,000 × *g* for 20 min using a JLA-8.1000 rotor or similar. Pour the supernatant back into the original culture flask and decontaminate using Virkon.
6. Remove traces of the medium from the cell pellet using a 1 mL pipette and transfer the cell pellet to a 50 mL tube. Record the wet-weight of the cells (generally 12–30 g from 1 L of culture) and store the pellets at –80 °C until required for purification. The cell pellets can be stored at –80 °C for many months (*see Note 23*).

3.5 Protein Extraction

All the following steps of protein extraction and purification are performed at 4 °C or on ice. Prechill all buffers and centrifuges.

1. If protein purification is performed straight after harvesting the cells, transfer the cell pellets to ice or if the cells were frozen, thaw the pellets in a water bath set at 37 °C for as long as required to thaw, then immediately transfer to ice.
2. Resuspend the cells in 1 volume of 2× Lysis buffer (1 mL/g wet-weight) and mix thoroughly using a glass rod or serological pipette. Add 2 to 3 more volumes of 1× Lysis buffer until the sample is manageably fluid with no cell lumps.
3. Prechill the cell disruptor and lyse the cells resuspended in **step 2** above. Refer to the manufacturer's instructions of the instrument that is used (e.g., for the basic Z model cell disruptor, two to three rounds at ~15,000 psi are sufficient for cell lysis). Recover the lysate in the disruptor by flushing it with Lysis buffer (20–40 mL). Save 10 µL of the lysate which represents the Total fraction (*see Note 24*).
4. Add PEI to the cell lysate to a final concentration of 0.15% and mix thoroughly by inverting the tube several times or using a pipette. At this stage the lysates turn milky (*see Note 25*).
5. Transfer the lysates to centrifuge tubes, balance the tubes pairwise, and centrifuge at 39,000 × *g* in a JA-17 rotor (or similar) for at least 30 min at 4 °C.
6. Transfer the clear supernatant into a clean tube taking care to avoid transferring any pelleted material. This clarified supernatant represents the Soluble fraction (*see Note 26*).

3.6 Large-Scale Protein Purification

The purification scheme described here for histidine-tagged proteins is generic and applied to a diverse set of proteins; however, it may not be applicable to every individual protein. Other buffer

compositions may be substituted to address issues such as protein instability and requirements of final applications. Careful optimization of the buffer composition with respect to the buffering system, pH, salt concentrations, and additives is particularly critical for difficult to purify proteins (*see Note 27*). The protein purification scheme described here is a two-step procedure (a) IMAC and (b) SEC. Manual IMAC provides the flexibility to use a specific volume of resin to the amount of lysate and collection of several elutions with gradual increase in imidazole concentration. Automated protein purification systems allow for rapid purification of target proteins while using multiple chromatography steps with minimal intervention. An important point to mention when working with large culture volumes are the problems associated with applying large volumes of lysate to small IMAC columns which can result in reduced protein binding capacity due to depletion of nickel ions from the column [10].

1. To perform manual IMAC, prepare the Ni-IDA (or Ni-NTA) resin as described in Subheading 2.2, item 19.
2. Add the resin to the clarified cell lysate in a 50 mL tube. Depending on the estimated protein expression level, add 0.5–2 mL of the 50% (w/v) resin to the clarified lysate obtained per L of culture and rotate the tubes gently for 30 min to 1 h at 4 °C.
3. Centrifuge at $500 \times g$ for 10 min, remove and save the supernatant in a fresh tube which represents the Unbound fraction, taking care not to lose the resin while removing the supernatant.
4. Resuspend the resin in 2–3 column volumes (CV) of Affinity buffer and transfer to an empty chromatography column (such as an Econo-Column). Alternatively, prepare a proportionate resin bed in an empty column, apply the clarified cell lysate and collect the Unbound fraction by gravity flow through the column.
5. Wash the resin in the column with 10 CV of Affinity buffer and save the flow through for gel analysis.
6. Wash the resin with 20 CV of Wash buffer, again saving the flow through for gel analysis.
7. Elute the bound protein in fractions of at least 5 elutions of 2 CV of Elution buffer, generally a total of 10–15 CV. Analyze 15 μ L of each elution by SDS-PAGE (*see Fig. 4a*) prior to proceeding to the next step (*see Note 28*).
8. To prepare the sample for SEC, pool the fractions and concentrate using an Amicon Ultra protein concentrator according to the manufacturer's instructions. Transfer the concentrated sample into a 50 mL tube and centrifuge at $4,000 \times g$ for

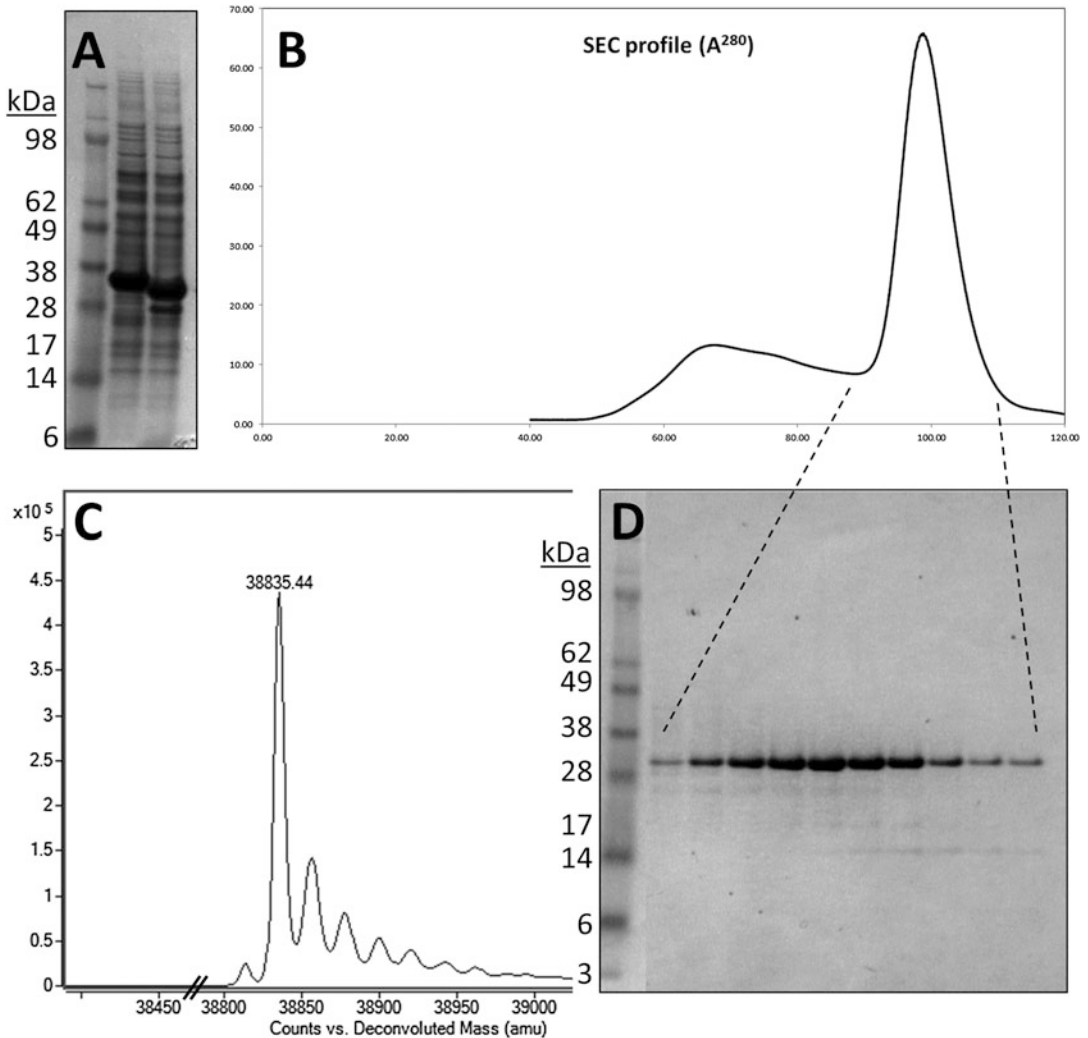


Fig. 4 Image showing quality assurance measures in protein purification. **(a)** The initial IMAC elutions are analyzed by SDS-PAGE to determine approximate size and yield. In the example shown, gel A lane 3 shows the product of TEV-mediated cleavage of the his-tag following IMAC purification. **(b)** The cleaved protein is then purified by SEC which in this example shows some degree of aggregation by the presence of two peaks. **(d)** The resulting fractions from SEC are assessed for purity by SDS-PAGE before pooling and concentrating the protein. **(c)** The identity of the purified protein is then confirmed by intact mass spectrometric (MS) analysis. Note that in the example shown the expected mass of the protein is 38.8 kDa, as confirmed by MS analysis. The size discrepancy shown in inserts A and D is due to the inaccuracy of size determination by SDS-PAGE

10 min or filter through a 0.22 μm low protein binding filter to remove aggregates and particulates before loading onto the SEC column.

9. To perform SEC, follow the method from **step 16** onward; however, the protein sample will have to be injected onto the SEC column manually (*see Note 29*).

10. To perform automated IMAC and SEC using an ÄKTA-Xpress chromatography system, prepare the system in a cold cabinet or cold room in advance by employing the desired number of IMAC columns (e.g., HisTrap FF crude) and a SEC column (e.g., HiLoad 16/60 Superdex 75 prep grade, HiLoad 16/60 Superdex 200 prep grade or Superose 6 Increase 10/300 GL).
11. Prepare the HisTrap columns by washing first with 10 CV of water and then by equilibrating with 10 CV of Affinity buffer at 0.8 mL/min flow rate.
12. Prepare the SEC column by washing first with 2 CV of water (inlet A5) and then with 2 CV of SEC buffer (inlet A4).
13. Set up an IMAC and SEC purification method. Change the default parameters as described in the notes (*see Note 30*). **Steps 14–17** are performed automatically on the ÄKTA-Xpress.
14. Apply the clarified and filtered cell lysate to the preequilibrated IMAC column at 0.8 mL/min flow rate.
15. Wash the IMAC column with 5–10 CV of Affinity buffer using inlet A1 until the A_{280} stabilizes. Wash with 10 CV of Wash buffer using inlet B1. Elute with 5 CV of Elution buffer using inlet A3. The eluted peak is automatically identified by detection of an increased A_{280} and is collected into the reinjection loop.
16. The peak is then automatically injected onto the SEC column at a flow rate of 1.2 mL/min, followed by running 1.2 CV of SEC buffer at the same flow rate using inlet A4.
17. Collect 2 mL fractions based on the A_{280} peak (*see Fig. 4b*) into a 96-deep-well block.
18. Analyze the fractions by SDS-PAGE for purity and homogeneity (*see Fig. 4d*). Avoid high molecular weight aggregates and pool peaks corresponding to different oligomeric forms (e.g., monomers, dimers) separately (*see Note 31*).
19. If the purified protein is to be used at a later date, concentrate the protein using an Amicon Ultra protein concentrator and aliquot in small volumes, flash-freeze in liquid nitrogen and store at $-80\text{ }^{\circ}\text{C}$ until needed. To prevent damage to the protein during freezing and thawing, add glycerol, if not already included in the buffer (*see Note 32*).
20. If required purity is not achieved in this two-step purification scheme, additional steps such tag removal followed by IMAC purification or ion exchange chromatography may be included to obtain pure and homogeneous protein (*see Note 33*).

3.7 Quality Assurance

In addition to SDS-PAGE and SEC, if available, mass spectrometric analysis of every purified protein is highly recommended (*see* Fig. 4c). This confirms the molecular weight of the protein, with mass discrepancies indicating mutations or cloning artifacts and potential posttranslational modifications. The protein is loaded into a small C3 HPLC column for desalting and eluted onto an in-line electrospray ionization time-of-flight analyzer. Any discrepancy needs to be explained, either by sequencing the DNA, by enzymatic removal of suspected modifications or by MS/MS analysis of proteolytic fragments.

4 Notes

1. It is advisable to prepare small batches (10–100 mL) of SOC medium as this can become contaminated very quickly.
2. On the day of purification, only prepare the required amount of buffer for the number of samples to be purified; for example, for one 96-well plate you will need about 25 mL of Lysis buffer, 70 mL of Wash buffer, and 5 mL of Elution buffer. The stock buffers can be stored at 4 °C for at least 1 month.
3. Be careful not to splash the cells up the sides of the wells while using the repeat pipettor and also check that the cells are at the bottom of the well before continuing. The cells can be added first using the repeat pipettor, followed by the DNA using the multichannel pipette which may reduce the risk of cross-contamination.
4. The SOC medium can be added using a multichannel pipette with the medium in a reagent reservoir.
5. The transformation can be performed for individual clones. In this case, plate 80 μ L of the transformation mixture onto a 50 mm petri dish and spread with a sterile spreader.
6. Multiple colonies are selected at this stage in order to account for clone-to-clone variation in expression levels of the protein.
7. We recommend not adding chloramphenicol at this stage as the pRARE2 plasmid is not lost during expression; however, its addition may significantly slow down the bacterial growth.
8. Using the 96-well plate reader to determine the OD_{600nm} of the cultures across the whole 96-well block indicates any inconsistencies with growth for particular targets, or constructs, and can therefore be used to identify proteins which failed to express because of a lack of proper growth. However, this step is not essential.
9. The cell pellets can also be stored at –80 °C for 1–2 weeks if necessary.

10. It is useful to set up two 96-well plates of test proteins in parallel to provide a balance for the centrifugation steps; however, a balance block can be used containing water instead.
11. At this point, you can remove 15 μL of clarified lysate (as the Soluble fraction) and mix with 5 μL of SB in a PCR plate before transferring to the filter plate.
12. Take care to avoid transferring insoluble material to the resin as it may block the filter plate in subsequent steps. To avoid disturbing the Insoluble fraction, tilt the plate and drive the tips down the side of the wells at an angle. Stop just above the pellet, on most plates there is a ridge just off the bottom—feel for this with the tips. Gently pipette up the supernatant and then transfer to the new plate. Do not go back into the wells as this will resuspend the pellets, if this happens then respin the sample and try again.
13. Alternatively, incubate at RT on a shaking platform. It is advisable to place the filter plate on top of a 96-well microtiter plate to avoid any drips coming through onto the shaker.
14. This step can also be done using centrifugation ($200 \times g$ for 1 min).
15. Removing all trace of Wash buffer is essential to ensure that the subsequent elution step does not become diluted with Wash buffer.
16. As standard, we only run the Purified fractions on gels to identify which proteins are expressed, soluble and purified. We will only analyze the Total and Soluble fractions if we want to determine whether or not a protein has been expressed but is insoluble or if the control protein has failed to purify.
17. Alternatively, retransform the expression plasmid into BL21 (DE3)-R3-pRARE2 as described in Subheading 3.1, except plate 80 μL of the transformation mixture onto a 50 mm petri dish and spread with a sterile spreader.
18. One 10 mL starter culture is required per liter of culture scaled up. If you are planning to scale up to more than 1 L, prepare starter cultures proportionately. We generally find that 1 L scale is sufficient to obtain milligram quantities of highly expressed proteins, that is, those having large visible bands on Coomassie SDS-PAGE after test purification (*see* Fig. 3). If the bands are weak, you may need to scale up to 3 L of culture or more.
19. The flasks can be autoclaved with the media in them but do not use porous seals during autoclaving; instead, cover the flasks with a piece of aluminum foil and use porous seals only during cell growth. The bacterial growth is an important determinant for protein expression and is mainly affected by aeration, stirring, and temperature. Efficient aeration in shaker flasks can be

achieved by optimizing the ratio of culture volume to the total capacity of flask and shaking speed. The wide mouth design of the 2.5 L Ultra Yield flasks with straight walls and baffles at the bottom of the flasks facilitate good oxygenation for culture volumes up to 1 L. Conventional baffled Erlenmeyer flasks provide comparable aeration but with lower culture to vessel ratios (typically 1:4).

20. Using a 5 mL serological pipette, remove 1 mL of sample and measure OD_{600nm}. OD measurements above 0.5 are not linear, dilute the culture if it is at higher OD before measurement and use the corrected value to obtain the precise OD. When cells are grown in TB medium, induction at an OD_{600nm} value of 1.5–4.0, followed by overnight growth at reduced temperature is optimum for protein expression. However, this may need to be optimized for individual proteins.
21. A concentration of 0.1 mM IPTG is sufficient for most strains; however, others, such as pLysS, may require higher concentrations in the range of 1–2 mM for efficient induction. We find that the optimum temperature of induction for the majority of the human proteins that we express in *E. coli* is 18–25 °C. It may be beneficial to test a number of temperatures (ranging from 15–37 °C) at the test expression stage for specific proteins.
22. At this stage you can remove a 5 mL sample and harvest the cell pellet by centrifugation in a 15 mL tube to perform a test purification which will confirm if the scale-up expression has been successful, before proceeding to large-scale purification.
23. If the cell pellets are not used for protein purification immediately, they can be frozen directly or after resuspension in a small volume of Lysis buffer at –80 °C. If the cells are frozen after resuspension in buffer, thawing may result in a very viscous solution because of cell lysis and release of nucleic acids. Viscosity can be reduced by the addition of Benzonase nuclease to the cell lysate at a concentration of 25–50 U/mL. The addition of protease inhibitors is important when freezing pellets in Lysis buffer to reduce protein degradation. However, it is advisable to test your protein by purification first to determine how sensitive it is to degradation. Some proteins require purification immediately from cell harvesting to prevent them from proteolytic degradation.
24. Although many methods are available to lyse cells, high-pressure cell disruption is a very efficient way of lysing large volumes of cell suspensions. For smaller volumes (<100 mL) sonication can be used effectively. However, both methods can cause localized heating which can result in protein precipitation or denaturation; therefore, it is important to keep samples on

ice at all times and prechill the cell disruptor. Cell disruption by sonication can also help in reducing viscosity by shearing nucleic acids. Use of detergents should be avoided for cell lysis if its presence will interfere with the downstream applications such as protein crystallization and cryo-EM. To lyse your cells by sonication, transfer the cell suspension to a 50 mL conical tube or a beaker depending on the volume and place the container on ice. Sonicate the cell suspension using 10–15 bursts of 5 s on, 10 s off. Generally, an amplitude of 35% using a 750 W Sonics Vibra-Cell sonicator is sufficient for lysis of a 50 mL cell suspension. The sonication time may need to be adjusted depending on the volume of the cell suspension. Avoid excessive foaming and heating of the suspension by keeping the cell suspension on ice at all times.

25. PEI is a highly positively charged polymer at neutral pH and can be used to remove negatively charged nucleic acids from cell lysates by precipitation in the presence of high salt. At lower ionic strength, nucleic acid binding proteins may remain bound to the nucleic acid and the use of PEI may precipitate proteins of interest along with the nucleic acid. Therefore, it is crucial to maintain a high salt concentration (>0.5 M NaCl) during this step. Alternatively, a preequilibrated anion exchange column such as DEAE-cellulose (DE52) may be used prior to IMAC purification.
26. If the supernatant is still turbid after centrifugation or if the pellet dislodges, add additional PEI to a final concentration of 0.05% and repeat the centrifugation step to obtain clear supernatant. If the lysate is still turbid, before proceeding to the IMAC step, filter the supernatant using a $0.80\ \mu\text{m}$ syringe filter first, followed by a $0.45\ \mu\text{m}$ syringe filter to remove large particulates and cell debris which can delay the binding, washing, and elution steps in the chromatography procedure.
27. Phosphate and HEPES buffers with 0.5 M NaCl concentration work equally well for the IMAC; HEPES is preferred as divalent ions (e.g., Mg^{2+} , Ca^{2+} , or Zn^{2+}) are included to avoid precipitation. If the purified protein is to be used for crystallization, care must be taken to exchange the buffer from phosphate to HEPES during later stages of purification (such as SEC) because phosphates may form salt crystals with many of the crystallization solutions. A commonly observed problem in IMAC is copurification of intrinsic proteins from host cells due to affinity of exposed histidines or metal binding moieties toward immobilized metal ions. Success of the technique depends on buffer composition, pH, and ionic strength. The binding of his-tagged proteins to the resin is optimal at physiological pH; therefore, it is important to keep the Lysis buffer pH close to 7.5–8.0. Higher salt concentration (> 0.5 M

NaCl) is also responsible for avoiding nonspecific binding of proteins to IMAC resin. Salt concentration also plays an important role in protein stability in solution; therefore, it is crucial that the ionic strength of the buffers should not be reduced too far, as this may promote protein precipitation. The presence of 5–10% glycerol is useful to promote protein stability; however, it may inhibit protein crystallization in some cases and should be completely absent for cryo-EM studies.

28. It is important to collect the eluates in fractions and analyze them by SDS-PAGE before pooling them together. Pooling the fractions together before SDS-PAGE analysis can result in mixing of the purified sample with other contaminated fractions or dilution of concentrated fractions. Protein purified through IMAC may be pure enough for some functional studies, but it is rarely pure enough for structural studies. Many host proteins bind nonspecifically during affinity chromatography which can be separated from the target protein by introducing a size exclusion chromatography step. This step also gives important information on oligomeric state of the protein and is useful in separating any contaminant proteins as well as aggregates.
29. To obtain high resolution separation on the SEC column, load a maximum volume of 5 mL (for column size mentioned in this protocol). It may be necessary to concentrate your protein before applying to the SEC column to reduce the volume and remember to filter the sample before loading to remove any aggregates.
30. If using an ÄKTA-Xpress system for purification, the detection parameters should be changed to accommodate varying protein loads. We recommend using the default parameters with the following changes:
Affinity peak collection:
Start: Watch level greater than 20 mAU, slope greater than 25 mAU/min.
Stop: Peak max factor 0.5, watch level less than 20 mAU, watch stable plateau for 0.5 min,
delta plateau 5 mAU/min.
Gel filtration peak collection:
Elution volume before fractionation: 0.3 CV.
Elution volume with fractionation: 0.8 CV.
Peak fractionation algorithm: level_OR_slope.
Start level 10 mAU, start slope 5 mAU/min.
Peak max factor 0.5, minimum peak width 0.5 min.
Stop level 10 mAU, slope 5 mAU/min.

31. Care must be taken while pooling protein fractions. Pay special attention to the concentration of the target protein and the level of contaminant proteins on the gel, analyze the SEC spectra and compare with molecular weight standards (which have been separated on the same column). Eliminate aggregated proteins (eluted in the void volume of SEC) and pool together fractions corresponding to monomer or oligomer peaks separately. Pool fractions from well-formed and symmetrical peaks and avoid mixing fractions from long tails which may represent some heterogeneity.
32. Protein aggregation can occur at any stage of the expression/purification procedure, but is very common during the process of concentration. This becomes clearly apparent when attempting to concentrate by ultrafiltration as protein aggregates rapidly block the filter and it becomes impossible to further concentrate the protein. Therefore, it is important to test a small volume of protein for its ability to concentrate before committing to concentrate the whole protein sample. Measure the protein concentration using a NanoDrop™ spectrophotometer or similar before starting to concentrate. Choose an appropriate protein concentrator with molecular weight cut off size that is two times smaller than the protein molecular weight. Transfer 200–500 μL of the sample to a concentrator that fits into a 1.5 mL microcentrifuge tube. Centrifuge according to the manufacturer's instructions at 4–15 $^{\circ}\text{C}$. Check the volume every 10–15 min and more regularly when the volume is low. The sample should concentrate quite rapidly to a protein concentration of at least 5–10 mg/mL. If the process is stuck with no apparent reduction in volume, it is likely that the protein is aggregating. The aggregates can be detected by analytical SEC or light scattering. If the protein aggregates easily, change in buffer pH, NaCl concentration or use of additives should be considered. Once the concentration conditions are established, the remaining protein can be concentrated using those parameters.
33. Impurity can be a result of copurification of contaminant proteins. To improve the purity of such samples, additional chromatographic steps can be employed. An effective general purification step is removal of the tag by cleavage with TEV protease followed by rebinding to Ni-IDA/NTA resin which is an efficient way to remove contaminants. An overnight digestion with TEV protease at 4 $^{\circ}\text{C}$ removes the his-tag (*see* Fig. 4a), the cleaved protein is then applied to Ni-IDA/NTA resin, which will isolate the cleaved his-tag as well as other contaminant proteins by their affinity for the beads and the target protein is collected in the flow through. In order for this protocol to work, the protein solution must not contain more

than 25 mM imidazole; this can be achieved by SEC (before or after cleavage), or by performing the TEV cleavage during dialysis of the protein. Further purification can be achieved using ion exchange and other chromatographic methods that need to be specifically tailored for each protein.

Acknowledgments

We would like to thank all the SGC scientists (past and present) who contributed toward the development of the method. The SGC is a registered charity (number 1097737) that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genome Canada, Innovative Medicines Initiative (EU/EFPIA), Janssen, Merck KGaA, MSD, Novartis Pharma AG, Ontario Ministry of Economic Development and Innovation, Pfizer, São Paulo Research Foundation-FAPESP, Takeda, and Wellcome.

References

1. Sorensen HP (2010) Towards universal systems for recombinant gene expression. *Microb Cell Factories* 9:27
2. Graslund S, Nordlund P, Weigelt J et al (2008) Protein production and purification. *Nat Methods* 5(2):135–146
3. Burgess-Brown NA, Sharma S et al (2008) Codon optimization can improve expression of human genes in *Escherichia coli*: a multi-gene study. *Protein Expr Purif* 59(1):94–102
4. Tegel H, Tourle S, Ottosson J et al (2010) Increased levels of recombinant human proteins with the *Escherichia coli* strain Rosetta (DE3). *Protein Expr Purif* 69(2):159–167
5. Cornvik T, Dahlroth SL, Magnusdottir A et al (2005) Colony filtration blot: a new screening method for soluble protein expression in *Escherichia coli*. *Nat Methods* 2(7):507–509
6. Cornvik T, Dahlroth SL, Magnusdottir A et al (2006) An efficient and generic strategy for producing soluble human proteins and domains in *E. coli* by screening construct libraries. *Proteins* 65(2):266–273
7. Savitsky P, Bray J, Cooper CD et al (2010) High-throughput production of human proteins for crystallization: the SGC experience. *J Struct Biol* 172(1):3–13
8. Gileadi O, Burgess-Brown NA, Colebrook SM et al (2008) High throughput production of recombinant human proteins for crystallography. *Methods Mol Biol* 426:221–246
9. Hanahan D, Jessee J, Bloom FR (1991) Plasmid transformation of *Escherichia coli* and other bacteria. *Methods Enzymol* 204:63–113
10. Magnusdottir A, Johansson I, Dahlgren LG et al (2009) Enabling IMAC purification of low abundance recombinant proteins from *E. coli* lysates. *Nat Methods* 6(7):477–478



Screening and Production of Recombinant Human Proteins: Protein Production in Insect Cells

Pravin Mahajan, Claire Strain-Damerell, Shubhashish Mukhopadhyay, Alejandra Fernandez-Cid, Opher Gileadi, and Nicola A. Burgess-Brown

Abstract

This chapter describes the step-by-step methods employed by the Structural Genomics Consortium (SGC) for screening and producing proteins in the baculovirus expression vector system (BEVS). This eukaryotic expression system was selected and a screening process established in 2007 as a measure to tackle the more challenging kinase, RNA–DNA processing, and integral membrane protein families on our target list. Here, we discuss our platform for identifying soluble proteins from 3 mL of insect cell culture and describe the procedures involved in producing protein from liter-scale cultures.

Key words Insect cells, Baculovirus, BEVS, Expression, Recombinant, Protein, Purification, IMAC, SEC chromatography, Gel filtration

1 Introduction

Availability of a pure protein is essential for obtaining information on protein structure and function. Heterologous protein production in *E. coli* has remained the preferred system for many research laboratories as it is low-cost, fast, and easy to handle. However, there is no guarantee that *E. coli* cells will produce eukaryotic proteins in a soluble and biologically active form because of a number of limitations such as codon bias, lack of posttranslational modifications (PTMs), or disulfide bond formation. Exploring other protein expression hosts such as mammalian cells, yeast, and insect cells is often required if *E. coli* fails to produce soluble protein after attempting different strains, solubility enhancing tags, and so on. Among the alternatives available, the baculovirus expression vector system (BEVS) is increasingly becoming popular for expression of recombinant proteins as it is nonpathogenic to humans [1], capable of producing high levels of soluble proteins with PTMs similar to those observed in mammalian cells and easily scalable in

suspension culture [2]. This system is also proving popular for the production of large protein complexes, production of virus-like particles, gene delivery, viral vector vaccines, expression of proteins in mammalian cells, and display of proteins and peptides on the baculovirus envelope [3]. Baculoviruses are double-stranded DNA viruses [4] most of which infect insects of the order Lepidoptera [5]. The most widely used baculovirus used as a BEVS is *Autographa californica* multiple nuclear polyhedrosis virus (AcMNPV). Two major genes that express in the very late phase of baculovirus infection of insects are p10 and polyhedrin which are strong expressers but dispensable for viral replication. This discovery has allowed for exploitation of the p10 and polyhedrin promoters to be used for driving recombinant protein expression in BEVS; the polyhedrin promoter in particular has been described as a workhorse promoter of BEVS [6]. The most common insect cell lines utilized as hosts of BEVS are Sf9 and Sf21 derived from pupal ovarian tissue of the fall army worm, *Spodoptera frugiperda* [7] and High Five cells (BTI-Tn-5B1-4) derived from ovarian cells of the cabbage looper, *Trichoplusia ni* [8].

Since the first use of baculoviruses for protein expression in 1983 [9], the system has gone through numerous technological advances that have allowed it to be widely accessible. Various baculovirus expression systems are commercially available to produce baculoviruses, most notably Bac-to-Bac[®] (Invitrogen), flashBAC (Oxford Expression Technologies), BaculoDirect[™] (Invitrogen), BacVector[®]-3000 (Novagen), BacPAK (Clontech), and Bac-n-Blue[™] (Invitrogen). About 12 years ago, it became evident in our laboratory that the bacterial expression system was unable to cope with more challenging proteins on our target list such as many protein kinases, RNA–DNA processing proteins, and integral membrane proteins (IMPs). To address this issue, we established an efficient process based on the Bac-to-Bac[®] system [10] for screening multiple versions of each protein in insect cells to identify those that were amenable to purification and crystallization. The 96-well cloning procedure is described in detail in Chapter 3. In this chapter we continue the methodologies for expression screening and scaling up expression of proteins in suspension culture. To describe our series of standardized protocols for protein production in insect cells, this chapter is broadly divided into the following stages: (a) transposition, bacmid production and PCR screen; (b) growth and maintenance of insect cell lines in adherent and suspension culture; (c) transfection into Sf9 cells, baculovirus generation, and small-scale test expression/purification; and (d) large-scale protein expression and purification. The screening process has been miniaturized to 24-well format. The steps involved in the pipeline from cloning to large-scale expression are outlined in Fig. 1.

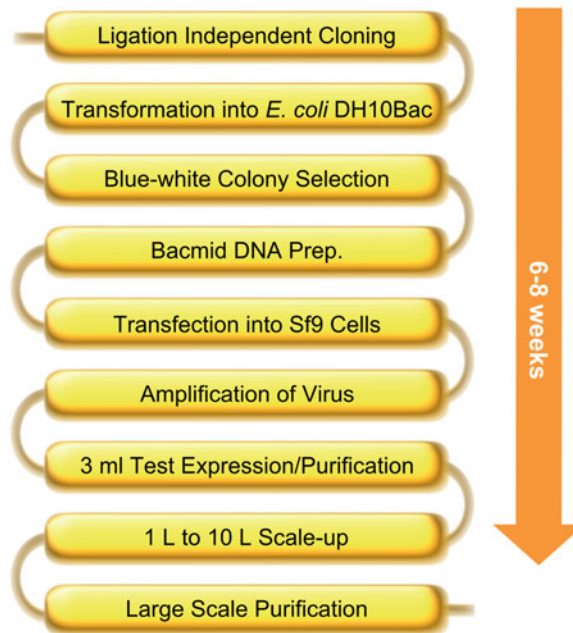


Fig. 1 Overview of the Baculovirus expression process. The process takes ~6–8 weeks from LIC to scale-up

2 Materials

Unless otherwise stated, all solutions are prepared using ultrapure water (prepared by purifying deionized water to reach a resistivity of 18 M Ω cm at 25 °C) and analytical grade reagents.

2.1 Transposition and Bacmid Preparation

1. *E. coli* DH10Bac (Invitrogen) or DH10EMBacY (Geneva Biotech) chemically competent bacterial cells are prepared in house as described [11] (*see Note 1*).
2. Primers: Primers are supplied by Eurofins and are HPSF purified at 0.01 or 0.05 μ mol scale. Primer stocks are either supplied at or diluted (in 10 mM Tris-HCl buffer, pH 8.0) to 100 μ M and stored at -20 °C.
3. MyTaq™ Red DNA Polymerase (5 unit/ μ L, Bioline).
4. Molecular biology grade water.
5. 10 mM dNTP solution: 2.5 mM dATP, 2.5 mM dTTP, 2.5 mM dGTP, and 2.5 mM dCTP (prepare from 100 mM dNTP set) diluted in molecular biology grade water and stored at -20 °C.
6. 50 \times TAE buffer (1 L): Dissolve 242 g of Tris base, 57.1 mL of glacial acetic acid, and 100 mL of 0.5 M EDTA, pH 8.0 in water and adjust pH to 8.5. Filter through a 0.2 μ m membrane filter and use as a 1 \times solution.

7. 96-Well 1.5% TAE-agarose gels: Dissolve 3 g of agarose powder in 200 mL of 1× TAE buffer using a microwave. Once cooled to hand-hot, add 8 μL of SYBR-safe DNA gel stain (Invitrogen), mix by swirling and cast in a Sub-cell Model 96 (Bio-Rad or similar) gel cast.
8. DNA ladder: 1 kb Plus DNA Ladder (Invitrogen) prepared in 1× BlueJuice™ (Invitrogen) diluted in molecular biology grade water.
9. TE Buffer: Prepare a solution of 10 mM Tris-HCl and 1 mM EDTA, pH 8.0, filter through a 0.20 μm syringe filter and store at room temperature (RT).
10. 60% (v/v) glycerol: Autoclave to sterilize.
11. 70% (v/v) ethanol.
12. 50 mg/mL kanamycin: Prepare in water, filter through a 0.20 μm syringe filter and store at -20 °C.
13. 10 mg/mL tetracycline: Prepare in ethanol and store -20 °C.
14. 7 mg/mL gentamycin: Prepare in water, filter through a 0.20 μm syringe filter, and store at -20 °C.
15. 100 mg/mL Blue-gal (Glycosynth): Prepare in dimethyl sulfoxide (DMSO) and store -20 °C.
16. 40 mg/mL IPTG: Prepare in water, filter through a 0.20 μm syringe filter, and store at -20 °C.
17. LB agar: Dissolve 22.5 g of premixed LB broth and 13.5 g of agar in 800 mL of ultrapure water. Adjust volume to 900 mL and autoclave on the same day.
18. Recombinant bacmid selection plates: Melt LB agar slowly in a microwave and add 5% (w/v) sucrose. Once cooled to hand-hot, add the appropriate antibiotic and swirl vigorously to mix. Pour 10 mL of the molten agar into each 50 mm petri dish and once set, upturn and leave open to dry. These can be prepared ahead of time and stored for up to a month at 4 °C, sealed in a plastic bag to prevent overdrying.
19. 2× LB: Dissolve 45 g of premixed LB broth in 800 mL of water. Adjust volume to 900 mL and autoclave on the same day.
20. Virkon.
21. Montage Plasmid Miniprep_{HTS} 96 Kit (Millipore, *see* **Note 6**).
22. 50 mm petri dishes.
23. 96-well PCR plates.
24. 96-well microtiter plates that can hold up to 200 μL of sample.
25. 96-deep-well blocks.
26. Adhesive tape pads.

27. 96-well filter plates, 25 μm .
28. Adhesive PCR seals.
29. AirOtop porous seals (Thomson or VWR).
30. Silicone 96-Square-Well AxyMat (Axygen).
31. Disposable sterile spreaders or 2 mm autoclaved glass balls (VWR).
32. Disposable sterile inoculation loops (1 μL).
33. Reagent reservoirs for multichannel pipetting.
34. Minisart syringe filters, 0.20 μm .
35. Supor® PES Membrane Disc Filters, 0.2 μm and unit (Pall).
36. Multichannel pipettes and repeat pipettors are used to dispense reagents into a 96-well format.
37. 96-well PCR thermocycler with heated lid.
38. 96-well gel cast and tank (Subcell Model 96 Bio-Rad or similar).
39. All gels are imaged on a Gel Doc™ XR+ (Bio-Rad).
40. A UV spectrophotometer for measuring DNA and protein concentration (e.g., The NanoDrop™ spectrophotometer allows for measurements from as low as 1.5 μL volumes).
41. Scanlaf Mars recirculating class II biological safety cabinet (BSC).
42. Micro-Express Glas-Col shaker (Glas-Col, Indiana, USA) or alternative that ranges in temperature from 18 °C to 37 °C and shakes up to 800 rpm.
43. 96-well block mixer (Eppendorf MixMate or similar).
44. Water bath set at 42 °C.
45. Incubator set at 37 °C.
46. Centrifuge suitable for 96-deep-well blocks (3,000 $\times g$).

2.2 Transfection and Cell Growth

The following reagents, consumables, and equipment are required in addition to those listed above:

1. Cell lines: Sf9 insect cells, SFM adapted (Invitrogen); High Five cells, SFM adapted (Invitrogen).
2. Media: Sf-900™ II SFM (1 \times) (Invitrogen).
3. Reagents: fetal bovine serum (FBS), insect cell culture tested (Invitrogen); Insect GeneJuice® (Merck), Pen/Strep (use at 50 units penicillin and 50 μg streptomycin per mL of medium); 0.4% Trypan Blue Stain.
4. DMSO, Molecular Biology grade (DNase/RNase free).
5. Cryovials.

6. 24-well tissue culture plates.
7. 24-well blocks (Microplate Devices Uniplate[®] or similar).
8. 250, 500, and 1,000 mL flasks with vented cap (Corning).
9. Stripette pipettes.
10. Inverted light microscope (Axiovert 25, CarlZeiss).
11. Hemocytometer, improved Neubauer (VWR International).
12. Static incubator set at 37 °C.
13. Multitron shaker-incubators with cooling capacity (Infors HT).

**2.3 Virus
Amplification and Test
Expression**

All reagents, consumables, and equipment listed above.

2.4 Test Purification

The following reagents, consumables, and equipment are required in addition to those listed above:

1. Benzonase (Novagen, HC, 250 units/ μ L).
2. Protease Inhibitor Cocktail Set III (Calbiochem).
3. 0.5 M Tris(2-carboxyethyl)phosphine (TCEP): Prepare in water, filter through a 0.20 μ m syringe filter, and store at -20 °C.
4. 1 M dithiothreitol (DTT): Prepare in water, filter through a 0.20 μ m syringe filter, and store as 1 mL aliquots at -20 °C.
5. SeeBlue[®] Plus2 Pre-Stained Standard (Invitrogen).
6. InstantBlue[™] (Expedeon Protein Solutions).
7. 20 \times NuPAGE[™] MES SDS Running Buffer (Invitrogen).
8. PBS: Dissolve 5 tablets of PBS in 1 L of water, filter through a 0.2 μ m membrane filter, and store at 4 °C.
9. 1 M HEPES, pH 7.5: Prepare in water, filter through a 0.2 μ m membrane filter, and store at RT.
10. 5 M NaCl: Prepare in water, filter through a 0.2 μ m membrane filter, and store at RT.
11. 3 M imidazole, pH 8.0: Prepare in water, filter through a 0.2 μ m membrane filter, and store at RT.
12. 50% (v/v) glycerol: Autoclave and store at RT.
13. Lysis buffer (1 L): 50 mM HEPES, pH 7.5, 300 mM NaCl, 5% (v/v) glycerol, and 10 mM imidazole prepared in advance, filtered through a 0.2 μ m membrane filter and stored at 4 °C. On the day of purification, add 0.2 μ L/mL Benzonase, 1 μ L/mL protease inhibitor cocktail, and 0.5 mM TCEP.

14. Wash buffer (1 L): 50 mM HEPES, pH 7.5, 300 mM NaCl, 5% (v/v) glycerol, and 30 mM imidazole prepared in advance, filtered through a 0.2 μ m membrane filter and stored at 4 °C. Add 0.5 mM TCEP on the day of purification.
15. Elution buffer (0.1 L): 50 mM HEPES, pH 7.5, 300 mM NaCl, 5% (v/v) glycerol, and 500 mM imidazole prepared in advance, filtered through a 0.2 μ m membrane filter and stored at 4 °C. Add 0.5 mM TCEP on the day of purification.
16. Affinity buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, 10% glycerol, and 10 mM imidazole prepared in advance, filtered through a 0.2 μ m membrane filter, and stored at 4 °C.
17. 50% (w/v) Ni-IDA Metal Chelate Resin (Generson) or Ni-NTA-agarose (Qiagen): The IMAC resins are generally supplied in 20% ethanol. To equilibrate, wash the resin twice in water and then three times in Affinity buffer in a 50 mL tube, by inverting to resuspend the resin and centrifuging at $500 \times g$ for 1 min. After the final wash, resuspend the resin in Affinity buffer as 50% (w/v) slurry and store at 4 °C when not in use.
18. SB: Prepare a stock of NuPAGE LDS sample buffer (Invitrogen) containing DTT (1:4 dilution of 1 M DTT in NuPAGE LDS sample buffer) and store at -20 °C.
19. 96-Well filter plates.
20. Precast 26-Lane SDS-PAGE gradient gels (4–12% Bis-Tris) (Invitrogen).
21. Protein gel electrophoresis apparatus (Invitrogen).
22. 96-Well thermocycler with heated lid.
23. Vibra-Cell Sonicator with 24-well probe (Sonics[®]).
24. General purpose benchtop centrifuge (Sorvall Legend RT, Kendro).

2.5 Large-Scale Expression

The following reagents, consumables, and equipment are required in addition to those listed above:

1. Media: Sf-900[™] II SFM (1 \times) (Invitrogen); Insect-XPRESS serum-free and protein-free medium (Lonza).
2. Nonbaffled Erlenmeyer flasks: glass or polycarbonate in various sizes 250 mL, 500 mL, and 1 L and glass flasks of 3 L capacity for large-scale expression.
3. Cell freezing container: Mr. Frosty (Nalgene).
4. Avanti J-20XP or Avanti J-26XP centrifuge or similar (Beckman Coulter) with a JLA 8.1000 rotor for harvesting large volumes of cells.
5. Chemgene.
6. Alconox[®].

2.6 Protein Extraction and Large-Scale Purification

The following reagents, consumables, and equipment are required in addition to those listed above.

1. Complete EDTA-free protease inhibitor (Roche).
2. $2\times$ Lysis buffer: 100 mM HEPES buffer, pH 7.5, 1 M NaCl, 20% (v/v) glycerol, and 20 mM imidazole. Filter through a 0.2 μm membrane filter and store at 4 $^{\circ}\text{C}$. On the day of purification, add Benzonase (0.2 $\mu\text{L}/\text{mL}$ of cell lysate), Protease inhibitor cocktail (2 $\mu\text{L}/\text{mL}$ of cell lysate) or Complete EDTA-free protease inhibitor cocktail (1 tablet/25 mL of cell lysate), and 1 mM TCEP.
3. Lysis buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, 10% (v/v) glycerol, and 10 mM imidazole. Filter through a 0.2 μm membrane filter and store at 4 $^{\circ}\text{C}$. On the day of purification, add Benzonase (0.1 $\mu\text{L}/\text{mL}$ of cell lysate), Protease inhibitor cocktail (1 $\mu\text{L}/\text{mL}$ of cell lysate) or Complete EDTA-free protease inhibitor cocktail (1 tablet/50 mL of cell lysate), and 0.5 mM TCEP.
4. Affinity buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, 10% glycerol, and 10 mM imidazole. Filter through a 0.2 μm membrane filter and store at 4 $^{\circ}\text{C}$. Add 0.5 mM TCEP on the day of purification.
5. Wash buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, 10% glycerol, and 30 mM imidazole. Filter through a 0.2 μm membrane filter and store at 4 $^{\circ}\text{C}$. Add 0.5 mM TCEP on the day of purification.
6. Elution buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, 10% glycerol, and 300 mM imidazole. Filter through a 0.2 μm membrane filter and store at 4 $^{\circ}\text{C}$. Add 0.5 mM TCEP on the day of purification.
7. Size Exclusion Chromatography buffer (SEC): 20 mM HEPES, pH 7.5, 500 mM NaCl, and 5% glycerol. Filter through a 0.2 μm membrane filter and store at 4 $^{\circ}\text{C}$. Add 0.5 mM TCEP on the day of purification.
8. Minisart syringe filters, 0.20 μm , 0.45 μm , and 0.80 μm .
9. Amicon Ultra protein concentrators.
10. Sonicator (Sonics Vibra-Cell, VCX 750, Sonics & Materials INC) or basic Z model cell disruptor (Constant Systems Ltd).
11. ÄKTA-Xpress or ÄKTA-Purifier liquid chromatography system.
12. HiTrap 5 mL FF columns for his-tagged protein purification.
13. Ion exchange chromatography columns such as HiTrap 5 mL Q FF and SP FF.

14. HiLoad Superdex columns for preparative size exclusion chromatography such as HiLoad 16/600 Superdex™ S75 pg, S200 pg, or Superose™ 6 Increase 10/300 GL.
15. JA-25.50 rotor for centrifugation of cell lysates.

3 Methods

3.1 Transposition in *E. coli* DH10Bac or DH10EMBacY

The transposition process is outlined in Fig. 2.

1. Prepare at least 100 petri dishes (50 mm) containing approximately 10 mL of LB agar, supplemented with 50 µg/mL kanamycin, 7 µg/mL gentamycin, 10 µg/mL tetracycline, 40 µg/mL IPTG, and 100 µg/mL Blue-gal (*see Note 2*) and once set, allow to dry, inverted at RT.
2. Using a multichannel pipette, add 3 µL of recombinant DNA to a 96-well PCR plate.
3. On ice, add 30 µL of chemically competent *E. coli* DH10Bac or DH10EMBacY cells using a repeat pipettor (*see Note 3*), cover with an adhesive tape pad and incubate for 30 min. It is advisable to include a positive control (i.e., a construct that has previously shown soluble protein expression in BEVS) in position H12 of the 96-well plate.
4. In the meantime, add 900 µL of prewarmed 2× LB medium containing 50 µg/mL kanamycin, 10 µg/mL tetracycline, and 1 µg/mL gentamycin to each well of a 96-deep-well block using a reagent reservoir.
5. Heat shock the cells in the PCR plate for 45 s in a 42 °C water bath and return briefly to ice.
6. Transfer the bacterial suspension into the prewarmed medium block (**step 4**), cover with a porous seal and incubate in a Glas-Col shaker (or equivalent) at 37 °C with shaking at 700 rpm for 5 h.
7. Dilute the culture (10 µL into 90 µL) into LB (or 2× LB) medium in a 96-well microtiter plate and spread 50 µL onto the recombinant bacmid selection plates (*see Note 4*).
8. Incubate the plates at 37 °C for 48 h, covered with foil (*see Note 5*).
9. White colonies contain the recombinant bacmid DNA and the blue ones do not (*see Fig. 2*). To ensure that the colonies are white, divide a selective plate into 6 or 8 sectors using a marker pen and label with the well position (e.g., A1). Pick single colonies, streak to dilution using a sterile loop and incubate at 37 °C overnight.

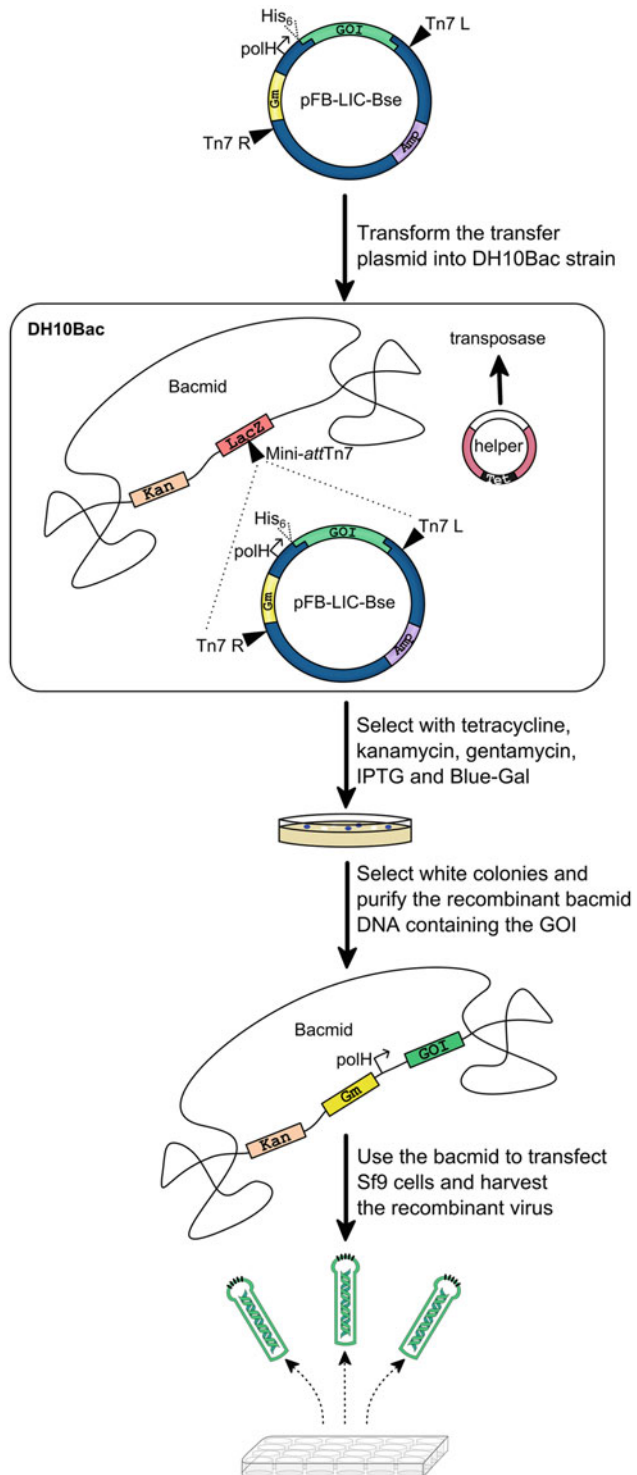


Fig. 2 Diagram describing the transposition process. The construct DNA is transformed into the DH10Bac *E. coli* strain, which contains both bacmid DNA and a helper plasmid. The transposase, expressed from the helper plasmid, will facilitate transfer of the transposable element including the gene of interest (GOI) into the bacmid. The recombinant bacmid DNA can then be purified and used directly to transfect Sf9 insect cells

3.2 *Bacmid* Production

1. Inoculate the recombinant white colonies (isolated from the restreaked plates) into the corresponding wells of two 96-deep-well blocks, each containing 1 mL of 2× LB medium per well, supplemented with 50 µg/mL kanamycin, 7 µg/mL gentamycin, and 10 µg/mL tetracycline (*see Note 6*).
2. Cover with a porous seal and incubate at 37 °C overnight at 700 rpm in a Glas-Col shaker.
3. The following morning, prepare one or two glycerol stocks by mixing 120 µL of the overnight culture and 30 µL of 60% (v/v) glycerol in a 96-well microtiter plate, and store at –80 °C.
4. Centrifuge the deep-well blocks at 3,000 × *g* for 30 min. Decant the supernatant into a suitable container for Virkon decontamination. Invert the blocks and tap gently on absorbent paper.
5. Add 250 µL of the Solution 1 from the 96-well miniprep kit to each well of one block using a multichannel pipette (*see Note 7*).
6. Seal the block with a silicone sealing mat (*see Note 8*) and mix in the Glas-Col incubator for 2 min at 700 rpm or a 96-well MixMate (or equivalent) at 1,000–1,500 rpm. If necessary, resuspend using a multichannel pipette.
7. Transfer the suspension to the corresponding wells of the second block. Seal and repeat the mixing process.
8. Add 250 µL of Solution 2 to each well, seal with a silicone sealing mat, invert gently 5 times and incubate at RT for 10 min.
9. Add 300 µL of Solution 3, seal with a silicone sealing mat and mix gently but thoroughly by inverting 5 times.
10. Place the sample on ice for 20 min, then centrifuge at 3,000 × *g* for 30 min at 4 °C.
11. Transfer the clear supernatant to a fresh 96-deep-well block, cover with an adhesive tape pad and centrifuge again at 3,000 × *g* for 30 min at 4 °C (*see Note 9*).
12. In another fresh 96-deep-well block, dispense 0.8 mL of isopropanol into each well and add 0.8 mL of the clarified supernatant to the corresponding wells (*see Note 10*).
13. Using a 1 mL capacity multichannel pipette, gently mix up and down, cover with an adhesive tape pad and then incubate on ice for 30 min (*see Note 11*).
14. Centrifuge at 3,000 × *g* for 30 min at 4 °C.
15. Spray the outside of the 96-deep-well block with 70% (v/v) ethanol (*see Note 12*) and inside the biological safety cabinet (BSC), remove the cover from the block and discard the supernatant by decanting into a suitable container and blotting on absorbent paper.

16. Add 500 μL of 70% (v/v) ethanol to each well and tap the block gently to wash the pellets. Cover with an adhesive tape pad and then centrifuge at $3,000 \times g$ for 30 min at 4°C .
17. Inside the BSC, open the block and discard the supernatant by decanting. Tap the block very gently on absorbent paper to remove the ethanol. Allow the block to dry inside the hood for approximately 2 h or cover with porous seal and leave overnight in the BSC with it switched on (*see* **Note 13**).
18. Inside the BSC, add 50 μL of sterile TE buffer, cover with an adhesive tape pad and allow to stand for about 1 h. Very gently resuspend the bacmid DNA using a multichannel pipette (*see* **Note 14**) and transfer to a 96-well microtiter plate. Remove a couple of microliters of DNA from a few wells to measure the concentration using a UV-spectrophotometer. Pipette 1 μL of each DNA into a PCR plate for the bacmid PCR screen, then seal with a fresh adhesive tape pad.
19. Store bacmid DNA at 4°C until the test purification is complete, then store at -20°C .

3.3 Bacmid PCR Screen

1. Prepare a 10 μM primer stock (50 μL each of the 100 μM forward and reverse primers added to 400 μL of molecular biology grade water) of the bacmid screening primers (*see* Table 1). Store at -20°C .
2. Dilute the bacmid DNA 1 in 50 in molecular biology grade water in a 96-well PCR plate (*see* **Note 15**).
3. Set up a PCR master mix as follows: 400 μL of $5\times$ MyTaqTM Reaction Buffer Red, 1.49 mL of water, 100 μL of 10 μM of bacmid screening primers (**step 1**) and 10 μL of MyTaqTM DNA Polymerase (5 unit/ μL). Using a repeat pipettor or a multichannel, pipette 20 μL into each well of a 96-well PCR plate.
4. Transfer 2 μL of the diluted bacmid (**step 2**) to the PCR plate (**step 3**) and mix well.
5. Seal the PCR reaction plate with an adhesive PCR seal and set a thermocycler with the following conditions making sure that the block is up to 95°C before placing your sample plate in the instrument:
 - 95 $^\circ\text{C}$, 5 min
 - (95 $^\circ\text{C}$, 45 s; 50 $^\circ\text{C}$, 45 s; 72 $^\circ\text{C}$, 2–5* min) \times 25 cycles
 - 72 $^\circ\text{C}$, 7 min
 - 15 $^\circ\text{C}$ hold

*Extension time dependent on length of PCR product—for example, 30 s per 1 kb. Please note that additional base pairs will be added to your products due to the positioning of the screening primers (*see* Table 1).

Table 1
Primers used to confirm correct insertions at the bacmid PCR screen stage

Primer Name	Primer Sequence
Fbac-1	TATTCATACCGTCCCACCA
M13bac_rev	CAGGAAACAGCTATGAC

Fbac-1 and M13bac_rev are used for the Baculovirus vectors

6. While the PCR cycle is running, prepare a 96-well 1.5% TAE-agarose gel.
7. Using a multichannel pipette, load 10 μL of the PCR reaction mixtures directly onto the gel. Note that the spacing of the wells means that samples will be interleaved. Load 6 μL of 1 kb Plus DNA Ladder and run the gel at 150 V for 1 h.
8. Confirm the sizing of the products and repeat the screen for any constructs that do not produce a band of the correct size in the first screen (*see* **Note 15**).

3.4 Growth and Maintenance of Insect Cell Lines

Insect cell lines can be maintained in adherent culture as well as in suspension culture. Their ability to grow in suspension at high densities allows for expression of recombinant proteins in large scale; however, their ability to grow in monolayers can be utilized for the initial stage of transfection to generate baculoviruses. The most widely used insect cell lines for BEVS-based protein expression are Sf9, Sf21, and High Five, all of which are adaptable to serum-free, protein-free medium. We routinely use Sf9 cells for all the steps from transfection to large scale protein expression simply because of their robustness and ease in manipulation; however, occasionally High Five cells are used for large scale expression of proteins. Use of Sf9 cells for all steps in routine protocols ensures that uniform parameters are applied to a number of protein targets initially and if needed other cell lines can be tested later on to improve protein expression. Insect cell culture methods are described previously in detail [6, 12, 13]. Some important points when working with insect cells are mentioned in **Note 16**.

3.5 Reviving Sf9 Cell Line from Frozen Stock

Sf9 cells can be revived straight into suspension culture without first reviving them into adherent culture, provided there are sufficient cryovials of cells available in liquid nitrogen. Alternatively revive cells into adherent culture using T-flasks, then transfer to suspension culture at a density of 1×10^6 cells/mL from 70% to 80% confluent flasks, using sloughing off method (i.e., washing off layers of cells, instead of using traditional dislodging methods such as trypsin solution). Cells can be kept in suspension culture for 6–8 weeks, after which time a new stock should be revived as

older cells may show a decline in protein expression. There are different commercial formulations of serum-free insect cell media available; however, we use Sf-900™ II SFM mainly for initial revival of cells, transfection, expression testing, virus amplification and large-scale protein expression. Insect-XPRESS can also be used for large-scale protein expression. Sf9 cells adapt quickly from one medium to another. All of the cell culture steps described below are performed in aseptic conditions inside a BSC.

1. Warm Sf-900™ II SFM medium to 27 °C in a water bath and pipette 30 mL of the medium into a 250 mL flask.
2. Remove a cryovial containing the cells (at 3×10^7 cells/mL) from liquid nitrogen and carefully release the cap to depressurize it, then tighten it (*see Note 17*).
3. Transfer the cryovial to a container with warm water (25–30 °C) and incubate until the sample is 70% thawed.
4. Decontaminate the outside of the vial by wiping with 70% (v/v) ethanol.
5. Using a 5 mL Stripette, transfer the thawed cells immediately into the 250 mL flask containing the medium and pour the remaining icy cells from the cryovial straight into the flask.
6. Gently mix the cell suspension and transfer the flask to a 27 °C incubator with shaking at 90–100 rpm.
7. Check the cells after 48 h for good health.

3.6 Suspension Culture of Sf9 Cells in Shake Flask

Cells previously cultured in an anchorage-dependent manner need complete adaptation to suspension culture. The cells can be grown in suspension using either shake flasks or spinner flasks; however, our method of choice is the former. The use of simple shake flasks makes the process of protein expression in insect cells easily scalable from 10 mL to more than 10 L volume and does not require specialized equipment, which would be needed for spinner flasks and bioreactors.

1. After growing a sufficient number of cells, determine the viable cell count using Trypan Blue Stain and a hemocytometer (*see Note 18*).
2. Seed the cells to a density of 1×10^6 cells/mL into a 500 mL nonbaffled polycarbonate or glass flask in Sf-900™ II SFM medium.
3. Incubate the flask at 27 °C with shaking set at 90–105 rpm (*see Note 19*).
4. When the cells reach a density of 4×10^6 cells/mL, dilute them back to 1×10^6 cells/mL and expand the cell volume depending on requirement of the cells (*see Note 20*).

3.7 Cell Freezing

Once the cells start doubling regularly after revival it is advisable to freeze down the low passage number cells in several cryovials.

1. Prepare freezing medium containing 92.5% (v/v) Sf-900™ II SFM medium and 7.5% (v/v) DMSO and store at 4 °C.
2. Label sterile cryovials with the name of the cell line, date of freezing and any other relevant information and store the vials at 4 °C until ready to use.
3. Take a small suspension of cells from a shake flask and count viable cells using a hemocytometer. Alternatively, cells from adherent cultures can be used for freezing.
4. Take the required volume of cell suspension for 3×10^7 cells per vial.
5. Centrifuge the cells at $500 \times g$ for 10 min and discard the supernatant.
6. Resuspend the cells in the freezing medium (prepared in **step 1**) so that after resuspension the cell density is $\sim 3 \times 10^7$ cells/mL.
7. Quickly aliquot 1 mL of the cell suspension into the cryovials (prepared in **step 2**).
8. Place the vials into a suitable freezing container (e.g., Mr. Frosty) and transfer the container to a -80 °C freezer overnight (*see Note 21*).
9. The following day transfer the vials to liquid nitrogen storage.

3.8 Decontamination and Cleaning of Shake Flasks

It is extremely important to clean the shake flasks properly so that they can be reused without affecting the cell health or cell growth. Any residual disinfectant or scum of dead cells can adversely affect the cells and protein expression.

1. Pour off any spent media into a waste container and add 1 tablet of Virkon per L of the spent media.
2. Completely fill the empty culture flask with a 1 in 200 dilution of Chemgene and leave for at least 20 min (but no longer than 30 min). Make sure that every surface of the flask that has come in contact with virus is covered with the diluted Chemgene (*see Note 22*).
3. Discard the decontaminated waste and rinse with tap water.
4. Add a scoopful of Alconox[®], fill the flask with water, incubate for 20 min and scrub with a laboratory bottle brush to make sure that there is no visible cell debris or dead cell scum remaining inside the flask.
5. Leave the flask with fresh water for minimum 1 h.

6. If available, wash the flasks using a washer-disinfectant according to the manufacturer's instructions.
7. Dry the flasks in a drying cabinet set at 50–60 °C, cover with two layers of aluminum foil and autoclave.

3.9 Transfection into Sf9 Cells

1. Prepare ~100 mL of Sf9 cells 1 day in advance by diluting the cell count to 1×10^6 cells/mL in Sf-900™ II SFM medium.
2. The next day dilute the mid-log phase Sf9 cells to 2×10^5 cells/mL in Sf-900™ II SFM medium.
3. Label four 24-well tissue culture (TC) plates with 'plate 1' to 'plate 4' to cover your 96 samples (*see* Fig. 3 for how to transfer samples between 96-well and 24-well blocks or plates).
4. Using a 1 mL 12-channel multichannel pipette (with 6 tips spaced two apart), dispense 1 mL of diluted culture (**step 2**) into each well of four 24-well TC plates. Include controls: one for Insect GeneJuice®-only and the other for untreated cells (*see* **Note 23**). Incubate the plates at 27 °C for 1 h to allow for cell attachment (*see* **Note 24**).
5. Mix 200 µL of Insect GeneJuice® with 4 mL of Sf-900™ II SFM medium in a sterile 15 mL tube (sufficient for 100 reaction wells). Gently vortex for 10 s.
6. Dispense 40 µL of the mixture prepared in **step 5** into a sterile 96-well microtiter plate (leaving a well empty for the cell-only control).
7. Transfer 2 µL of recombinant bacmid DNA (concentration should be 0.5–2 µg/µL) into each well and cover the microtiter plate with an adhesive tape pad. Mix by tapping the plate gently or pipetting (*see* **Note 14**).
8. Incubate the mixture inside the BSC for 30 min; this incubation time is critical and extensions should be avoided.
9. After incubation, add 160 µL of Sf-900™ II SFM medium to the mixture in **step 8**.
10. Remove the 24-well TC plates containing the cells (**step 4**) from the incubator and aspirate the medium from the cells using a multichannel pipette.
11. Add the 200 µL DNA–Insect GeneJuice® mixture from **step 9** dropwise onto the cells using a 12-channel multichannel pipette (with 6 tips spaced two apart) following the layout from Fig. 3 (*see* **Note 25**). Gently rock the plates back and forth and from side to side.
12. Incubate the cells for 4 h at 27 °C, in a humidified incubator.
13. Gently add 0.4 mL of Sf-900™ II SFM insect medium containing 2% (v/v) FBS to each well (*see* **Note 25**). Incubate the cells at 27 °C in a static incubator for 3 days.

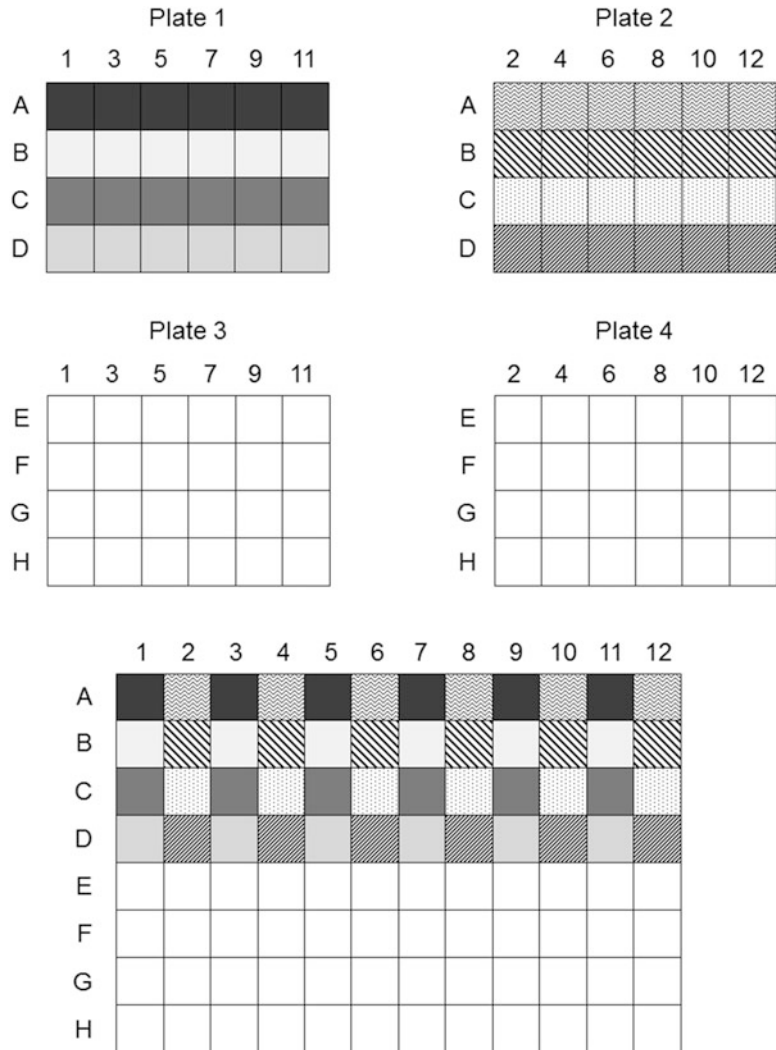


Fig. 3 The format for transferring samples between 24-well and 96-well blocks

14. Signs of infection should be seen in the transfected cells 2–3 days posttransfection, by comparing with the control cells under an inverted microscope. Confluent growth of cells will be seen in control wells, whereas areas of clearing will be prominent in wells with infected cells. Infected cells are usually larger and deformed or elongated compared to uninfected cells.
15. Harvest the viruses when the cells are well infected by transferring the liquid contents from the 24-well TC plate into a sterile 96-deep-well block (*see* Fig. 3 for layout) and centrifuging at $1,500 \times g$ for 20 min at RT. Collect the clear supernatant (<0.7 mL) in another sterile 96-deep-well block. This is the P0 baculovirus (BV) stock, which is stored at 4 °C, protected from light.

3.10 Virus Amplification and Test Expression

1. Using a 1 mL multichannel pipette, dispense 3 mL of Sf9 cells (in Sf-900™ II SFM medium, containing 2% (v/v) FBS, at a density of 2×10^6 cells/mL) into each well of four 24-deep-well blocks.
2. Following the layout shown in Fig. 3, infect the cells with 120 μ L of P0 BV stock (*see Note 26*) and incubate at 27 °C, with shaking at 450 rpm in a Glas-Col shaker for 66–72 h (i.e., set up late on day 1 and harvest early on day 4).
3. Pellet the cells by centrifugation at $1,500 \times g$ for 20 min and harvest the supernatant by pipetting into a 96-deep-well block in the BSC according to the layout shown in Fig. 3. Store as P1 BV stock at 4 °C in the dark.
4. Resuspend the pellets in 1 mL of Lysis buffer, supplemented with protease inhibitors, and store at -80 °C for test purification at later date (or preferably purify directly).

3.11 Test Purification

1. If frozen, thaw pellets in a water bath at RT, then sonicate on ice for 4 min (3 s on, 15 s off with 35% amplitude on a 750 watt sonicator) using a 24-head probe (check that the probe is level and all tips are in the liquid; after sonication check for clearing).
2. Remove 15 μ L of the total cell lysate into a 96-well PCR plate as the Total fraction, add 5 μ L of the 4 \times sample buffer, and store at 4 °C.
3. Transfer the remaining sample into a 96-deep-well block according to the layout shown in Fig. 3 and centrifuge at $3,000 \times g$ for 30 min at 4 °C.
4. Remove the clarified supernatant to a fresh 96-deep-well block using a multichannel pipette, taking care to avoid transferring any pelleted material (*see Note 27*).
5. Add 100 μ L of a previously washed and equilibrated 50% slurry (Ni-IDA or Ni-NTA) to each well using a multichannel pipette with cut tips, mixing well before each row (*see Note 28*).
6. Seal the block with a silicone mat and place another 96-deep-well block on top, tape together and incubate at 18 °C on a rotating wheel for 1 h, spinning at 10 rpm (*see Note 29*).
7. Centrifuge the block for 30 s at $200 \times g$ to remove the liquid from the lid and load the mixture on to a 96-well filter plate placed on top of a 96-deep-well waste collection block.
8. Allow the liquid to drip through the filter plate or centrifuge at $200 \times g$ for 1 min.
9. Add 800 μ L of Wash buffer to the resin block to wash out the remaining resin and then transfer to the corresponding wells of the filter plate. Allow the buffer to flow through or centrifuge briefly at $200 \times g$. Pour off the buffer from the waste block after this and all subsequent washing steps.

10. Add 800 μL of Wash buffer and allow the buffer to flow through or centrifuge briefly at $200 \times g$.
11. Repeat the wash step a further 3 times and after the final wash, spin the plate for 2 min at $300 \times g$ to remove any residual Wash buffer. Pour off Wash buffer from the waste block and spin for a further 1 min to remove all trace of Wash buffer (*see Note 30*).
12. Place the filter plate on top of a fresh 200 μL V-bottomed 96-well microtiter plate and add 50 μL of Elution buffer to each filter well.
13. Incubate at RT with shaking for 20 min, then centrifuge for 3 min at $300 \times g$ to collect the elution (Purified fraction).
14. In a 96-well PCR plate, mix 15 μL of each Purified fraction with 5 μL of $4\times$ sample buffer. Heat denature at 80°C for 10 min.
15. Prepare four SDS-PAGE precast gels by rinsing with water, adding $1\times$ MES buffer and rinsing the wells.
16. Using a multichannel pipette, load 15 μL of your samples onto the gels, note that samples will be interleaved (e.g., A1, B1, A2, B2, etc.). Also load 5 μL of a protein marker (e.g., SeeBlue[®] Plus2 Pre-Stained Standard) in the first lane of the gel.
17. Run the gel at 150 V for at least 1 h, or as long as required for the dye-front to reach the bottom of the gel.
18. Break open the cast and carefully remove the gel into a tray, rinse with water, and add half a cap full of InstantBlue[™]. Stain for ~ 1 h with shaking at RT.
19. Discard the stain and wash twice with water, taking care not to tear the gel. Leave in water with shaking to destain for as long as required.
20. Confirm the size of your protein of interest against the protein ladder (*see Note 31* and Fig. 4).

3.12 Virus Amplification

The volumes of P0 (0.7 mL) and P1 (3 mL) viruses generated as described previously are low in volume and insufficient to be used for large-scale expression experiments. Therefore, it is necessary to amplify the virus in a larger volume, typically to the scale of 50–100 mL. The virus can be stored at 4°C for months, but it is advisable to reamplify the virus, if stored at 4°C for a longer period of time. For virus amplification, insect cells are generally infected with low Multiplicity of Infection (MOI—number of virus particles per cell) to avoid generating noninfectious particles in the virus stocks. Use a healthy log phase culture of Sf9 cells with more than 95% viability. All of our virus stocks are made in Sf-900[™] II SFM, but other media formulations may work equally well.

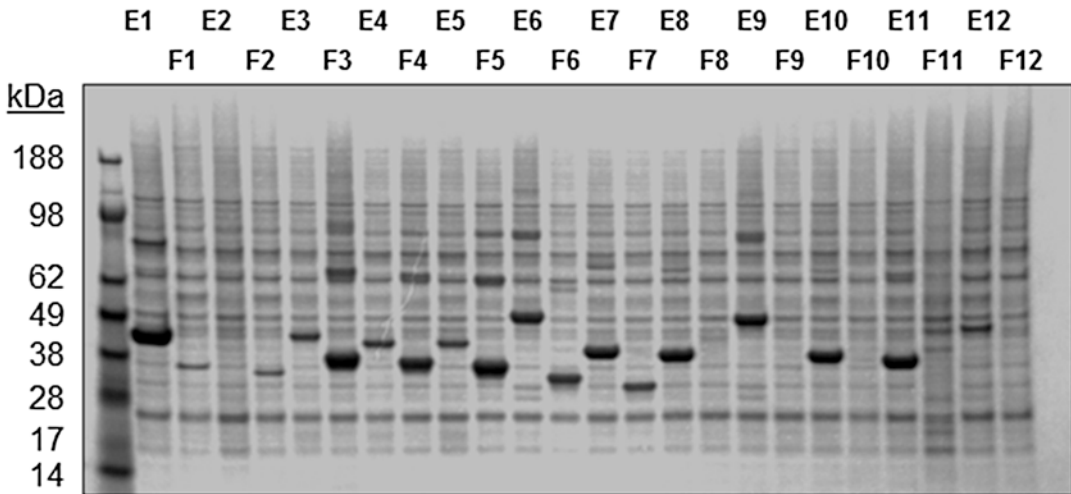


Fig. 4 Image showing the SDS-PAGE result of a test purification from insect cells. The gel shows a range of high, medium, and low expressions of various proteins of different molecular weights. Note that samples loaded using a multichannel pipette will be interleaved (e.g., A1, B1, A2, B2, etc.)

1. Take a sterile 250 mL or 500 mL flask and seed 50 mL of suspension-adapted Sf9 cells (2×10^6 cells/mL) in Sf-900™ II SFM.
2. Add FBS to the final concentration of 2% (*see Note 32*).
3. Add 100 μ L of the P1 BV stock to the cells and gently swirl the flask.
4. Transfer the flask to a 27 °C shaking incubator with shaking speed set at 100 rpm and incubate the flask for 72 h.
5. At 72 h postinfection take a small aliquot of cells and observe under the microscope for signs of infection (*see Note 33*) and absence of any form of microbial contamination.
6. Transfer the cells to a 50 mL tube and centrifuge at $900 \times g$ for 20 min.
7. Collect the supernatant into a fresh 50 mL tube and store at 4 °C. This represents P2 BV stock.
8. The cell pellet generated in the process of virus amplification can be utilized for protein purification using IMAC. Protein purified from this pellet can be used for any intended application. Moreover, this purification validates the ability of the virus stock to express protein.

3.13 Large-Scale Expression

This protocol is successfully applied for the expression of a broad range of proteins but for some proteins the expression time point, and MOI can be highly specific and will require optimization (*see Note 34*).

1. Seed log phase Sf9 cells to the density of 1×10^6 cells/mL in Insect-XPRESS or Sf-900™ II SFM. Keep the volume of culture to 1 L in a 3 L capacity flask. If more than 1 L scale-up is needed, use multiple 3 L flasks with 1 L culture volume in each (*see Note 35*).
2. Incubate flasks at 27 °C with shaking set at 100 rpm and allow the cells to grow for 24 h.
3. The next day, check the cell density using a hemocytometer and cell health and look for any signs of contamination. Cells should go through one doubling cycle in 24 h and the cell count should be $\sim 2 \times 10^6$ cell/mL.
4. Add 1.5–3.0 mL of P2 BV stock per L of the culture, swirl the flask gently, and transfer the culture to a 27 °C shaker-incubator set at 100 rpm (*see Note 36*).
5. Incubate the flask for 64–72 h.
6. Take a small sample of the infected culture and look under the microscope for signs of infection, but not lysis of the cells. In addition, look carefully for the absence of any bacterial or fungal contamination (*see Note 37*).
7. Take 3 mL out of the culture and centrifuge separately (at $900 \times g$ for 20 min) from the remaining culture for expression testing (*see Note 38*).
8. Without waiting for results from **step 7** above, transfer the remaining cells to 1 L centrifuge pots, balance pairwise and centrifuge at $900 \times g$ for 20 min using JLA 8.1000 rotor on Avanti J-20XP or Avanti J-26XP centrifuge (*see Note 39*).
9. Pour the supernatant to a waste container for decontamination using Virkon.
10. Resuspend the cell pellet obtained from 1 L of the culture in 25–30 mL of PBS by swirling and pipetting gently and transfer to 50 mL tubes.
11. Balance the tubes pairwise and centrifuge at $900 \times g$ for 20 min using a benchtop centrifuge.
12. Discard the PBS in a Virkon solution and proceed to purify the protein from the cell pellet or freeze the cell pellets at -80 °C for purification at a later date.

3.14 Protein Extraction

All the following steps of protein extraction and purification are performed at 4 °C or on ice. Prechill the buffers and centrifuges.

1. If protein purification is performed straight after harvesting the cells, transfer the cell pellets to ice or if the cells were frozen, thaw the pellets in a water bath set at RT or 37 °C. Do not leave pellets in the water bath for any longer than is required to thaw them and transfer onto ice immediately once thawed.

2. Resuspend the cells in 1 volume of ice cold $2\times$ Lysis buffer (1 mL per g wet-weight of cells) using a pipette and add additional Lysis buffer until the suspension is homogeneous.
3. Place the cell suspension container on ice. Set the amplitude to 35% on a 750 watt Sonics Vibra-Cell sonicator and sonicate with 10–15 bursts of 5 s on, 10 s off (*see Note 40*). Save 10 μ L of the lysate which represents the Total fraction.
4. Transfer the lysates to centrifuge tubes, balance the tubes pairwise, and centrifuge at $53,000 \times g$ using a JA-25.50 rotor for at least 30 min at 4 °C.
5. Transfer the clear supernatant into a clean tube taking care to avoid transferring any pelleted material. This clarified supernatant represents the Soluble fraction.

3.15 Large-Scale Protein Purification

The protein purification scheme for insect cells is similar to protein purification from *E. coli* as described in Chapter 4, Subheading 3.6. However, we recommend paying particular attention to the following points while purifying proteins from insect cells:

1. The buffer composition described here works for a diverse set of proteins but the buffer can be substituted to address issues such as protein instability and requirements of final applications. Careful optimization of the buffer composition with respect to the buffering system, pH, salt concentrations, and additives is particularly critical for difficult to purify proteins.
2. In comparison to *E. coli* cell lysates, insect cell lysates are denser because of higher background protein concentration. This can result in clogging of prepacked IMAC columns; therefore, we recommend doing manual IMAC using the gravity-flow procedure for purification of proteins from insect cells.
3. Often intrinsic proteins from insect cells copurify due to the affinity of exposed histidines or metal binding moieties of endogenous proteins toward the immobilized metal ions. Therefore, it is often the case that IMAC followed by SEC is not enough to obtain very pure protein from insect cells, which necessitates inclusion of additional purification steps such as ion exchange chromatography or tag cleavage and rebinding to IMAC.

3.16 Quality Assurance

If available, mass spectrometric analysis of every purified protein is highly recommended. This confirms the molecular weight of the protein, with mass discrepancies indicating mutations or cloning artifacts and potential posttranslational modifications. The protein is loaded into a small C3 HPLC column for desalting and eluted onto an in-line electrospray ionization time-of-flight analyzer. Any discrepancy needs to be explained, either by sequencing the DNA, by enzymatic removal of suspected modifications or by MS/MS analysis of proteolytic fragments.

4 Notes

1. The EMBacY backbone contains a constitutively expressing YFP expression cassette that allows for easy monitoring of viral titers via fluorescence without plaque assays.
2. X-gal does not produce sufficiently dark blue nonrecombinant colonies in our hands; therefore, we use Blue-gal instead. The plates can be stored for up to 1 month at 4 °C, covered with foil to prevent exposure to light.
3. Be careful not to splash the cells against the sides of the wells while using the repeat pipettor and check that the liquid is at the bottom of the well before continuing. This step can also be done using a single channel pipette but will take more time.
4. When there are no colonies, plate 50 μ L of undiluted culture instead.
5. This step can be performed at RT on the bench over the weekend if necessary.
6. One 96-well block should provide sufficient bacmid DNA for transfection. However, we find it useful to set up two blocks to provide a balance for the centrifugation step.
7. We only use the reagents from the Montage Plasmid Mini-prep_{HTS} 96 Kit for purifying the recombinant bacmid DNA, not the filter plates. The reagents can also be purchased from Merck individually.
8. Covering the block with an adhesive tape pad or alternative will result in leaking and cross-contamination of wells. Make sure the silicone sealing mats are suitable for either round or square 96-deep-well blocks, depending on which 96-well blocks you use.
9. This second centrifugation step is important to remove as much of the insoluble pelleted material as possible in order to obtain clean bacmid DNA at the end of the prep.
10. It is recommended not to remove all of the supernatant to avoid transferring insoluble material.
11. Incubation can also be done overnight at 4 °C and will result in a higher yield of bacmid DNA but is not necessary.
12. If you have more than 1 block, be careful not to remove the marker labels when using 70% (v/v) ethanol.
13. Do not allow the pellets to dry out completely.
14. The bacmid DNA is very fragile so mix it gently, do not over-pipette. If the concentration of the DNA is less than 0.5 μ g/ μ L, use up to 5 μ L.

15. High concentrations of bacmid DNA will inhibit the bacmid PCR screen so we dilute the bacmid prior to addition. Where the yields of bacmid are low it may be necessary to use a lower dilution instead.
16. All cell culture steps must be performed under aseptic conditions in a BSC, making sure that sterility is maintained throughout the procedures. To keep the cultures free from contamination by bacteria, yeast, fungi and viruses, it is crucially important to keep the benches, BSC and incubators clean. Use 70% (v/v) ethanol to wipe the cabinet before and after use, also wipe the outside of media bottles, pipettors, flasks and other containers with 70% (v/v) ethanol before transferring them into the cabinet. Wear clean lab coats and gloves and wash hands before and after working with cell culture. Any spillage inside the BSC, incubators, and so on should also be cleaned immediately with 70% (v/v) ethanol or MicroSol. Use separate media bottles for general cell culture maintenance and for virus work. We recommend adding penicillin and streptomycin to the final concentration of 50 units/mL and 50 $\mu\text{g}/\text{mL}$ respectively to the cell culture media to prevent bacterial contamination during culture growth.
17. Always wear protective clothing (lab coat, gloves, and safety specs) when thawing vials containing frozen cells as they sometimes explode on contact with the water. Do not dilute cells below $1 \times 10^6/\text{mL}$. Final DMSO concentration in suspension should not exceed 0.5%.
18. The % Cell viability is calculated by counting the number of viable cells and also the number of total cells on the hemocytometer grid. Viable cells do not take up Trypan Blue Stain; however, nonviable cells take up the stain and appear blue under the microscope. To determine cell viability, mix 0.1 mL of Trypan Blue Stain with 1 mL of cell suspension and load a hemocytometer. Count the number of blue-stained cells and the total number of cells and then calculate the number of viable cells per mL and correct for the dilution factor. Cell viability should be at least 95% for a healthy log phase culture before it can be used for transfection, virus amplification or protein expression.
19. For better aeration of the cells, it is important to keep the culture volume between 25% and 35% of the total volume capacity of shake flask and shaking between 90 and 105 rpm. Cells form clumps initially but cells should start growing in single cell suspension within a week or so.
20. Cells can be transferred gradually to 1 L and then 3 L flasks, keeping the culture volume between 25% and 35% of the total volume capacity of shake flask. Ideally do not allow the cell

density to exceed 6×10^6 cells/mL or fall below 0.7×10^6 cells/mL. Cell growth may slow down if diluted to the density of less than 0.7×10^6 cells/mL. Cells should not be diluted by more than 1 in 5.

21. If a freezing container is not available, vials can be transferred to a -20 °C freezer for 2–3 h followed by transfer to -80 °C overnight.
22. It is not necessary to keep Chemgene solution in flasks for more than 20 min. Leaving Chemgene for longer may make it difficult to remove the traces from flasks. Glass flasks are easier to clean than the polycarbonate flasks. Polycarbonate flasks for suspension culture are meant to be disposable but they can be reused several times if cleaned properly after treatment.
23. To keep it cost effective and to express most of our recombinant proteins, we tested and compared a range of transfection reagents and decided to use Insect GeneJuice[®].
The Insect GeneJuice[®]- and cell-only controls are important for determining the success of the transfection as they allow the user to distinguish cytotoxic effects and uninfected cells from infected cells.
24. Cell attachment can be observed using an inverted microscope by focusing through the sample; the cells should be visible in one plane of view once successfully attached.
25. Pipette the mixture gently and avoid touching the bottom of the plate so as to not disturb the cells.
26. For some targets it may be necessary to use the P1 virus to infect for test expression. However, we have found that there is little difference in the yields when expressing from P1 rather than P0. We therefore use P0 virus, which shortens the expression process by at least 3 days.
27. To avoid disturbing the Insoluble fraction, tilt the plate and drive the tips down the side of the wells at an angle. Stop just above the pellet, on most plates there is a ridge just off the bottom—feel for this with the tips. Gently pipette up the supernatant and then transfer to the new plate. Do not go back into the wells as this will resuspend the pellets; if this happens then respin the sample and try again.
28. The resin tends to clump and settles quickly. We recommend using 200 μ L tips with ~ 5 mm cut from the ends to prevent clogging the tips and ensure even loading. Also, continually mix the resin by pipetting up and down as well as shake the reservoir from side to side to prevent settling.
29. When the silicone matting seal is pressed down firmly and held in place with another deep-well block, the block will not leak

when placed on its side. If you prefer you can incubate the plate upright, but the resin tends not to mix as well when done this way; we would therefore recommend keeping the samples in a 24-well format for this step, as this provides greater surface area for binding.

30. Removing all trace of Wash buffer is essential to ensure that the subsequent elution step does not become diluted with Wash buffer.
31. It is beneficial to grade the expression level of your proteins to more easily identify ones that you may wish to scale up. At this point we also recommend confirming the targets using quality control steps such as intact mass (if quantities are sufficient) or by in-gel tryptic digest MSMS analysis.
32. Baculovirus stability is known to improve in the presence of FBS. As Sf-900™ II SFM is a serum-free and protein-free medium, addition of FBS to the final concentration of 2% is recommended to stabilize the virus and maintain its infectivity when it is stored at 4 °C.
33. Signs of baculovirus infection: baculovirus infected insect cells look swollen, nuclei appear to fill the cells and the cells do not show any clumps when compared to a healthy cell control. If the cells are in very late phase of infection, they will start to lyse.
34. Availability of healthy viable cells is very important for successful scale up of a broad range of targets. Culture conditions such as temperature, pH, dissolved oxygen, osmolality, and nutrient composition of the culture medium can influence the infection of the insect cells. In addition, factors such as cell line, expression time point, MOI, and cell density at the time of infection can have significant effects on protein expression in insect cells. This protocol is generically applied to large number of proteins; however, occasionally for some proteins, optimization at protein expression level is necessary to improve the results. Optimization experiments should be performed on a small scale initially and can be later applied to large-scale expressions. The following conditions could be tested for expression optimization: range of MOI, two harvesting time points (48 and 72 h), two cell lines (Sf9 and High Five), or different cell densities (2×10^6 cells/mL and 4×10^6 cells/mL). It should be noted that baculoviruses are lytic viruses for insect cells and will eventually lyse the cells if left long enough after infection. This also means that a harvesting time of 48 or 72 h is also determined by the volume of virus added. The cells can be infected with low MOI (0.05–0.3 pfu/cell) and harvested at 72 h or they can be infected with a high MOI (>1 pfu/cell) and harvested at 48 h. Cells infected with high MOI and harvested at 72 h may show significant lysis.

35. Before diluting the cells, check for the health of the cells and absence of any signs of infection or contamination under a microscope. If less than 1 L scale-up is enough, smaller flasks should be used. However, remember to use a culture volume of only 25–35% of the total volume capacity of the flask.
36. The amount of virus added is determined by the titer of virus stock. We do not routinely measure viral titers but various methods for baculovirus titration have been developed based on cell viability, plaque formation, antibody-based assays, and so on [14]. For the 72 h expression time point, we recommend an MOI of 0.05–0.3 pfu/cell. If the titer of virus stock is 1×10^8 pfu/mL and 2 mL of virus is added to 1 L of the cells (total of 2×10^9 cells), that would be an MOI of 0.1. Addition of more virus can affect the expression and can also cause cell lysis.
37. It should be noted that good signs of infection are desirable but more than 10% lysis of cells can be detrimental to the protein purification.
38. This small volume of cells can be used for expression testing before committing to purify a large batch of cells. This can give a quick estimate of protein expression levels or any failure of the batch to express the protein of interest. To purify the protein from 3 mL of culture, follow the protocol as described in Subheading 3.11.
39. Sf9 cells become very fragile after infection and can rupture if centrifuged at very high speed resulting in loss of protein in the medium itself. We recommend harvesting the cells by centrifugation at $900 \times g$ for 20 min and handling cell pellets gently.
40. Sonication time may need to be adjusted depending on volume of the cell suspension. Avoid excessive foaming and heating of the suspension by adjusting the instrument settings and keeping the cell suspension on ice all the time to reduce the potential for protein precipitation or denaturation. Cell disruption by sonication can also help in reducing viscosity by shearing nucleic acids.

Acknowledgments

We would like to thank all the SGC scientists (past and present) who contributed toward the development of the method. The SGC is a registered charity (number 1097737) that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genome Canada, Innovative Medicines Initiative (EU/EFPIA), Janssen, Merck KGaA, MSD, Novartis Pharma AG, Ontario Ministry of

Economic Development and Innovation, Pfizer, São Paulo Research Foundation-FAPESP, Takeda, and Wellcome. The BacMam vector backbone (pHTBV1.1) was kindly provided by Professor Frederick Boyce (Massachusetts General Hospital, Cambridge, MA).

References

1. Ignoffo CM (1975) Entomopathogens as insecticides. *Environ Lett* 8(1):23–40
2. Invitrogen (2002) Guide to Baculovirus expression vector systems (BEVS) and insect cell culture techniques. Invitrogen Life Technologies, Carlsbad
3. Kost TA, Condreay JP, Jarvis DL (2005) Baculovirus as versatile vectors for protein expression in insect and mammalian cells. *Nat Biotechnol* 23(5):567–575
4. Summers MD, Anderson DL (1972) Granulosis virus deoxyribonucleic acid: a closed, double-stranded molecule. *J Virol* 9(4):710–713
5. Matthews REF (1982) Classification and nomenclature of viruses. *Intervirology* 17(1–3):1–181
6. O'Reilly D, Miller L, Luckow V (1992) Baculovirus expression vectors: a laboratory manual. Oxford University Press, New York
7. Vaughn JL, Goodwin RH, Tompkins GJ et al (1977) The establishment of two cell lines from the insect *Spodoptera frugiperda* (Lepidoptera; Noctuidae). *In Vitro* 13(4):213–217
8. Granados RR, Guoxun L, Derksen ACG et al (1994) A new insect cell line from *Trichoplusia ni* (BTI-Tn-5B1-4) susceptible to *Trichoplusia ni* single enveloped nuclear polyhedrosis virus. *J Invertebr Pathol* 64(3):260–266
9. Smith GE, Summers MD, Fraser MJ (1983) Production of human beta interferon in insect cells infected with a baculovirus expression vector. *Mol Cell Biol* 3(12):2156–2165
10. Invitrogen (2010) Bac-to-Bac[®] Baculovirus expression system. Invitrogen Life Technologies, Carlsbad
11. Vijayachandran LS, Viola C, Garzoni F et al (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. *J Struct Biol* 175(2):198–208
12. Shrestha B, Smee C, Gileadi O (2008) Baculovirus expression vector system: an emerging host for high-throughput eukaryotic protein expression. *Methods Mol Biol* 439:269
13. Invitrogen (2010) Growth and maintenance of insect cell lines. Invitrogen Life Technologies, Carlsbad
14. Roldao A, Oliveira R, Carrondo MJ et al (2009) Error assessment in recombinant baculovirus titration: evaluation of different methods. *J Virol Methods* 159(1):69–80



Expression Screening of Human Integral Membrane Proteins Using BacMam

Pravin Mahajan, Katherine Ellis, Shubhashish Mukhopadhyay, Alejandra Fernandez-Cid, Gamma Chi, Henry Man, Katharina L. Dürr, and Nicola A. Burgess-Brown

Abstract

This chapter describes the step-by-step methods employed by the Structural Genomics Consortium (SGC) for screening and producing proteins in the BacMam system. This eukaryotic expression system was selected and a screening process established in 2016 to enable production of highly challenging human integral membrane proteins (IMPs), which are a significant component of our target list. Here, we discuss our recently developed platform for identifying expression and monodispersity of IMPs from 3 mL of HEK293 cells.

Key words Insect cells, Mammalian cells, Baculovirus, BacMam, Transduction, Protein purification, IMAC, SEC chromatography, Gel filtration, FSEC

1 Introduction

Baculovirus vector-mediated gene transfer in mammalian cells was first described in 1995 [1] and in 1996 [2]. Since then its entry into different types of vertebrate cells has been confirmed [3–7] and various constitutively active promoters have been used to drive recombinant gene expression in mammalian cells. Attempts to improve gene delivery by means of vector engineering and transduction methods have been evolving. The initial discovery has prompted exploration of this gene delivery tool for many uses including cancer gene therapy [8–10], heart [11], cartilage [12, 13], and bone [14] tissue engineering, vaccination [14, 15], modification of stem cells for therapeutic applications [16, 17], VLP production [18], eukaryotic protein display [19], and cell-based assays [20–22].

Recently, BacMam started gaining traction as a tool for large-scale protein production for structural or functional studies as well.

This has particularly been the case for integral membrane proteins (IMPs), which account for roughly a quarter of functional human proteins [23, 24] and half of drug targets [25] in humans. Because IMPs have transmembrane domains which transverse the hydrophobic lipid bilayer, they require unique sets of chaperones and posttranslational modifications for proper protein folding and stability. Moreover, lipid composition varies between cell types and organelles [26–28] and the environments across the membrane can differ [27, 28], presenting additional challenges for the stability of IMPs. For these reasons, expression systems as close to the source organism as possible are often required for large-scale expression of IMPs. In addition, it is preferable for the system's expression level to be as high as possible, since IMPs tend to have lower yields than soluble proteins and the reagents for purification are much more expensive (due to detergents). BacMam system addresses these challenges by using mammalian cells as expression hosts, hence maintaining the benefits of Baculovirus Expression Vector System (BEVS) while providing even more native environments for delicate proteins [29, 30].

BacMam also has advantages over competing systems for protein expression in mammalian cells. Baculoviruses cannot replicate in human cells [29, 30], making it a safer option than the lentivirus system as well as the transduced cells being less susceptible to the virus-mediated lysis. BacMam has higher infectivity than transient transfection and is less constrained by the size of genes than either of the systems [29]. Some notable publications on the use of baculoviruses for large-scale expression in mammalian cells are described for secreted proteases [31], for recombinant soluble and membrane glycoproteins in suspension culture [32], for a ligand-gated ion channel [33, 34], and for the ABC transporter CFTR [35].

In this chapter, we describe recently developed methods for expression screening and scaling up of the production of human IMPs using BacMam. To describe our series of standardized protocols for protein production in mammalian cells, this chapter is divided into the following stages: (a) transposition, bacmid production, and PCR screen; (b) growth and maintenance of mammalian cell lines in adherent and suspension culture; (c) transfection into Sf9 cells, baculovirus generation, and small-scale transduction/purification; and (d) fluorescence size exclusion chromatography (FSEC) screening. The screening process has been miniaturized to 24-well format. The steps involved in the pipeline from cloning to large-scale expression are outlined in Fig. 1.

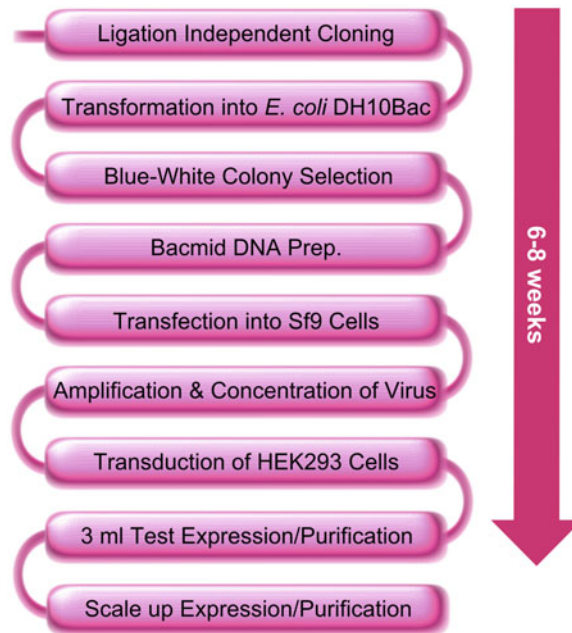


Fig. 1 Overview of the BacMam expression process. The process takes ~6–8 weeks from LIC to scale-up

2 Materials

Unless otherwise stated, all solutions are prepared using ultrapure water (prepared by purifying deionized water to reach a resistivity of 18.2 M Ω cm at 25 °C) and analytical grade reagents.

2.1 Transposition, Bacmid Preparation, and Transfection

All materials are listed in Chapter 5, Subheadings 2.1 and 2.2.

2.2 Virus Amplification, Transduction, and Test Expression in Expi293FTM and HEK293S GnTI⁻ Cells

The following reagents, consumables, and equipment are required in addition to those listed in Chapter 5, Subheadings 2.1 and 2.2.

1. Cell lines: Expi293FTM cells (Thermo Fisher Scientific); HEK293S GnTI⁻ (ATCC[®] CRL-3022TM) or Expi293FTM GnTI⁻ (Thermo Fisher Scientific).
2. Media: FreeStyleTM 293 Expression Medium (Thermo Fisher Scientific).
3. 1 M sodium butyrate: In a fume hood, dissolve 55 g of sodium butyrate in PBS, make up to 500 mL and filter through a 0.2 μ m syringe filter. Store at 4 °C (short term) and -20 °C (long term).
4. Dulbecco's PBS (Ca-Mg free).

5. 20% polyethylene glycol (PEG) 10,000: Dissolve 200 g of PEG 10,000 and 12 g of NaCl in 600 mL of water, mix, and make up to 1 L. Autoclave and store at room temperature (RT) (*see Note 1*).
6. Roller bottles, 2 L (Avidity Science).
7. Fume hood.
8. CO₂ incubator (Panasonic Sanyo CO₂ incubator, or similar).
9. Celltron shaker (Infors HT).
10. Class II Microbiological Safety Cabinet (ScanLAF-Mars600/1200 or similar).
11. Dimethyl sulfoxide (DMSO), Molecular Biology grade (DNase/RNase free).
12. Infors Multitron Cell CO₂ shaker incubator (25 mm throw).

2.3 Test Purification

The following reagents, consumables, and equipment are required in addition to those listed above.

1. Complete EDTA-free tablet (Roche).
2. 0.5 M Tris(2-carboxyethyl)phosphine (TCEP): Prepare in water, filter through a 0.2 μ m syringe filter, and store at -20°C .
3. 1 M dithiothreitol (DTT): Prepare in water, filter through a 0.2 μ m syringe filter, and store in 1 mL aliquots at -20°C .
4. SeeBlue[®] Plus2 Pre-Stained Standard (Invitrogen) or Precision Plus Prestained Protein[™] Standards (Bio-Rad or similar).
5. InstantBlue[™] (Expedeon Protein Solutions).
6. 20 \times NuPAGE[™] MES SDS Running Buffer (Invitrogen).
7. PBS: Dissolve 5 tablets of PBS in 1 L of water, filter through a 0.2 μ m membrane filter, and store at 4°C .
8. 1 M HEPES, pH 7.5: Prepare in water, filter through a 0.2 μ m membrane filter, and store at RT.
9. 5 M NaCl: Prepare in water, filter through a 0.2 μ m membrane filter, and store at RT.
10. 50% (v/v) glycerol: Autoclave and store at RT.
11. 10% n-dodecyl- β -D-maltoside (DDM) (Anatrace) and 1% cholesteryl hemisuccinate (CHS) (Merck KGaA): Dissolve 5 g of DDM in 50 mL of water in a Falcon tube and add 0.5 g of CHS into the DDM solution. Filter through a 0.2 μ m membrane filter and store at -20°C (*see Note 2*).
12. Lysis buffer (1 L): 50 mM HEPES, pH 7.5, 300 mM NaCl, and 5% (v/v) glycerol prepared in advance, filtered through a 0.2 μ m membrane filter and stored at 4°C . On the day of purification, add 1 μ L/mL Protease inhibitor cocktail, 0.5 mM TCEP, and 1% DDM–0.1% CHS.

13. Wash buffer (1 L): 50 mM HEPES, pH 7.5, 300 mM NaCl, and 5% (v/v) glycerol prepared in advance, filtered through a 0.2 μm membrane filter and stored at 4 °C. Add 0.5 mM TCEP and 0.03% DDM–0.003% CHS on the day of purification.
14. Elution buffer (0.1 L): 50 mM HEPES, pH 7.5, 300 mM NaCl and 5% (v/v) glycerol prepared in advance, filtered through a 0.2 μm membrane filter and stored at 4 °C. Add 0.5 mM TCEP, 0.03% DDM–0.003% CHS, and 50 mM biotin (*see Note 3*) on the day of purification.
15. Affinity buffer: 50 mM HEPES, pH 7.5, 500 mM NaCl, and 10% glycerol, prepared in advance. Filter through a 0.2 μm membrane filter and store at 4 °C.
16. 50% (w/v) Strep-Tactin[®] XT resin (IBA Lifesciences): To equilibrate, wash the resin twice in water and then three times in Affinity buffer in a 50 mL tube, by inverting to resuspend the resin and centrifuging at $500 \times g$ for 1 min. After the final wash, resuspend the resin in Affinity buffer as 50% (w/v) slurry and store at 4 °C when not in use.
17. SB: Prepare a stock of NuPAGE LDS sample buffer (Invitrogen) containing DTT (1:4 dilution of 1 M DTT in NuPAGE LDS sample buffer) and store at –20 °C.
18. 96-well filter plates.
19. 96-well plate, 0.5 mL, round wells, U-shaped, polypropylene, 14 mm (Agilent).
20. pH Strips.
21. Square-shaped silicone mat (AxyMat[™] VWR).
22. Precast 26-Lane SDS-PAGE gradient gels (4–12% Bis-Tris) (Invitrogen).
23. Protein gel electrophoresis apparatus (Invitrogen or similar).
24. Vibra-Cell Sonicator with 24-well probe (Sonics[®]).
25. General purpose benchtop centrifuge (Sorvall Legend RT, Kendro).

**2.4 Dionex
Fluorescence Size
Exclusion
Chromatography
(FSEC) Screening
(Fig. 2)**

1. Dionex[™] Ultimate 3000 UHPLC (Thermo Scientific[™]). Equipped with the following modules: SR-3000 solvent rack, LPG-3400RS pump, WPS-3000TBRS autosampler, TCC-3000SD Thermostatted column compartment, MWD-3000RS multi wavelength UV-VIS detector, FLD-3400RS multiwavelength fluorescence detector, and VF-F11-A-01 fraction collector. The PC associated with the equipment has Chromeleon 7 software installed for interfacing and data analysis.
2. Zenix[™]-C SEC 300 column, particle size: 300 Å, ID \times length: 7.8 \times 300 mm (Sepax Technologies).

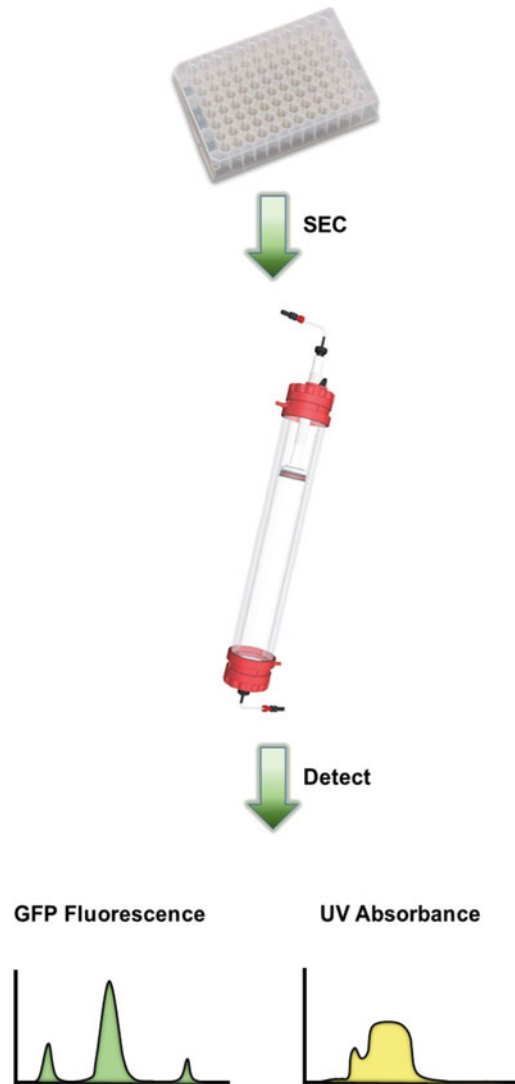


Fig. 2 Diagram describing the FSEC screening process. The 96-well elutions from the Strep-Tactin[®]XT resin purification are loaded onto the Dionex system using an autosampler and separated on a Zenix[™]-C SEC 300 column. Fluorescence measurements are recorded for both tryptophan and GFP

3. Zenix[™]-C SEC 300 guard column, particle size: 300 Å, ID × length: 7.8 × 50 mm (Sepax Technologies).
4. SEC buffer (2 L): 50 mM HEPES, pH 7.5, 150 mM NaCl prepared in advance. Filter through a 0.2 μm membrane filter and store at 4 °C. Add 0.03% DDM–0.003% CHS on the day of purification.

3 Methods

3.1 Transposition, Bacmid Production, and Bacmid PCR Screen

The transposition process, bacmid production, and bacmid PCR screen methods are described in Chapter 5, Subheadings 3.1, 3.2 and 3.3 respectively and do not differ for the BacMam process with the only exception being the specific bacmid PCR screening primers which are listed in Table 1.

3.2 Growth and Maintenance of Expi293F™ and HEK293S GnTI– Cell Lines

1. The Expi293F™ and HEK293S GnTI– mammalian cell lines are maintained in suspension culture adapted to FreeStyle™ 293 medium.
2. The HEK293S GnTI– cells are suspension-adapted by growing them initially as per ATCC's protocol in DMEM-F12 medium in T-flasks, which is then serially adapted into FreeStyle™ and transferred to vented baffled shaker flasks. We grow the cells to densities of $3\text{--}4 \times 10^6$ cells/mL (*see Note 4*). The conditions of growth for suspension cells are 8% CO₂, 75% humidity, and 37 °C, and speed of shaking varies from flask to flask (*see Notes 5 and 6*).
3. Once Expi293F™ cells have reached a density of 2×10^6 cells/mL, split to 0.2×10^6 cells/mL and maintain at densities of $0.2\text{--}2 \times 10^6$ cells/mL for up to 6–8 weeks.
4. For protein expression, cells can be grown to higher densities, but should be maintained at $0.2\text{--}2 \times 10^6$ cells/mL. HEK293S GnTI– cells should be maintained at densities of $0.3\text{--}2 \times 10^6$ cells/mL for 6–8 weeks, growing to higher densities for protein expression.

3.3 Reviving Expi293F™ and HEK293S GnTI– Cell Lines from Frozen Stocks

1. Warm FreeStyle™ 293 medium to 37 °C in a water bath and pipette 50 mL of the medium into a 500 mL sterile polycarbonate shake flask (*see Note 4*).
2. Remove a cryo-vial containing the Expi293F™ or HEK293S GnTI– cells ($1\text{--}2 \times 10^7$ cells/mL) from liquid nitrogen and rapidly thaw in a plastic beaker of water at 37 °C (not in a water bath).

Table 1
Primers used to confirm correct insertions at the bacmid PCR screen stage

Primer name	Primer sequence
pFBM-fwd	CAAATGTCGTAACAACCTCCGC
pFBM-rev	TAGTTAAGAATAACCAGTCAATCTTTCAC

3. Thaw the cells to approximately 70%, and transfer using a 5 mL stripette into the 500 mL flask containing 50 mL of the medium.
4. Gently mix the cell suspension and transfer the flask to a 37 °C CO₂ shaker incubator, with shaking set at 100 rpm.
5. Check the cells every day after thawing until they start doubling every 24 h for good health, then dilute the cells to 0.5×10^6 cells/mL when the density reaches 2×10^6 cells/mL.
6. Always ensure that viability is $\geq 90\%$ by counting the cells on a hemocytometer under the microscope and maintain the cells in mid-logarithmic phase (density between 0.2 and 2.5×10^6 cells/mL), changing to a fresh flask after 7–10 days (*see Note 4*).
7. Revive a new vial of cells from frozen stocks every 6–8 weeks and discard the old cells.

3.4 Suspension Culture of Expi293F™ and HEK293S GnTI– in Shake Flask

1. After growing a sufficient number of cells as described above, determine the viable cell count using Trypan Blue Stain and a hemocytometer (*see Note 7*).
2. Seed the cells to a density of 0.5×10^6 cells/mL into a 500 mL non-baffled polycarbonate flask in FreeStyle™ 293 medium.
3. Incubate the flask at 37 °C with shaking set at 90–105 rpm (*see Note 6*).
4. When the cells reach a density of 2×10^6 cells/mL, dilute them back with medium to 0.2×10^6 cells/mL and expand the cell volume depending on requirement of the cells (*see Note 4*).

3.5 Cell Freezing

Once the cells start doubling regularly after reviving them, it is advisable to freeze down the low passage number cells in several cryo-vials.

1. Label sterile 2-mL internal thread cryo-vials with the name of the cell line, date of freezing, and density of cells, and store the vials at 4 °C until ready to use.
2. Take a small sample of cells from a shake flask and count viable cells using a hemocytometer. Alternatively, cells from adherent cultures can be used for freezing.
3. Take the required volume of mid-logarithmic phase cells with viability $\geq 90\%$ so that $1\text{--}3 \times 10^7$ cells are available for each vial to be prepared for storage.
4. Harvest cells by centrifugation at $200 \times g$ for 10 min and keep the supernatant (which is the conditioned medium), to prepare cryopreservation medium.
5. Prepare cryopreservation medium containing 92.5% (v/v) FreeStyle™ 293 Expression medium (50:50 ratio of fresh media to conditioned media) and 7.5% DMSO.

6. Resuspend the cell pellet in the predetermined volume of cryopreservation medium.
7. Rapidly dispense 1 mL aliquots of this cell suspension into prelabeled 2-mL cryovials.
8. Transfer the vials into a Mr. Frosty freezing container and keep it in a -80°C freezer for 24–72 h. Transfer the vials to a liquid nitrogen Dewar.

3.6 Decontamination and Cleaning of Shake Flasks

Decontamination of flasks is described in Chapter 5, Subheading 3.8.

3.7 Transfection into Sf9 Cells

This method is identical for insect and mammalian cells so please follow Chapter 5, Subheading 3.9.

3.8 Test Expression (Transduction) in Expi293F™ or HEK293S GnTI– Cells

1. Dispense 75 μL of PEG 10,000 solution to each well of four 24-deep-well blocks.
2. Add 300 μL of PI virus (harvested in the previous section; *see Note 8*) into the corresponding wells of the 24-deep-well blocks containing the PEG 10,000 solution following the layout shown in Fig. 3.
3. Cover with a sticky seal and incubate at 18°C and 300 rpm for 5 min in a Glas-Col shaker.
4. Store the deep-well blocks for up to 1 week (ideally overnight) at 4°C .
5. The following morning, incubate the 24-deep-well blocks in a Glas-Col set at 18°C and 300 rpm for 30 min.
6. Centrifuge the deep-well blocks at $3,000 \times g$ for 45 min.
7. Inside a microbiological safety cabinet, carefully pipette the supernatant out and discard. The concentrated virus (translucent-looking pellet) is settled at the bottom of the blocks.
8. Prepare the required amount of Expi293F™ or HEK293S GnTI– cells at 2×10^6 cells/mL supplemented with 5 mM sodium butyrate.
9. Using a 1 mL 12-multichannel pipette, dispense 3 mL of Expi293F™ or HEK293S GnTI– cells in Expi293F™ medium, at a density of 2×10^6 cells/mL into each well of four 24-deep-well blocks.
10. Incubate at 37°C on a Celltron shaker at 200 rpm, placed in a CO_2 incubator for 48 h (*see Note 9*).
11. Pellet the cells by centrifugation at $900 \times g$ for 20 min.
12. Wash the pellets with 1 mL of cold PBS, resuspending the cells slowly to avoid damaging them.

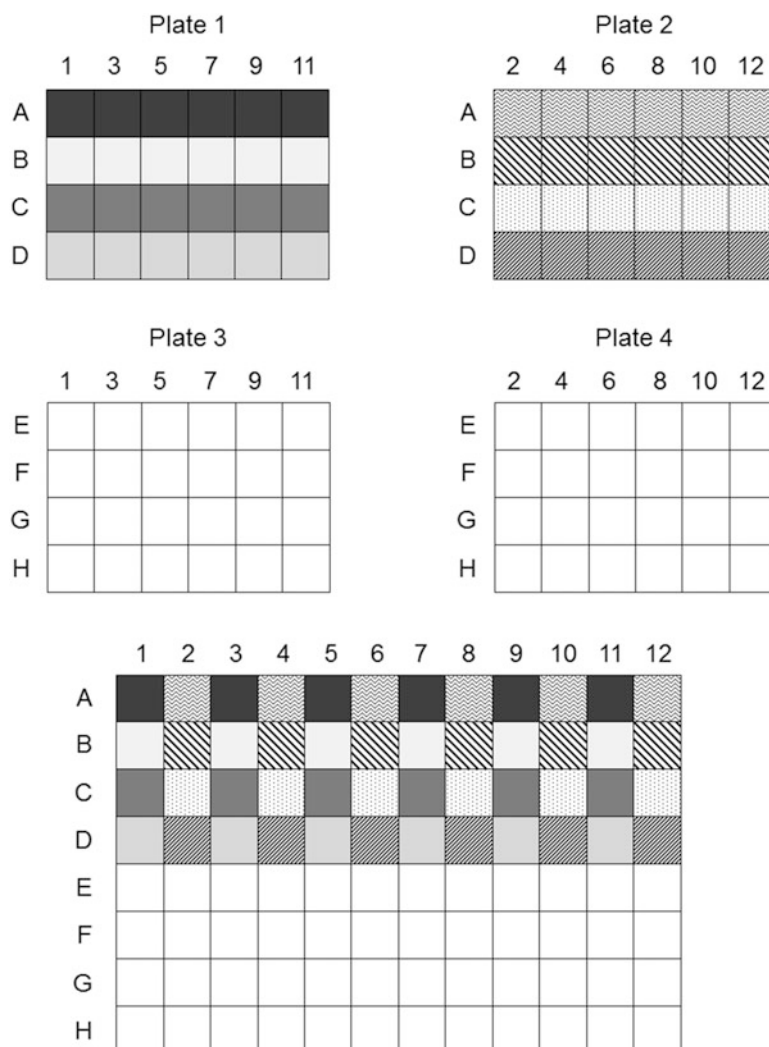


Fig. 3 The format for transferring samples between 24-well and 96-well blocks

13. Spin the cells down at $900 \times g$ for 20 min.
14. Store pellets at -80°C or proceed to the next step.

3.9 Test Purification

1. If frozen, thaw the pellets in a water bath at RT, then resuspend the pellets in 1 mL of Lysis buffer.
2. Sonicate on ice for 4 min (3 s on, 15 s off with 35% amplitude on a 750 watt sonicator) using a 24-head probe (checking that the probe is level and all tips are in the liquid; after sonication check for clearing).
3. Transfer the lysate into a 96-deep-well block following the layout shown in Fig. 3. Add $100\ \mu\text{L}$ of the DDM-CHS stock to each well, seal with a square-shaped silicone mat and rotate gently in the cold room for 1 h (*see Note 10*).

4. Remove 15 μL of the total cell lysate into a 96-well PCR plate as the Total fraction, add 5 μL of the 4 \times sample buffer, and store at 4 $^{\circ}\text{C}$ (*see Note 11*).
5. Centrifuge the remaining sample at 3,000 $\times g$ for 30 min at 4 $^{\circ}\text{C}$.
6. Remove the clarified supernatant to a fresh 96-deep-well block using a multi-channel pipette, taking care to avoid transferring any pelleted material (*see Note 12*).
7. Add 100 μL of a previously washed and equilibrated 50% slurry (Strep-Tactin[®]XT resin) to each well using a multichannel pipette with cut tips, mixing well before each row (*see Note 13*).
8. Seal the block with a silicone mat and place another 96-deep-well block on top, tape together and incubate at 4 $^{\circ}\text{C}$ on a rotating wheel for 1 h, spinning at 10 rpm (*see Note 14*).
9. Centrifuge the block for 30 s at 200 $\times g$ to remove the liquid from the lid and load the mixture on to a 96-well filter plate placed on top of a 96-deep-well waste collection block.
10. Allow the liquid to drip through the filter plate or centrifuge at 200 $\times g$ for 1 min.
11. Add 800 μL of Wash buffer to the resin block to wash out the remaining resin and then transfer to the corresponding wells of the filter plate. Allow the buffer to flow through or centrifuge briefly at 200 $\times g$. Pour off the buffer from the waste block after this and all subsequent washing steps.
12. Add 800 μL of Wash buffer and allow the buffer to flow through or centrifuge briefly at 200 $\times g$.
13. Repeat the wash step a further 3 times and after the final wash, spin the plate for 2 min at 300 $\times g$ to remove any residual Wash buffer. Pour off the Wash buffer from the waste block and spin for a further 1 min to remove all trace of Wash buffer (*see Note 15*).
14. Place the filter plate on top of a fresh 200 μL 96-well (U-shaped) plate and add 50 μL of Elution buffer to each filter well.
15. Incubate at RT for 20 min, then centrifuge for 3 min at 300 $\times g$ to collect the elution (Purified fraction).
16. In a 96-well PCR plate, mix 15 μL of each Purified fraction with 5 μL of 4 \times sample buffer (*see Note 11*).
17. Prepare four SDS-PAGE precast gels by rinsing with water, adding 1 \times MES buffer and rinsing the wells.
18. Using a multichannel pipette load 15 μL of your samples onto the gels, note that samples will be interleaved (e.g., A1, B1, A2, B2, etc.). Also load 5 μL of the Precision Plus Prestained Protein[™] Standards protein ladder in one lane of the gel.
19. Run the gel at 150 V for at least 1 h, or as long as required for the dye-front to reach the bottom of the gel (*see Note 16*).

20. Break open the cast and carefully remove the gel into a tray, rinse with water, and add half a cap full of InstantBlue™. Stain for ~1 h with shaking at RT.
21. Discard the stain and wash twice with water, taking care not to tear the gel. Leave in water with shaking to destain for as long as required.
22. Confirm the sizing of your products against the protein ladder (see **Note 17** and Fig. 4).

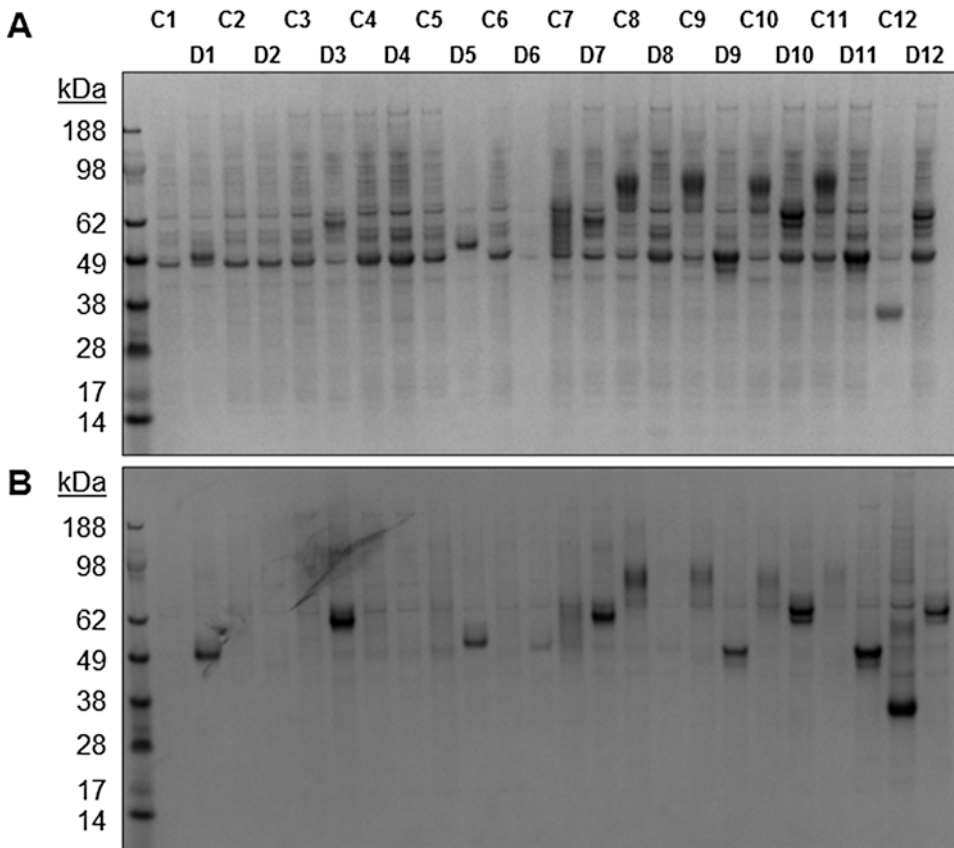


Fig. 4 Image showing the Coomassie SDS-PAGE result of a test purification from HEK293 cells. **(a)** Purification using a decahistidine tag and Talon resin. **(b)** Purification using a twin-strep tag and Strep-Tactin[®]XT resin. The gel shows a range of high, medium, and low expressions of various proteins of different molecular weights. Note that samples loaded using a multichannel pipette will be interleaved (e.g., C1, D1, C2, D2, etc.). When screening for purification of proteins from HEK293 cells, we discovered that a host protein contaminant, NONO (Non-POU Domain-Containing Octamer-Binding Protein) and its binding partner SFPQ (Splicing factor Proline and Glutamine Rich) were copurifying along with the target proteins during purification with Talon resin **(a)** but were not visibly detected during purification with Strep-Tactin[®]XT resin **(b)**. These contaminants were confirmed by tryptic digest tandem mass spectrometry analysis. Consequently, our preferred affinity purification tag from mammalian cells is the twin-strep tag

**3.10 Dionex
Fluorescence Size
Exclusion
Chromatography
(FSEC) Screening**

The following Thermo Scientific™ Chromeleon™ method is specific for the 96-well (Dionex) plate and the Dionex™ Ultimate 3000 UHPLC setup along with the SEC column listed in Subheading 2.4.

1. Place the elution plate from Subheading 3.9, **step 15** into the autosampler of the Dionex™ Ultimate 3000 UHPLC.
2. Attach the SEC buffer to an eluent line and purge the line at 5.0 mL/min for 90 s, to remove any air and to prime the system.
3. An injection sequence is written using the method listed in Table 2 for injecting samples from the elution plate onto the Dionex™ Ultimate 3000 UHPLC via the autosampler.
4. Export the entire run as a single PDF using Chromeleon Studio containing all the chromatograms for the injected samples, displaying Fluorescence Emission 1 and Fluorescence Emission 2. An example of an FSEC run is shown in Fig. 5.

4 Notes

1. The PEG 10,000 solution needs to be mixed during the cooling period to prevent a phase separation.
2. To dissolve the DDM, rotate the mixture at RT until completely dissolved and add the CHS to this solution, with rotation until fully dissolved. Alternatively, rotate the DDM overnight in the cold room, remove, then add the CHS to this solution and dissolve by rotation at RT, then filter.
3. The biotin will not dissolve in the Elution buffer until the pH is adjusted to approximately 7.5–8.0 by the addition of 0.5 M NaOH.
4. Expi293F™ cells should be maintained at $0.2\text{--}2 \times 10^6$ cells/mL to maintain doubling times; HEK293S GnTI– cells should be maintained at $0.3\text{--}2 \times 10^6$ cells/mL.
5. To keep it cost-effective and to express most of the difficult recombinant proteins, we have tested and compared a range of media and concluded to use FreeStyle™ 293 medium without any serum. Expi293™ medium which is recommended for use with Expi293F™ cells can also be used in this method and may give increased yields of protein.
6. Shaking speeds: Erlenmeyer flasks, baffled (for HEK293S GnTI–) or nonbaffled (for Expi293F™), filled to 15–30% with cells and shaking speeds of 90–105 rpm; Roller bottles, for the 2 L bottles filled with 500 mL to 1 L of cells, shaking speeds are 170 rpm on a shaker with throw radii at 25 mm.

Table 2
Injection sequence for injecting samples from the elution plate onto the Dionex™ Ultimate 3000 UHPLC via the autosampler

Time (min)	Instrument setup	
	FractionCollector.Valve	Drain
	UV.SlitWidth	Wide
	UV.ResponseTime	1.000 (s)
	UV.Data_Collection_Rate	5 (Hz)
	Sampler.InjectWash	AfterDraw
	Sampler.WashSpeed	20.000 (μL/s)
	Sampler.WashVolume	50.000 (μL)
	Sampler.SampleHeight	0.200 (mm)
	Sampler.SampleHeightOffset_96	3.000 (mm)
	Sampler.WasteSpeed	32.000 (μL/s)
	Sampler.DispenseDelay	0.000 (s)
	Sampler.DispSpeed	20.000 (μL/s)
	Sampler.DrawSpeed	0.5 (μL/s)
	Sampler.DrawDelay	3.000 (s)
	Sampler.InjectMode	Normal
	Sampler.PumpDevice	Pump
	Sampler.LoopWashFactor	2
	Sampler.TempCtrl	Off
	FLD.FLD_FlowCell.TempCtrl	On
	FLD.FLD_FlowCell. ReadyTempDelta	0.50 (°C)
	FLD.FLD_FlowCell.Temperature. Nominal	30.00 (°C)
	PumpModule.Pump.%A.Equate	%A
	PumpModule.Pump.%B.Equate	%B
	PumpModule.Pump.%C.Equate	%C
	PumpModule.Pump.%D.Equate	%D
	PumpModule.Pump.Pressure. LowerLimit	5 (bar)
	PumpModule.Pump.Pressure. UpperLimit	250 (bar)

(continued)

Table 2
(continued)

Time (min)	Instrument setup	
	PumpModule.Pump. MaximumFlowRampUp	0.500 (mL/min ²)
	PumpModule.Pump. MaximumFlowRampDown	1.000 (mL/min ²)
	UV.UV_VIS_1.Wavelength	280 (nm)
	UV.UV_VIS_1.Bandwidth	1 (nm)
	UV.UV_VIS_1.RefWavelength	Off
	UV.UV_VIS_1.RefBandwidth	1 (nm)
	UV.UV_VIS_2.Wavelength	340 (nm)
	UV.UV_VIS_2.Bandwidth	1 (nm)
	UV.UV_VIS_2.RefWavelength	Off
	UV.UV_VIS_2.RefBandwidth	1 (nm)
	UV.UV_VIS_3.Wavelength	420 (nm)
	UV.UV_VIS_3.Bandwidth	1 (nm)
	UV.UV_VIS_3.RefWavelength	Off
	UV.UV_VIS_3.RefBandwidth	1 (nm)
	UV.UV_VIS_4.Wavelength	490 (nm)
	UV.UV_VIS_4.Bandwidth	1 (nm)
	UV.UV_VIS_4.RefWavelength	Off
	UV.UV_VIS_4.RefBandwidth	1 (nm)
	FLD.Emission_1.ExWavelength	480.0 (nm)
	FLD.Emission_1.EmWavelength	510.0 (nm)
	FLD.Emission_1.Sensitivity	3
	FLD.Emission_1.FilterWheel	280nm
	FLD.Emission_2.ExWavelength	280.0 (nm)
	FLD.Emission_2.EmWavelength	310.0 (nm)
	FLD.Emission_2.Sensitivity	1
	FLD.Emission_2.FilterWheel	280nm
	FLD.Emission_3.ExWavelength	280.0 (nm)
	FLD.Emission_3.EmWavelength	350.0 (nm)
	FLD.Emission_3.Sensitivity	1

(continued)

Table 2
(continued)

Time (min)	Instrument setup
	FLD.Emission_3.FilterWheel 280nm
	FLD.Emission_4.ExWavelength 280.0 (nm)
	FLD.Emission_4.EmWavelength 350.0 (nm)
	FLD.Emission_4.Sensitivity 4
	FLD.Emission_4.FilterWheel 280nm
	FractionCollector. FractionCollection. CollectFractions No
	FLD.BaselineBehavior Append
	Column_Switch.Position 1
	Detector_Switch.Position 1
	Valve_A.Position System.Injection.CustomVariables.Eluent_line
0	Inject Preparation
	UV.Autozero
	FLD.AutoZero
	Wait UV.Ready And Sampler.Ready And FLD.Ready And PumpModule.Pump.Ready
0	<i>Inject</i>
	Sampler.Inject
0	<i>Start run</i>
	UV.UV_VIS_1.AcqOn
	UV.UV_VIS_2.AcqOn
	UV.UV_VIS_3.AcqOn
	UV.UV_VIS_4.AcqOn
	FLD.FLD_FlowCell.AcqOn
	PumpModule.Pump. Pump_Pressure.AcqOn
	FLD.Emission_1.AcqOn
	FLD.Emission_2.AcqOn
	FLD.Emission_3.AcqOn
	FLD.Emission_4.AcqOn

(continued)

Table 2
(continued)

Time (min)	Instrument setup
0	<i>Start flow</i>
	PumpModule.Pump.Flow.Nominal 0.6 (mL/min)
	PumpModule.Pump.%B.Value 0.0 (%)
	PumpModule.Pump.%C.Value 0.0 (%)
	PumpModule.Pump.%D.Value 0.0 (%)
	PumpModule.Pump.Curve 5
25	<i>Stop flow</i>
	PumpModule.Pump.Flow.Nominal 0.6 (mL/min)
	PumpModule.Pump.%B.Value 0.0 (%)
	PumpModule.Pump.%C.Value 0.0 (%)
	PumpModule.Pump.%D.Value 0.0 (%)
	PumpModule.Pump.Curve 5
25	<i>Stop run</i>
	UV.UV_VIS_1.AcqOff
	UV.UV_VIS_2.AcqOff
	UV.UV_VIS_3.AcqOff
	UV.UV_VIS_4.AcqOff
	FLD.FLD_FlowCell.AcqOff
	PumpModule.Pump. Pump_Pressure.AcqOff
	FLD.Emission_1.AcqOff
	FLD.Emission_2.AcqOff
	FLD.Emission_3.AcqOff
	FLD.Emission_4.AcqOff

- The %Cell viability is calculated by counting the number of viable cells and also the number of total cells on the hemocytometer grid. Viable cells do not take up Trypan Blue Stain however, nonviable cells take up the stain and appear blue under the microscope. To determine cell viability, mix 0.1 mL of Trypan Blue Stain with 1 mL of cell suspension and load a hemocytometer. Count the number of blue-stained cells and the total number of cells and then calculate the number of viable cells per mL and correct for the dilution factor. Cell

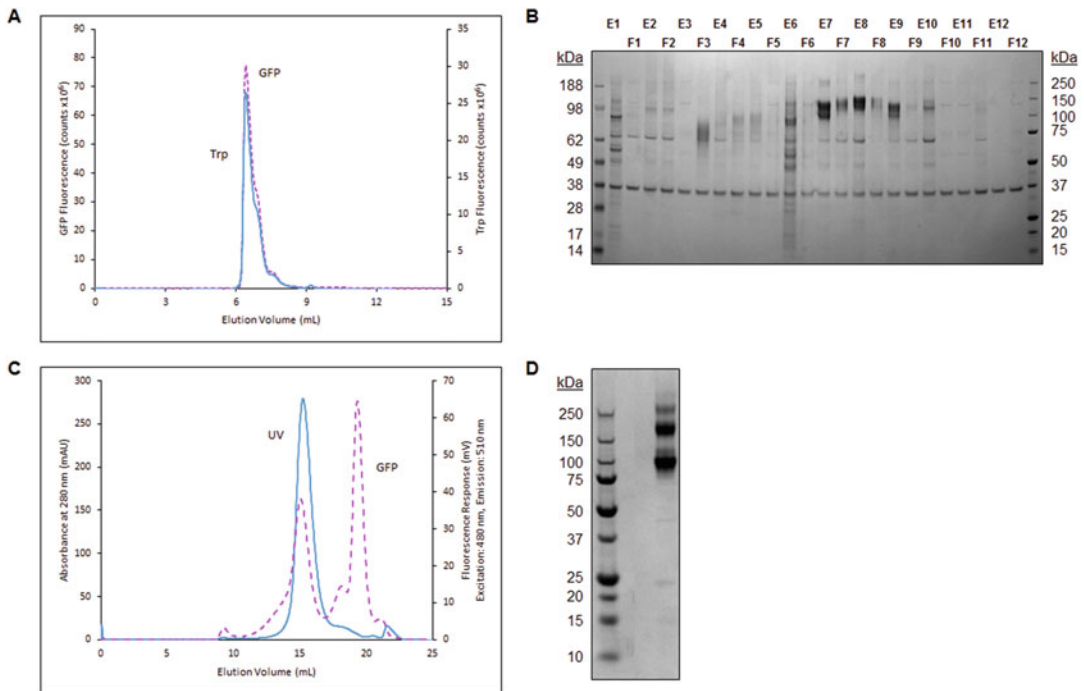


Fig. 5 FSEC analysis of a test purification from HEK293 cells. **(a)** Chromatogram showing fluorescence measurements for both tryptophan and GFP, corresponding to lane E8 of the test purification shown in **(b)**. **(b)** Coomassie SDS-PAGE result of a twin-strep test purification from mammalian cells. **(c)** FSEC chromatogram of a scale-up purification of sample lane E8. **(d)** Scale-up purification using a twin-strep tag and Strep-Tactin[®] XT resin followed by FSEC

viability should be at least 95% for a healthy log phase culture before it can be used for transfection, virus amplification or protein expression.

8. Depending on the protein of interest, it may be beneficial to use P2 virus instead of P1 for the transduction step to improve the expression levels and SDS-PAGE readout. If no expression is observed at all, determining viral titers using an endpoint dilution assay [36] and optimizing multiplicity of infection (MOI) could be helpful. Alternatively, a transient expression test with PEI (polyethyleneimine) is recommended to rule out a lack of expression due to low BacMam titers.
9. The expression conditions of 37 °C for 48 h can be varied depending on the proteins being tested. For proteins prone to degradation, toxicity or to aid protein folding, alternative conditions used are 30 °C for 72 h. Both conditions can be tested in parallel if best expression conditions are not known for a particular protein.
10. The block will not leak when placed on its side when the silicone matting seal is pressed down firmly and held in place with another deep well block.

11. Do not boil membrane protein samples as they may aggregate when heated up. They are loaded directly onto gels in the absence of boiling as this is likely to create a smear of proteins on the gel otherwise.
12. Pipette out the mixture gently and avoid touching the bottom of the plate so as to not disturb the cells.
13. The resin tends to clump and settles quickly. We recommend using 200 μ L tips with \sim 5 mm cut from the ends to prevent clogging the tips and ensure even loading. Also, continually mix the resin by pipetting up and down as well as shake the reservoir from side to side to prevent settling.
14. If you prefer to incubate the plate upright instead of by rotation, we would recommend keeping the samples in a 24-well format to provide optimal mix of the sample and resin.
15. Removing all trace of Wash buffer is essential to ensure that the subsequent elution step does not become diluted with Wash buffer.
16. When screening targets fused to GFP, we recommend running the gel at 100 V for 2 h in order to be able to scan it and detect the GFP signal prior to staining it with InstantBlue™.
17. Membrane proteins tend to run a bit smaller than expected due to their physical–chemical characteristics.

Acknowledgments

We would like to thank all the SGC scientists (past and present) who contributed toward the development of the method. The SGC is a registered charity (number 1097737) that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genome Canada, Innovative Medicines Initiative (EU/EFPIA), Janssen, Merck KGaA, MSD, Novartis Pharma AG, Ontario Ministry of Economic Development and Innovation, Pfizer, São Paulo Research Foundation-FAPESP, Takeda, and Wellcome. The BacMam vector backbone (pHTBV1.1) was kindly provided by Professor Frederick Boyce (Massachusetts General Hospital, Cambridge, MA).

References

1. Hofmann C, Sandig V, Jennings G et al (1995) Efficient gene transfer into human hepatocytes by baculovirus vectors. *Proc Natl Acad Sci* 92 (22):10099–10103
2. Boyce FM, Bucher NL (1996) Baculovirus-mediated gene transfer into mammalian cells. *Proc Natl Acad Sci U S A* 93(6):2348–2352
3. Shoji I, Aizaki H, Tani H et al (1997) Efficient gene transfer into various mammalian cells, including non-hepatic cells, by baculovirus vectors. *J Gen Virol* 78(Pt 10):2657–2664
4. Condeary JP, Witherspoon SM, Clay WC et al (1999) Transient and stable gene expression in mammalian cells transduced with a

- recombinant baculovirus vector. *Proc Natl Acad Sci U S A* 96(1):127–132
5. Ho YC, Chen HC, Wang KC et al (2004) Highly efficient baculovirus-mediated gene transfer into rat chondrocytes. *Biotechnol Bioeng* 88(5):643–651
 6. Liang CY, Wang HZ, Li TX et al (2004) High efficiency gene transfer into mammalian kidney cells using baculovirus vectors. *Arch Virol* 149(1):51–60
 7. Gao R, McCormick CJ, Arthur MJ et al (2002) High efficiency gene transfer into cultured primary rat and human hepatic stellate cells using baculovirus vectors. *Liver* 22(1):15–22
 8. Wang S, Balasundaram G (2010) Potential cancer gene therapy by baculoviral transduction. *Curr Gene Ther* 10(3):214–225
 9. Luo WY, Shih YS, Hung CL et al (2012) Development of the hybrid sleeping beauty: baculovirus vector for sustained gene expression and cancer therapy. *Gene Ther* 19(8):844–851
 10. Luo WY, Shih YS, Lo WH et al (2011) Baculovirus vectors for antiangiogenesis-based cancer gene therapy. *Cancer Gene Ther* 18(9):637–645
 11. Paul A, Binsalamah ZM, Khan AA et al (2011) A nanobiohybrid complex of recombinant baculovirus and Tat/DNA nanoparticles for delivery of Ang-1 transgene in myocardial infarction therapy. *Biomaterials* 32(32):8304–8318
 12. Chen HC, Sung LY, Lo WH et al (2008) Combination of baculovirus-expressed BMP-2 and rotating-shaft bioreactor culture synergistically enhances cartilage formation. *Gene Ther* 15(4):309–317
 13. Chen HC, Chang YH, Chuang CK et al (2009) The repair of osteochondral defects using baculovirus-mediated gene transfer with de-differentiated chondrocytes in bioreactor culture. *Biomaterials* 30(4):674–681
 14. Chuang CK, Lin KJ, Lin CY et al (2010) Xenotransplantation of human mesenchymal stem cells into immunocompetent rats for calvarial bone repair. *Tissue Eng A* 16(2):479–488
 15. Tani H, Abe T, Matsunaga TM et al (2008) Baculovirus vector for gene delivery and vaccine development. *Futur Virol* 3(1):35–43
 16. Bak XY, Lam DH, Yang J et al (2011) Human embryonic stem cell-derived mesenchymal stem cells as cellular delivery vehicles for pro-drug gene therapy of glioblastoma. *Hum Gene Ther* 22(11):1365–1377
 17. Zeng J, Du J, Zhao Y et al (2007) Baculoviral vector-mediated transient and stable transgene expression in human embryonic stem cells. *Stem cells (Dayton, Ohio)* 25(4):1055–1061
 18. Chen YH, Wu JC, Wang KC et al (2005) Baculovirus-mediated production of HDV-like particles in BHK cells using a novel oscillating bioreactor. *J Biotechnol* 118(2):135–147
 19. Grabherr R, Ernst W (2010) Baculovirus for eukaryotic protein display. *Curr Gene Ther* 10(3):195–200
 20. Ames RS, Lu Q (2009) Viral-mediated gene delivery for cell-based assays in drug discovery. *Expert Opin Drug Discovery* 4(3):243–256
 21. Lotze MT, Kost TA (2002) Viruses as gene delivery vectors: application to gene function, target validation, and assay development. *Cancer Gene Ther* 9(8):692–699
 22. Kost TA, Condreay JP, Ames RS (2010) Baculovirus gene delivery: a flexible assay development tool. *Curr Gene Ther* 10(3):168–173
 23. Almen MS, Nordstrom KJ, Fredriksson R et al (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
 24. Uhlen M, Fagerberg L, Hallstrom BM et al (2015) Proteomics. Tissue-based map of the human proteome. *Science* 347(6220):1260419
 25. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? *Nat Rev Drug Discov* 5(12):993–996
 26. Boesze-Battaglia K, Schimmel R (1997) Cell membrane lipid composition and distribution: implications for cell function and lessons learned from photoreceptors and platelets. *J Exp Biol* 200(Pt 23):2927–2936
 27. Muro E, Atilla-Gokcumen GE, Eggert US (2014) Lipids in cell biology: how can we understand them better? *Mol Biol Cell* 25(12):1819–1823
 28. van Meer G, de Kroon AIPM (2011) Lipid map of the mammalian cell. *J Cell Sci* 124(1):5
 29. Kost TA, Condreay JP (2002) Recombinant baculoviruses as mammalian cell gene-delivery vectors. *Trends Biotechnol* 20(4):173–180
 30. Andréll J, Tate CG (2013) Overexpression of membrane proteins in mammalian cells for structural studies. *Mol Membr Biol* 30(1):52–63
 31. Scott MJ, Modha SS, Rhodes AD et al (2007) Efficient expression of secreted proteases via recombinant BacMam virus. *Protein Expr Purif* 52(1):104–116
 32. Dukkipati A, Park HH, Waghay D et al (2008) BacMam system for high-level expression of

- recombinant soluble and membrane glycoproteins for structural studies. *Protein Expr Purif* 62(2):160–170
33. Yin Y, Wu M, Zubcevic L et al (2018) Structure of the cold- and menthol-sensing ion channel TRPM8. *Science* 359(6372):237–241
 34. Mansoor SE, Lu W, Oosterheert W et al (2016) X-ray structures define human P2X(3) receptor gating cycle and antagonist action. *Nature* 538(7623):66–71
 35. Liu F, Zhang Z, Csanady L et al (2017) Molecular structure of the human CFTR ion channel. *Cell* 169(1):85–95.e88
 36. Gochring A, Lee CH, Wang KH et al (2014) Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. *Nat Protoc* 9(11):2574–2585



High-Throughput Expression Screening in Mammalian Suspension Cells

Susan D. Chapple and Michael R. Dyson

Abstract

Proteins naturally expressed in eukaryotic organisms often require host chaperones, binding partners, and posttranslational modifications for correct folding. Ideally the heterologous expression system chosen should be as similar to the natural host as possible. For example, mammalian proteins should be expressed in mammalian expression systems. However, this does not guarantee a protein will be expressed in a sufficient high yield for structural or biochemical studies or antibody generation. Often a screening process is undertaken in which many parameters including truncations, point mutations, investigation of orthologs, fusion to peptide or protein tags at the N- or C-terminus, the coexpression of binding partners, and even culture conditions are varied to identify the optimal expression conditions. This requires multiparallel expression screening in mammalian cells similar to that already described for *E. coli* expression. Here we describe in detail a multiparallel method to express proteins in mammalian suspension cells by transient transfection in 24-well or 96-well blocks.

Key words Expression screening, HEK293 cells, CHO cells, Transient transfection, Mammalian cell culture, Interaction assays, Antibodies

1 Introduction

Expression of human and mammalian proteins in *E. coli* often results in a poor soluble expression yield [1]. Expression in eukaryotic systems such as yeast or insect cells can aid expression. However, the most authentic chaperones, binding partners and posttranslational modifications for mammalian proteins will be found in mammalian expression systems. There are several reasons why one may wish to perform a multiparallel expression experiment. Firstly it is common to express single or tandem domains of multidomain containing proteins to both improve expression and to study their function. Unfortunately domain boundaries are not accurately predicted within the current protein databases [2] and so often several truncations are performed at the DNA level either by rational or combinatorial [3] design followed by expression

screening. Secondly individual expression domains can be stabilized and their yield improved by fusion at the N- or C-terminus with peptide or protein tags [4]. Each protein target is different and so it is likely that several fusion partners would need to be investigated. Thirdly, it is well known that protein orthologs and mutations can display different solubility and crystallization properties and so one may wish to investigate a panel of point mutations and orthologs. Lastly some proteins are only stable in the presence of their natural binding partners and so one may wish to investigate coexpression with candidate binding partners [5]. The variables described here soon multiply and a thorough investigation requires the use of a plate based mammalian expression screen.

The optimisation of expression parameters is not the only reason an investigator may wish to perform a multiparallel expression experiment. They may also, for example, need to express a panel of receptor ectodomains for interaction studies [6] or functional screening [7]. Also panels of recombinant antibodies can be expressed for screening in proteomic applications [8, 9] or to aid therapeutic antibody lead isolation and optimisation projects [10, 11].

Screening expression in suspension adapted HEK293 or CHO cells allows the convenience of fast scale-up of any hits discovered in a small scale expression screen [12, 13]. Here we describe a method for transfection of HEK293F cells in 24-well blocks and a dot-blot screen to identify secreted expression screen hits. The dot blot screen could be replaced by a standard Western blot procedure or ELISA. The methods described here can also be adapted to 96-well transfections as described in Subheading 3.3. Expression system cell lines with improved productivity, such as Expi293 (Thermo Fisher) can also be adapted for 24- or 96-well using lipid-based transfection reagents [14].

2 Materials

All chemicals are from Sigma, unless stated otherwise.

2.1 HEK293F Cell Maintenance

1. For maintenance of cells in Erlenmeyer flasks a humidified CO₂ shake incubator is required with a 25 mm orbital throw such as the Infors Multitron.
2. Vented sterile Erlenmeyer flasks (Corning).
3. HEK293F cells and Freestyle media (Life Technologies).
4. A hemocytometer for cell counting.

2.2 HEK293F Cell 24-Well Block Transfection

1. For maintenance of cells in 24-well blocks a humidified CO₂ plate shake incubator is required with a 3 mm orbital throw such as the Infors Multitron plate shaker incubator.

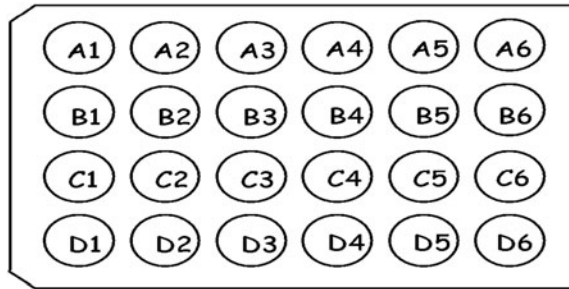


Fig. 1 24-well block for HEK293F cell culturing and transfection

2. Sterile 24-well blocks are from Qiagen (Fig. 1).
3. Linear 25 kDa polyethylenimine (PEI) is from Polysciences Inc. This is prepared at a concentration of 1 mg/ml in MilliQ water. Solubilization is achieved by first adding concentrated HCl to a stirred PEI solution until the pH is <2.0 and stirring continued for 2–3 h. The pH is then adjusted to 7.0 using concentrated NaOH. The PEI solution is finally filter-sterilized by passage through a 0.22 μm membrane and 1 ml aliquots are stored at $-20\text{ }^{\circ}\text{C}$.
4. Serum-free media (SFM).
5. Pluronic F68 reagent.

2.3 Expression Screening by Dot Blot

1. Dot blot apparatus from Schleicher & Schuell (Minifold I system dot blot apparatus).
2. Nitrocellulose from Schleicher & Schuell.
3. 8 M urea.

3 Methods

3.1 HEK293F Cell Maintenance

1. When the cell density reaches $1\text{--}4 \times 10^6$ [6] cells/ml passage the cells (*see Note 1*).
2. Centrifuge cells for 4 min at $150 \times g$ (Sorvall Legend centrifuge) at room temperature in a 50 ml Falcon tube.
3. Resuspend the cells in fresh medium (i.e., 1/4 the original culture volume) and pipet to break up any cell clumps.

4. Count viable cells by trypan blue exclusion using a 1:5 dilution (e.g., 200 μ l cells–100 μ l trypan blue (0.4%):700 μ l medium).
5. Seed the required culture volume with 2.5×10^5 [5] cells/ml using Freestyle medium (*see Note 2*).
6. Label flask with name, cell line name, passage number, date, and seeding density.
7. Incubate at 37 °C, 5% CO₂, 60% humidity, 125 rpm.
8. The cells will require splitting again 3–4 days later.

3.2 HEK293F Cell 24-Well Block Transfection

1. Split 200 ml of HEK293F cells at 2.5×10^5 cells/ml in a 1 l sterile vented Erlenmeyer flask for each 24-well block (i.e., for 96-well plate 4 \times 200 ml flasks are required), 48 h before the transfection.
2. On the day of transfection, add 400 μ l of serum-free media (SFM), warmed to room temperature (*see Note 3*) to the wells of the 24-well block followed by 4 μ g of plasmid DNA (*see Note 4*).
3. Add 4 μ l PEI to the walls of each well with a repeater pipettor or a multichannel pipette with a Varispan to allow pipetting into the 6-well row of the 24-well block. The PEI is placed approximately 0.5–1 cm from the meniscus of the SFM.
4. Vortex the 24-well block for 10 s on plate vortexer. Incubate for 10 min at room temp (*see Note 5*).
5. Add Pluronic F68 reagent into each 1 l vented Erlenmeyer flask, now containing 1×10^6 cells/ml (*see Note 6*) to a final concentration of 0.1% (*see Note 7*).
6. HEK293F cells are added (4 ml) to each well of the 24-well block containing the DNA – PEI complex. Cover with an air-pore plate sealer.
7. Incubate the 24-well block at 37 °C, 5% CO₂, 75% humidity, 400 rpm in a plate shake incubator with a 3 mm orbital throw. Check after 1 h that the cells are still in complete suspension.
8. Harvest after 5 days transfection (*see Note 8*), by centrifugation at $2500 \times g$ for 5 min, and analyze the supernatant (secreted proteins) or cell lysate (intracellular proteins) by Western blot or by dot blot.

3.3 HEK293F Cell 96-well Block Transfection

1. Split 55 ml of HEK293F cells at 2.5×10^5 cells/ml in a 500 ml sterile vented Erlenmeyer flask for each 96 well block, 48 h before the transfection.
2. On the day of transfection, add 50 μ l of serum-free media (SFM), warmed to room temperature (*see Note 3*) to the wells of the 96 square deep-well block followed by 0.5 μ g of plasmid DNA (*see Notes 4 and 9*).

3. Add 1 μl PEI to the walls of each well with a repeater pipettor or a multichannel pipette with a Varispan to allow pipetting into the 12-well row of the 96-well block. The PEI is placed approximately 0.5 cm from the meniscus of the SFM.
4. Vortex the 96-well block for 10 s on plate vortexer. Incubate for 10 min at room temp (*see Note 5*).
5. Add Pluronic F68 reagent into each 250 vented Erlenmeyer flask, now containing 1×10^6 cells/ml (*see Note 6*) to a final concentration of 0.1% (*see Note 7*).
6. HEK293F cells are added (0.5 ml) to each well of the 96-well block containing the DNA-PEI complex. Cover with an air-pore plate sealer.
7. Incubate the 96-well block at 37 °C, 5% CO₂, 75% humidity, 800 rpm in a plate shake incubator with a 3 mm orbital throw. Check after 1 h that the cells are still in complete suspension.
8. Harvest after 5 days transfection (*see Note 8*) by centrifugation at $2500 \times g$ for 5 min, and analyze the supernatant (secreted proteins) or cell lysate (intracellular proteins) by Western blot or by dot blot.

3.4 Expression Screening by Dot Blot

1. 8 M urea was added to cleared culture supernatants (or purified proteins) to give a final concentration of 5 M urea (i.e., 125 μl 8 M urea added to 75 μl culture supernatant) (*see Note 10*).
2. Incubate the culture supernatant-urea mix for 1 h at room temperature.
3. Set up dot blot apparatus during this time: Presoak Whatman 3MM filter paper (2–3 sheets) and nitrocellulose membrane in PBS.
4. Arrange Minifold I apparatus according to the Schleicher & Schuell protocol. In summary, place the middle unit (96 wells with small holes) on top of base collection unit according to the guide pins.
5. Place 2–3 sheets of PBS-soaked filter paper onto the unit, followed by the membrane.
6. Place the top unit (96-well plate with dispensing holes) in place over the filter paper and membrane using the line up pins.
7. Secure the whole dot blot apparatus in place using the four clips on the side (N.B.: make sure that they are fixed in place using clip 1 followed by clip 4 then clip 2 followed by clip 3 and NOT clips 1 + 2 followed by clips 3 + 4 as illustrated in Fig. 2).
8. When ready to load the samples: connect dot blot unit to vacuum source and turn on for a few seconds to clear the excess PBS from the wells.

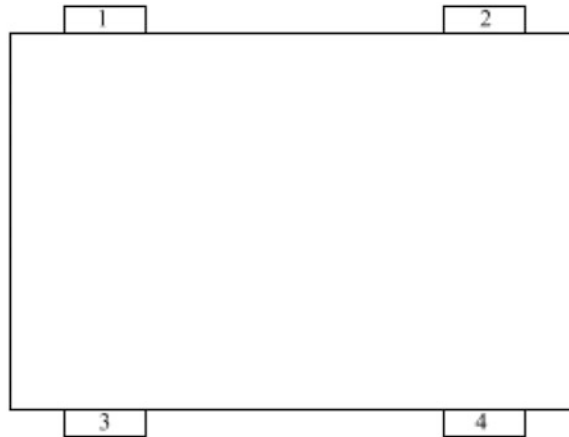


Fig. 2 Dot blot apparatus depicting clip numbering

9. Switch vacuum OFF, then load all samples to be analyzed (can use multichannel pipette).
10. Switch on vacuum and allow samples to move onto the membrane (this should take approx. 10–20 s). If there are small air bubbles trapping sample in a well, gently tap the apparatus on the bench to move the air bubble out the way and allow the sample to move onto the membrane.
11. Once finished, remove membrane and place in blocking solution (e.g., 3% milk/PBS/Tween).
12. Probe with antibody as detailed in standard Western blot protocols.
13. Finally: rinse dot blot apparatus in water to prevent anything clogging up the apparatus and allow to air-dry on bench.

4 Notes

1. Work at all times with good aseptic technique within a functioning tissue culture hood. Prewarm the culture media in hood for approx. 1 h. prior to use. Always clean (using ethanol spray) the inside of hood and any equipment to be used prior to use in the hood. Infection of mammalian cell cultures with bacteria or yeast results in poor expression yield and can be a major cause for delay.
2. HEK293F cells can be split as low as 1×10^5 cells/ml. It is important not to allow the cells to overgrow ($\geq 3 \times 10^6$ cells/ml) as dead cells can accumulate, resulting in a less healthy cell population. Maintaining cells in a good state is essential for high transfection efficiency and thus expression yield.

3. The serum-free media is the media the cells are normally propagated with, minus the addition of serum. For example for HEK293F cells, this would be Freestyle medium (Life Technologies).
4. The plasmid DNA to be used for transfection must be of sufficient purity to allow for an efficient transfection. The DNA should be prepared according to the NAPPA protocol [15], a standard midi- or maxi-prep method involving an isopropanol precipitation, or a commercially available transfection quality plasmid DNA kit from suppliers such as Qiagen or Macherey-Nagel. The OD260–OD280 ratio should be between 1.8 and 1.9. This ensure low protein and endotoxin contamination.
5. Ten minutes is the minimum time to allow for formation of the DNA–PEI complex. Up to 30 min still allows for efficient transfection, but from 30 min to 1 h transfection efficiency gradually decreases due to the formation of higher order DNA–PEI aggregates.
6. The cells should be as close to mid-logarithmic phase as possible (for HEK293 cells between 0.8×10^6 and 1.2×10^6 cells/ml) with a cell viability of $\geq 95\%$.
7. The antifoaming agent Pluronic is required to maintain the viability of the HEK293 suspension cells during growth in 24-well blocks.
8. The time required before harvesting depends on the protein being expressed. Intracellular and nuclear-localized proteins may require only 2–3 days for optimal expression, whereas secreted protein such as receptor ectodomains or antibodies typically require 4–5 days. The time required should be determined empirically for the target class of proteins being investigated.
9. For 96-well transfections it was found that both cell viability and cell productivity is affected both by the geometry of the 96-well plate and the volume of media in the wells. This is likely due to optimal gas exchange within the well. For example, a 1 ml volume within the well resulted in less cell productivity than a 0.5 ml or 0.25 ml volume (*see* Fig. 3).
10. It was found that the addition of urea enhanced the binding of glycoproteins to the nitrocellulose membrane [12].

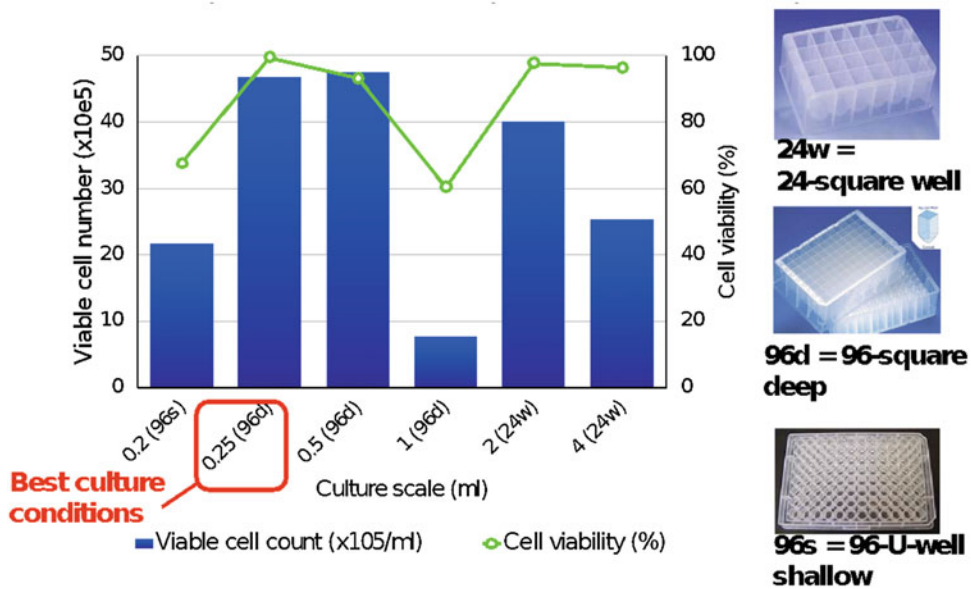


Fig. 3 Effect of plate geometry and culture volume on HEK293F cell growth and viability. HEK293F cells were seeded at 5×10^5 cells per ml. The 96-well plates were shaken in a CO₂ humidified plate shaker (Infors) for 3.5 days, total cells were counted and their viability determined by trypan blue staining. Cell viability (green circles) and cell count (blue blocks) are plotted for each plate type and culture scale on the x-axis. The plates employed include 96-well shallow plates with circular wells (96 s), 96-deep-well square block plates (96d), and 24-well block plates

References

- Dyson MR, Shadbolt SP, Vincent K, Perera R, McCafferty J (2004) Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression. *BMC Biotechnol* 4:32
- Dyson MR (2010) Selection of soluble protein expression constructs: the experimental determination of protein domain boundaries. *Biochem Soc Trans* 38:908–913
- Dyson MR et al (2008) Identification of soluble protein fragments by gene fragmentation and genetic selection. *Nucl Acids Res* 36:e51
- Brown MH, Barclay AN (1994) Expression of immunoglobulin and scavenger receptor superfamily domains as chimeric proteins with domains 3 and 4 of CD4 for ligand analysis. *Protein Eng* 7:515–521
- Trowitzsch S, Bieniossek C, Nie Y, Garzoni F, Berger I (2010) New baculovirus expression tools for recombinant protein complex production. *J Struct Biol* 172:45–54
- Bushell KM, Söllner C, Schuster-Boeckler B, Bateman A, Wright GJ (2008) Large-scale screening for novel low-affinity extracellular protein interactions. *Genome Res* 18:622–630
- Gonzalez R et al (2010) Screening the mammalian extracellular proteome for regulators of embryonic human stem cell pluripotency. *Proc Natl Acad Sci* 107:3552–3557
- Colwill K, Graslund S (2011) A roadmap to generate renewable protein binders to the human proteome. *Nat Methods* 8:551–558
- Dyson MR et al (2011) Mapping protein interactions by combining antibody affinity maturation and mass spectrometry. *Anal Biochem* 417:25–35
- Bradbury ARM, Sidhu S, Dubel S, McCafferty J (2011) Beyond natural antibodies: the power of in vitro display technologies. *Nat Biotech* 29:245–254
- Parthiban K et al (2019) A comprehensive search of functional sequence space using large mammalian display libraries created by gene editing. *mAbs* 11:884–898
- Chapple S, Crofts A, Shadbolt SP, McCafferty J, Dyson MR (2006) Multiplexed

- expression and screening for recombinant protein production in mammalian cells. *BMC Biotechnol* 6:49
13. Tom R, Bisson L, Durocher Y (2008) Transfection of HEK293-EBNA1 cells in suspension with linear PEI for production of recombinant proteins. *Cold Spring Harbor Protocols* 2008: pdb.prot4977
 14. Vazquez-Lombardi R et al (2018) Transient expression of human antibodies in mammalian cells. *Nat Protoc* 13:99–117
 15. Link AJ, LaBaer J (2008) Construction of nucleic acid programmable protein arrays (NAPPA) 3: isolating DNA plasmids in a 96-well plate format. *Cold Spring Harbor Protocols* 2008: pdb.prot5058



In Vitro Production of Perdeuterated Proteins in H₂O for Biomolecular NMR Studies

Lionel Imbert, Rachel Lenoir-Capello, Elodie Crublet, Alicia Vallet, Rida Awad, Isabel Ayala, Celine Juillan-Binard, Hubert Mayerhofer, Rime Kerfah, Pierre Gans, Emeric Miclet, and Jerome Boisbouvier

Abstract

The cell-free synthesis is an efficient strategy to produce in large scale protein samples for structural investigations. In vitro synthesis allows for significant reduction of production time, simplification of purification steps and enables production of both soluble and membrane proteins. The cell-free reaction is an open system and can be performed in presence of many additives such as cofactors, inhibitors, redox systems, chaperones, detergents, lipids, nanodisks, and surfactants to allow for the expression of toxic membrane proteins or intrinsically disordered proteins. In this chapter we present protocols to prepare *E. coli* S30 cellular extracts, T7 RNA polymerase, and their use for in vitro protein expression. Optimizations of the protocol are presented for preparation of protein samples enriched in deuterium, a prerequisite for the study of high-molecular-weight proteins by NMR spectroscopy. An efficient production of perdeuterated proteins is achieved together with a full protonation of all the amide NMR probes, without suffering from residual protonation on aliphatic carbons. Application to the production of the 468 kDa TET2 protein assembly for NMR investigations is presented.

Key words Cell-free, In vitro protein synthesis, Structural biology, Isotopic labeling, NMR, Perdeuteration

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is an established method to study interactions, structure, and dynamics of biomolecules at atomic resolution. This approach relies on the detection of NMR signals of natural hydrogen ¹H isotope, the most abundant nucleus in biomacromolecules, characterized by favorable NMR properties such as a high gyromagnetic ratio and a spin 1/2. Multidimensional ¹H-homonuclear NMR studies of proteins [1, 2] were

Lionel Imbert and Rachel Lenoir-Capello contributed equally to this work.

limited to systems with a molecular weight up to ca. 10 kDa, due to the complexity of NMR spectra. During the last three decades, introduction of isotopic labeling techniques considerably increased the maximum size of biochemical systems that can be addressed by NMR spectroscopy. The use of robust protocols to introduce stable ^{15}N and ^{13}C spin $\frac{1}{2}$ nuclei in recombinant proteins together with the development of triple resonance experiments have allowed spectroscopists to simplify and extend application of NMR to proteins with a molecular weight of ca. 25 kDa [3–6]. Solution NMR studies of larger biomolecules are challenging due to the inherent spectral overlap between all the ^1H , ^{13}C , and ^{15}N signals. Furthermore, rapid transverse relaxation (R_2) induces broadening and decreases intensity of NMR signals in high molecular weight proteins. This fast transverse relaxation is mainly due to the large number of intense dipolar interactions involving the abundant ^1H nuclei. As the magnitude of these dipolar interactions increases with the hydrodynamic radius of studied biomolecules, it is more complex to study larger proteins characterized by slow overall tumbling.

Perdeuteration of proteins [7–10] was shown to improve relaxation properties in order to study larger targets. Deuterium isotope (^2H) has indeed a low gyromagnetic ratio, the dipolar interactions involving hydrogen nuclei are decreased by a factor of 43 when deuteron (^2H) substitutes the proton (^1H), leading to slower transverse relaxation and concomitant increase of sensitivity and resolution. In combination with optimized NMR experiments [11, 12] using spectrometers operating at high magnetic field, perdeuteration allows for the study of monomeric proteins as large as 82 kDa [13] and, in favorable cases, complexes above 100 kDa [14, 15]. Common methods to produce perdeuterated proteins for NMR studies usually rely on the overexpression of the target protein in *E. coli* grown in minimal media containing 100% $^2\text{H}_2\text{O}$ as solvent and a deuterated carbon source [10] (*see* Fig. 1a). Such protocols enable protein perdeuteration up to 98% [16]. In order to observe backbone amide protons ($^1\text{H}_\text{N}$), the overexpressed protein is purified and finally dialyzed against $^1\text{H}_2\text{O}$ to allow for back-exchange of protein $^2\text{H}_\text{N}$ with ^1H nuclei from the solvent. While this simple approach is suitable for fast exchanging protons, amide protons located in the core of large proteins exchange too slowly with the solvent to allow for efficient back protonation. Generally, to reintroduce ^1H probes in these protected parts of the protein, the strategy consists in destabilizing the protein with chaotropic agents in $^1\text{H}_2\text{O}$ to speed up H_N exchange, before refolding the protein (*see* Fig. 1a). The drawback of such a strategy is that the refolding of large proteins or membrane proteins is particularly challenging. At best, the target proteins will be refolded with poor yields and loss of precious labeled materials, but a lot of proteins of biologic interest cannot be refolded *in vitro* in their native conformation in absence of cellular

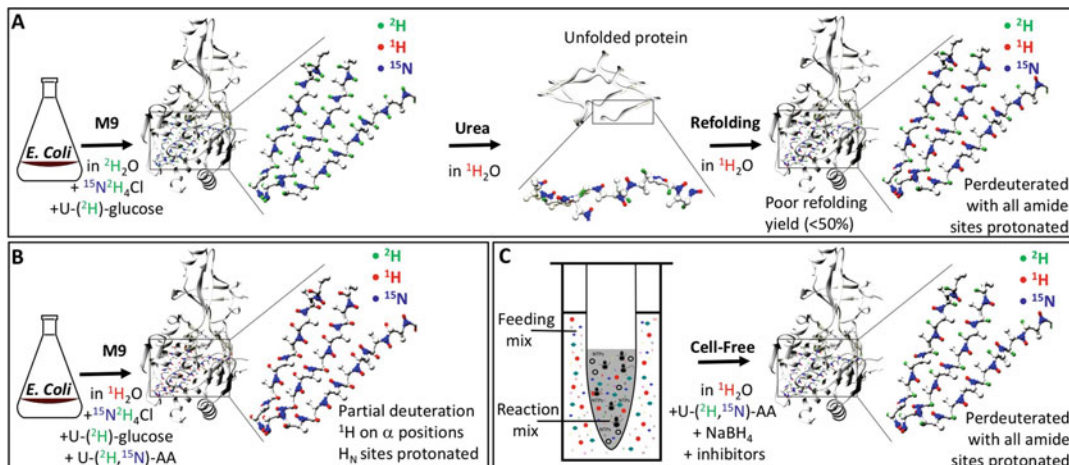


Fig. 1 Schematic illustration of strategies used to produce perdeuterated proteins in vivo or in vitro. **(a)** In vivo production of perdeuterated proteins in M9/ $^2\text{H}_2\text{O}$ media. Purified proteins need to be destabilized in $^1\text{H}_2\text{O}$ to allow back-protonation of amide sites before refolding. **(b)** In vivo production of perdeuterated proteins in M9/ $^1\text{H}_2\text{O}$ media supplemented with U- $^2\text{H}, ^{15}\text{N}$] or U- $^2\text{H}, ^{13}\text{C}, ^{15}\text{N}$] amino acids to ensure complete protonation of amide sites. α -sites are partially protonated because of the residual transaminase activities. **(c)** Cell-free production of perdeuterated proteins in M9/ $^1\text{H}_2\text{O}$ media supplemented with perdeuterated amino acids using NaBH_4 treated S30 cell extracts in the presence of transaminase inhibitors (DM, AOA) to suppress residual protonation on α -positions and ensure full protonation of amide sites

cofactors or chaperones, precluding production of a sample for NMR investigations.

The use of minimal medium prepared in 100% $^1\text{H}_2\text{O}$ buffer and supplemented with an excess of perdeuterated amino acids have been proposed to produce proteins fully labeled with ^1H nuclei on exchangeable sites of the overexpressed protein (ca. 20% of hydrogen in proteins), while nonexchangeable hydrogen covalently bound to aliphatic or aromatic carbons are deuterated at a level of ca. 85% [17] (see Fig. 1b). This level of deuteration allows for acquisition of high-quality 2D- $(^{15}\text{N}, ^2\text{H})$ spectra for large proteins [17, 18]. The residual protonation is, however, not homogeneously distributed on all the H_α , and a higher level of ^1H spin is observed on α sites (30–80% of ^1H [10]) because of abundant transaminases in living cells that are able to catalyze the exchange of $^2\text{H}_\alpha$ with protons from the solvent [10, 19] (see Fig. 1b). In small and medium size proteins, this heterogeneity results in the presence of extra $^{13}\text{C}_\alpha$ signals increasing complexity of NMR spectra, and in larger proteins the signals corresponding to $^{13}\text{C}_\alpha-^1\text{H}$ isotopomers disappear due to the unfavorable transverse relaxation. Such artifacts introduced by the residual protonation deteriorate the quality of 3D NMR spectra used to sequentially assign backbone NMR frequencies, a prerequisite step for the investigation of proteins by NMR spectroscopy.

In vitro protein synthesis has been shown to be an attractive alternative method to produce perdeuterated proteins in $^1\text{H}_2\text{O}$ buffer (*see* Fig. 1c). Cell-free, or in vitro expression of proteins, exploits transcription and translation machineries extracted from prokaryotic [20, 21] or eukaryotic [22] cells. The cell extract, containing the required protein synthesis machineries, is recovered after removal of the DNA/RNA of the original organism. These machineries are supplemented with T7 RNA polymerase, amino acids, and energy sources, in order to express the target protein from provided DNA template. Due to the absence of a biological membrane, the cell-free environment is an open system, offering the possibility of adding at any time compounds such as cofactors, ligands and stabilizers, in order to improve the synthesis of the target protein or to label the protein with isotopes without suffering from metabolic scrambling as observed in vivo. In this chapter, we describe protocols extensively applied by authors to produce *E. coli* S.30 extracts and T7 RNA polymerase, required for large scale in vitro synthesis of milligrams of protein samples for structural biology investigations. An optimized version of the protocol to produce perdeuterated proteins in $^1\text{H}_2\text{O}$ buffer for biomolecular NMR studies is also presented. Particular attention has been given to the quantification of residual protonation level on aliphatic carbons with and without treatment of S30 cell-free extract to inhibit transaminases activities. These protocols are illustrated by 2D- $(^1\text{H}, ^{15}\text{N})$ solution and solid-state NMR spectra of peptidase TET2, a homododecameric protein assembly of 468 kDa. We show that in vitro expression of such large proteins in $^1\text{H}_2\text{O}$ supplemented with deuterated amino acids allows for recovery of a substantial number of signals for important $^1\text{H}_\text{N}$ probes undetectable when proteins are produced using standard in vivo perdeuteration protocols.

2 Materials

The cell-free protein synthesis is a coupled reaction of transcription and translation starting from a DNA template. This step of transcription needs to be performed in RNase-free conditions. Careful consideration should therefore be given to the quality of plastic consumables and to the cleaning of glassware and devices used along the process. All surfaces, pipettes, and glassware should be cleaned with RNase remover (RNase-Off, Shimitek), and washed with RNase-free water before being dried with ethanol. Users have to wear clean gloves all the time and work on ice as much as possible. All buffers should be prepared with RNase-free water, sterilized using a 0.22 μm filter.

**2.1 *E. coli* Extracts
Preparation (S30
Extract)**

1. Fermenter, protocol is described here for Techfors-S 42 L (INFORS HT).
2. RNase-free water (such as Direct Q3 with Biopak as final filter, Millipore).
3. Centrifuge J26 (Beckman) with JLA 9.1000, bottles (catalog # 366751, Beckman).
4. Ultracentrifuge (Beckman) with 45Ti rotor and six 70 mL bottles (catalog # 355655 Beckman).
5. Benchtop centrifuge for 1.5, 15, and 50 mL tubes.
6. French press.
7. OD_{600nm} spectrophotometer.
8. Hybridization oven at 42 °C or thermostatic orbital shaker.
9. 4X Z-media: 165 mM KH₂PO₄, 664 mM K₂HPO₄, and 40 g/L yeast extract
10. Glucose–thiamine solution: 240 g of glucose and 12 mg of thiamine per liter.
11. *E. coli* strain BL21 DE3 or other.
12. 1 M dithiothreitol (DTT)
13. 10× S30 buffer: 100 mM HEPES–KOH, pH 7.5, 600 mM C₂H₃O₂K (potassium acetate); 140 mM C₄H₆MgO₄ (magnesium acetate), add fresh DTT at 10 mM
14. Luria–Bertani (LB) medium for overnight culture.
15. Antifoam 289 (Merck).
16. 5 M NaCl solution
17. Spectra/Por 4 dialysis tube (12–14 MWCO) and magnetic clamp.
18. pIVEX GFP plasmid as model protein (RTS 100 *E. coli*, catalog # BR1400106, biotechrabbit).

**2.2 T7 RNA
Polymerase
Expression
and Purification**

1. 10 L of LB ampicillin (100 mg/L final) for agar plates and culture medium.
2. 1 M isopropyl-beta-D-thiogalactopyranoside (IPTG) solution (catalog # EU0008-C, Euromedex)
3. *E. coli* strain BL21 DE3 (Invitrogen).
4. Vector containing the gene of the target T7RNAPol (pAR1219).
5. Centrifuge J26 (Beckman) with 6 × 1 L rotor, JA 25.5 and JA 14 and dedicated bottles (catalog# 363678, 357002 and 355673, Beckman).
6. Ultracentrifuge Beckman with 45Ti rotor and six 70 mL bottles (catalog # 355655, Beckman).

7. Ultrasonic liquid processors (Vibra-Cellell VC505, Sonics).
8. FPLC device at 4 °C (DuoFlow 10, Bio-Rad) with SP Sepharose High Performance packed column (catalog # 17108701, GE).
9. SDS-PAGE gels (Miniprotean gradient 4–20%, stainfree, Bio-Rad).
10. OD_{600nm} spectrophotometer.
11. Lysozyme (catalog # 5934C, Euromedex).
12. 0.8% sodium deoxycholate in water
13. 0.5 M NaOH and 2 M NaCl solution
14. 1 M DTT, 10 mg/mL benzamidine, 0.1 M phenylmethane-sulfonyl fluoride (PMSF) in EtOH 100%
15. Buffer W: 20 mM Tris-HCl, pH 8.1, 20 mM NaCl, 2 mM EDTA, 1 mM DTT, 50 μM PMSF, and 10 μg/mL benzamidine.
16. Spectra/Por 4 dialysis tube (12–14 MWCO) and magnetic-clamp.
17. Buffer R: 50 mM Tris-HCl, pH 8.1, 20 mM NaCl, 2 mM EDTA, 1 mM DTT, 50 μM PMSF, and 10 μg/mL benzamidine.
18. 2 M ammonium sulfate solution
19. 50% poly(ethyleneimine) (PEI) solution in H₂O (catalog # 03880, Merck)
20. 10% PEI solution in buffer A
21. 4.1 M saturated ammonium sulfate solution, pH 7.0
22. 99% glycerol solution (catalog # EU3550, Euromedex)
23. Buffer A: 20 mM NaH₂PO₄, pH 7.7, 50 mM NaCl, 1 mM EDTA, and 1 mM DTT.
24. Buffer B: 20 mM NaH₂PO₄, pH 7.7, 300 mM NaCl, 1 mM EDTA, and 1 mM DTT.
25. Buffer C: 20 mM NaH₂PO₄, pH 7.7, 100 mM NaCl, 1 mM EDTA, and 1 mM DTT.

2.3 Cell-Free Expression of Proteins

1. 20 amino acids (Merck)
2. 1 M HCl (Merck)
3. 1 M KOH (Merck)
4. 100 mM solution of ribonucleotides CTP, GTP, UTP, and ATP, pH 7.0 (catalog # NU-1014, Jena Bioscience)
5. 2 M HEPES-KOH (Merck), pH 7.5
6. 10 mM folinic acid (Merck)
7. 100 mM 3',5'-cyclic AMP (Merck)

8. 1 M DTT (Merck)
9. 9.2 M ammonium acetate (Merck)
10. 1 M spermidine (Merck)
11. 1 M creatine phosphate (Merck)
12. 4 M potassium glutamate (Merck)
13. 1.07 M magnesium acetate
14. 17.5 mg/mL MRE600 tRNA (Roche)
15. 10 mg/mL creatine kinase (Roche)
16. Home-made T7RNAPol (*see* Subheading 3.2).
17. Home-made S30 extract (*see* Subheading 3.3).
18. 1 µg/µL in water of target protein DNA cloned in pIVEX 2.3d or 2.4d vectors (RTS pIVEX *E. coli* vector set distributed by biotechrabbit, catalog # BR1400701)
19. GeBAflex dialysis device (Euromedex).

2.4 In Vitro Production of Perdeuterated protein in H₂O

1. NaBH₄ (Merck).
2. Dimethylformamide (DMF).
3. 100 mM aminooxyacetate (AOA), pH 7.5
4. 500 mM D-malate (DM), pH 7.5
5. 10× S30 buffer: 100 mM HEPES-KOH, pH 7.5, 600 mM C₂H₃O₂K (potassium acetate), 140 mM C₄H₆MgO₄ (magnesium acetate), and fresh 10 mM DTT.
6. Algal Amino Acids Mix (AAAM) as Celtone[®] (Cambridge Isotope Labeling) or Isogro[®] (Merck).
7. Protease from *Streptomyces griseus* (catalog # P5147, Merck).
8. 20 mM Tris-HCl, pH 7.5
9. Water bath at 95 °C.
10. Benchtop centrifuge for 1.5, 15, and 50 mL tubes (Eppendorf).

3 Methods

3.1 E. coli Extracts Preparation (S30 Extract)

S30 extracts can be prepared in shake-flasks or a fermenter (*see Note 1*). The protocol detailed herein corresponds to a 12-L-fermenter culture of *E. coli* BL21 DE3 strain, enabling the preparation of 200 mL of S30 extract (*see Note 2*). RNase contamination can occur anytime during the protocol, so strict cleaning procedures with RNase remover should be applied for glassware and RNase-free plastic consumables should be used. Moreover, as ribosomes are temperature sensitive, S30 extract preparation must be performed on ice with chilled glassware/plastics.

1. Fermenter preparation: The day before the culture, prepare all the solutions listed in the dedicated Materials section, autoclave the glucose–thiamine solution and a funnel. Fill the fermenter tank with 9 L of H₂O and 3 L of 4× Z-media and autoclave the 12 L in situ (*see Note 3*). Inoculate 400 mL of LB with 50 μL of BL21 DE3 cells (glycerol stock) at 37 °C overnight.
2. *E. coli* culture: In the morning, warm up the medium at 37 °C, set up the stir at 550 rpm and the airflow at maximum level (*see Note 4*). Add the 1 L glucose–thiamine solution with a sterile funnel and inoculate the fermenter with the 400 mL overnight culture at an initial OD₆₀₀ of 0.1. Follow the pH, pO₂, and OD₆₀₀ every 30 min. A decrease of pO₂ is expected together with an increase of the optical density. Culture should be stopped when the value of pO₂ is close to zero or the OD₆₀₀ equal to 3, usually occurring 3–4 h after the beginning of the culture (*see Note 5*). The temperature must be quickly reduced from 37 to 16 °C, and the OD₆₀₀ should not exceed 3.2 (*see Note 6*). Harvest the cells by centrifugation for 15 min at 5000 × *g* and 4 °C.

3. Preparation of S30 extracts.

Wash: Perform 3 washing steps of the cell pellets with cooled S30 buffer, the first one with 2 liters, the second one with 1 L and finally with 0.5 Liter. At each step, pellet the cells at 4 °C for 15 min at 5000 × *g* (*see Note 7*). Weigh the wet cells and store the pellet overnight on ice in a cold room (*see Note 8*).

Lysis: The next day, cells are resuspended in cold S30 buffer (1.27 mL of buffer per gram of wet cells) and then disrupted using French press (only one pressure cycle). The supernatant is clarified by two centrifugation steps at 4 °C in six 70 mL ultracentrifuge tubes, using a 45 Ti rotor (30,000 × *g*, 30 min). After each centrifugation step, recover only the supernatant (corresponding to 80% of the initial volume).

Maturation: Process the endogenous mRNA by incubating the S30 extract supernatant at 42 °C for 45 min after the addition of a 5 M NaCl solution to reach a final concentration of 400 mM [23].

Dialysis: Glassware (a 2 L cylinder as dialysis tank, a 500 mL one for measuring the S30 volume), S30 buffer (12 L prepared without DTT) and a 12 kDa cutoff dialysis membrane should be stored in advance at 4 °C for dialysis (*see Note 9*). S30 extract is dialyzed in six steps each time against 2 L of S30 buffer at 4 °C. Perform two successive 1 hour-dialysis steps, one overnight dialysis bath, and three extra 1 hour-dialysis steps the next morning. The DTT

(1 mM) should be added at the beginning of each dialysis step. Centrifuge the S30 extract in 50 ml tubes at $5000 \times g$ and 4°C . Prepare 2 mL safe-lock tubes to aliquot the supernatant in 1 mL fractions and freeze the tubes immediately in liquid nitrogen. These S30 extract aliquots can be stored for several years at -80°C .

Quality control: Expression yields are very dependent on Mg^{2+} concentration and the precise Mg^{2+} concentration already present in the S30 extract can vary from batch to batch. For each S30 extract batch, the Mg^{2+} concentration has to be optimized by performing expression tests, in duplicate, using a model protein. We usually perform six in vitro expression tests in small volumes, with additional magnesium concentration in the reaction and feeding mixes ranging from 5 to 15 mM. This optimization is routinely performed using GFP as a model protein, and we determine the optimal concentration of Mg^{2+} from highest fluorescence intensity.

3.2 T7 RNA polymerase Expression and Purification

T7 RNA polymerase (T7RNAPol) needs to be purified in RNase-free conditions at 4°C . It is an essential component of the Cell-free reaction and high concentrations are needed for the reaction. It can be purchased from different providers or produced in-house by overexpression in *E. coli* and a one-step ion exchange purification (*see Note 10*) [24, 25].

1. Culture: Spread BL21DE3 cells transformed with the plasmid pAR1219 coding for T7RNAPol on an LB Ampicillin (LBA) agar plate and incubate overnight at 37°C . Inoculate a 30 mL LBA culture with a single colony and let grow for 8 h at 37°C with vigorous shaking. In the evening, inoculate an overnight culture of 200 mL at OD_{600} : 0.1. The next day, warm up 10 L at 37°C of culture medium for overexpression of T7RNAPol (ten 3 L flasks filled with 1 L of LBA media). Inoculate with the overnight culture at OD_{600} : 0.1, let grow and induce with IPTG at 0.3 mM when OD_{600} reaches 0.8. After 2 h at 37°C , collect the cells by centrifugation ($5000 \times g$, at 4°C for 15 min). In order to centrifuge only 5 L of culture simultaneously, the inoculation for 5 out of the 10 flasks can be shifted from 30 min. Wash the cells with buffer W and pellet the cells by centrifugation. Weigh the wet cells, and store pellets on ice overnight in a cold room. Around 40 g of wet cells are expected from a 10 L culture using this protocol.
2. Lysis: resuspend the pellet with 2.88 mL of buffer R per gram of wet cells, for instance 115.2 mL for a 40 g cell pellet. The lysis is initiated by adding 1.5 mg of lysozyme per gram of wet cells, followed by an incubation of 20 min on ice with

occasional shaking. The detergent sodium deoxycholate (0.8% stock solution in water) is subsequently added at 0.3 mL/g of wet cells (i.e., 12 mL for a 40 g pellet) and incubated for 20 min on ice with occasional shaking. The viscosity is reduced by sonication in a beaker on ice. The supernatant is clarified by two centrifugation steps at 4 °C. The first step is performed at $10,000 \times g$ for 15 min, and the supernatant is then ultracentrifuged for 3 h at $140,000 \times g$.

3. Ammonium sulfate precipitation: Adjust the supernatant volume to 5.76 mL per gram of wet cells with buffer A in a chilled beaker. Addition of a 2 M ammonium sulfate solution is performed slowly, drop by drop, on ice under stirring up to a final concentration of 0.2 M. To initiate precipitation of high molecular weight nucleic acids, add slowly, with stirring, on ice, polyethyleneimine at a final concentration of 0.5%. Incubate for 20 min on ice in a rocker-shaker. Centrifuge for 10 min at $30,000 \times g$, at 4 °C, collect the supernatant, and measure its volume (noted V) in a chilled cylinder (*see Note 11*). The T7RNAPol is precipitated using a drop-by-drop addition of an ammonium sulfate saturated solution (total volume added $0.82 \times V$). The addition is performed in a centrifuge bottle under stirring on ice, and followed by extra stirring for 15 min. The sample is centrifuged for 20 min at $15,000 \times g$ and 4 °C, and the pellet is solubilized with 4 mL of buffer C per gram of wet cells. In order to eliminate ammonium sulfate, the solution is dialyzed once against 5 L of chilled buffer A, using a 12 kDa cutoff membrane.
4. Ion exchange purification (*see Note 12*): The supernatant is clarified by centrifugation for 10 min at $30,000 \times g$ and loaded at a flow rate of 2–3 mL/min on the SP-Sepharose column, previously equilibrated with buffer A. The column is washed with 1.5 column volume (CV) of buffer A and the T7RNAPol is eluted by a gradient from 0 to 100% of buffer B in 20 CV. The enzyme is recovered around 100 mM NaCl and collected by 5 mL fractions. The fractions are analyzed on SDS-PAGE gel and the ones containing T7 RNAPol at highest concentration (center of the chromatography peak) are pooled, and dialyzed against 1 L of cold buffer C containing 50% glycerol. After an overnight dialysis, the volume naturally decreases by 2 (because of the glycerol), to reach a concentration of ca. 5 mg/mL. The T7RNAPol can be stored for at least 1 year at -20 °C (*see Note 13*).

3.3 Cell-Free Expression of Proteins

We present here the Continuous Exchange Cell-free (CECF) reaction using a dialysis membrane system [26, 27]. This protocol is optimized for the large-scale expression of unlabeled proteins typically used in structural biology projects. This section describes the preparation of the reaction mix (1 mL) and the feeding mix

(10 mL) required to synthesize in vitro the target protein. Particular attention should be paid to the design and purity of the DNA vectors coding for the target protein. pIVEX vectors [28] have been developed by Roche and are optimized for in vitro expression using T7RNAPol and *E. coli* S30 extracts. They are designed with a T7 promoter, Ribosome Binding Site, T7-terminator sequences to allow efficient protein synthesis, His-tag sequence at N or C-terminus to facilitate detection or purification of expressed protein, and Multi Cloning Sites for insertion of target protein DNA sequence. The DNA sequence for N-terminal His-tag is optimized to ensure efficient initiation of transcription and translation. Alternative optimized expression tags [29] can be incorporated to increase production yield of the target protein. The DNA vector should be prepared in large scale (ca. 500 µg) using a commercial plasmid preparation kit and eluted in RNase-free water at a concentration of 1 mg/mL. In the following protocol, vortex all the mixtures, except at point 6.

1. Solubilize amino acid mixtures in three 50 mL tubes, the concentration of each amino acid being 50 mM. Alanine, arginine, glycine, histidine, lysine, proline, serine, threonine, and valine are resuspended together in water. Acidic soluble amino acids (asparagine, aspartate, cysteine, glutamine, glutamate, leucine, methionine, tryptophan, and tyrosine) are resuspended in HCl 1 M. Isoleucine and phenylalanine are resuspended in KOH 1 M.
2. Wash extensively the dialysis Gebaflex tube with water and keep it in water during the CECF reaction preparation (*see Note 14*).
3. Prepare the amino acid solution (AA mix) by adding 247.5 µL of each amino acid mix (water-soluble, acid-soluble and base-soluble amino acids), complement with 82.5 µL RNase free water.
4. Prepare the 10X reaction mix containing HEPES–KOH (55 mM, pH 7.5); DTT (3.4 mM); ATP (1.2 mM); 3', 5'-cyclic AMP (0.64 mM); 0.8 mM of each CTP, GTP, and UTP ribonucleotides; folinic acid (68 µM); ammonium acetate (27.5 mM); and spermidine (2 mM).
5. Prepare the feeding mix with 1.05 mL of the 10× reaction mix, creatine phosphate at final concentration of 80 mM, potassium glutamate (208 mM), AA mix (1 mM for each amino acid), and magnesium acetate (14.4 mM).
6. Prepare the reaction mix as described for the feeding mix, with further addition of 250 mg/mL of creatine kinase, 175 µg/mL of total tRNA *E. coli* MRE600, 50 µg/mL of T7 RNAPol, 40% of the final volume of S30 extract, and 16 µg/mL of the target protein vector. Do not vortex the reaction at this stage.
7. Load the reaction mix (1 mL) in the dialysis Gebaflex and the feeding mix (10 mL) in a 25 mL cylinder and incubate

overnight with stirring at a temperature optimal for the target protein (*see Note 15*).

8. Next morning, recover the supernatant by centrifuging 20 min at $10,000 \times g$ and proceed to purification according to the standard protocol of the protein. A dilution by a factor 4 is usually required in order to decrease sample ionic strength before loading it on a purification column.

3.4 *In Vitro* Production of Perdeuterated Protein in H₂O

The above described protocol enables mgs scale protein production for structural investigations. Dynamics and structural studies of proteins using NMR require uniform enrichment of the protein with stable ¹⁵N and ¹³C isotopes. Such uniform labeling schemes

Table 1
Preparation of reaction mix and feeding mix for CECF *in vitro* synthesis of protein

Amino acid mix	Volume (μL)
Water soluble AA: A, G, H, K, P, R, S, V, T 50 mM concentration for each AA	247.5
In 1 M HCl: C, D, E, L, M, N, Q, W, Y 50 mM concentration for each AA	247.5
In 1 M NaOH: I, F 50 mM concentration for each AA	247.5
RNase-free H ₂ O	82.5
Total	825.0

10× reaction mix	(μL)
100 mM rCTP	96.0
100 mM rGTP	96.0
100 mM rUTP	96.0
2.0 M HEPES–KOH pH 7.5	330.0
100 mM ATP	144.0
10 mM folinic acid	81.6
100 mM cyclic AMP	76.8
1 M DTT	40.8
1 M spermidine	24.0
9.2 M NH ₄ OAc	35.9
RNase-free H ₂ O	178.9
Total	1200

Reaction Mix (RM) and Feeding Mix (FM)	RM (μL)	FM (μL)
10 \times reaction mix	100.0	1000.0
1 M creatine phosphate	80.0	800.0
Amino acid mix	66.7	666.7
4 M potassium glutamate	52.0	520.0
1.07 M mg(OAc) ₂	7.9	130.8
17.5 mg/mL MRE 600 tRNA	10.0	0
10 mg/mL creatine kinase	25.0	0
T7 RNA polymerase (1/100e)	10.0	0
Adjust pH of FM to 7.5 with KOH		
S30 extract	400.0	0
Target DNA (16 $\mu\text{g}/\text{mL}$)	16.0	0
RNase-free H ₂ O	232.5	6882.5
Volume (μL) of mix	1000	10,000

are easily obtained using protocols described in part 3.3 by substituting unlabeled amino acids (Table 1) by U-(¹⁵N), or U-(¹⁵N, ¹³C)-labeled mix of amino acids or labeled algal extract available from isotope suppliers (*see* **Notes 16** and **17**). With the protocol described in this chapter, we typically obtain U-(¹⁵N), or U-(¹⁵N, ¹³C)-labeled protein in CECF mode with a yield of ca. 2 mg/mL of reaction mix. Production of perdeuterated proteins can be achieved by substitution of ¹H₂O by ²H₂O solvent [30], lyophilization of all components used in the cell-free reaction (Table 1 with exception of T7RNAPol) and use of perdeuterated amino acids mix (*see* **Note 16**). Experimentally, we observed that performing cell-free reactions in ²H₂O buffer reduces the protein yield by a factor of ca. 2. Furthermore, for NMR studies of large proteins, the destabilization of proteins with chaotropic agents is required to allow back-exchange of the backbone ¹H_N proton. Prior to structural investigation, the protein needs to be refolded in its native fold and this tedious step is usually associated with a low protein recovery yield. Alternatively, the use of U-(²H, ¹⁵N) or U-(²H, ¹⁵N, ¹³C)-labeled amino acids for in vitro protein synthesis in ¹H₂O solvent results in perdeuterated proteins fully protonated on all the exchangeable hydrogen positions [30] without requiring protein refolding. However, in this case, as the S30 extract contains active transaminases, perdeuterated amino acids added for the cell-free reaction are processed by enzymes catalyzing the exchange of α -deuteron by the solvent proton [30–32]. Although the level of

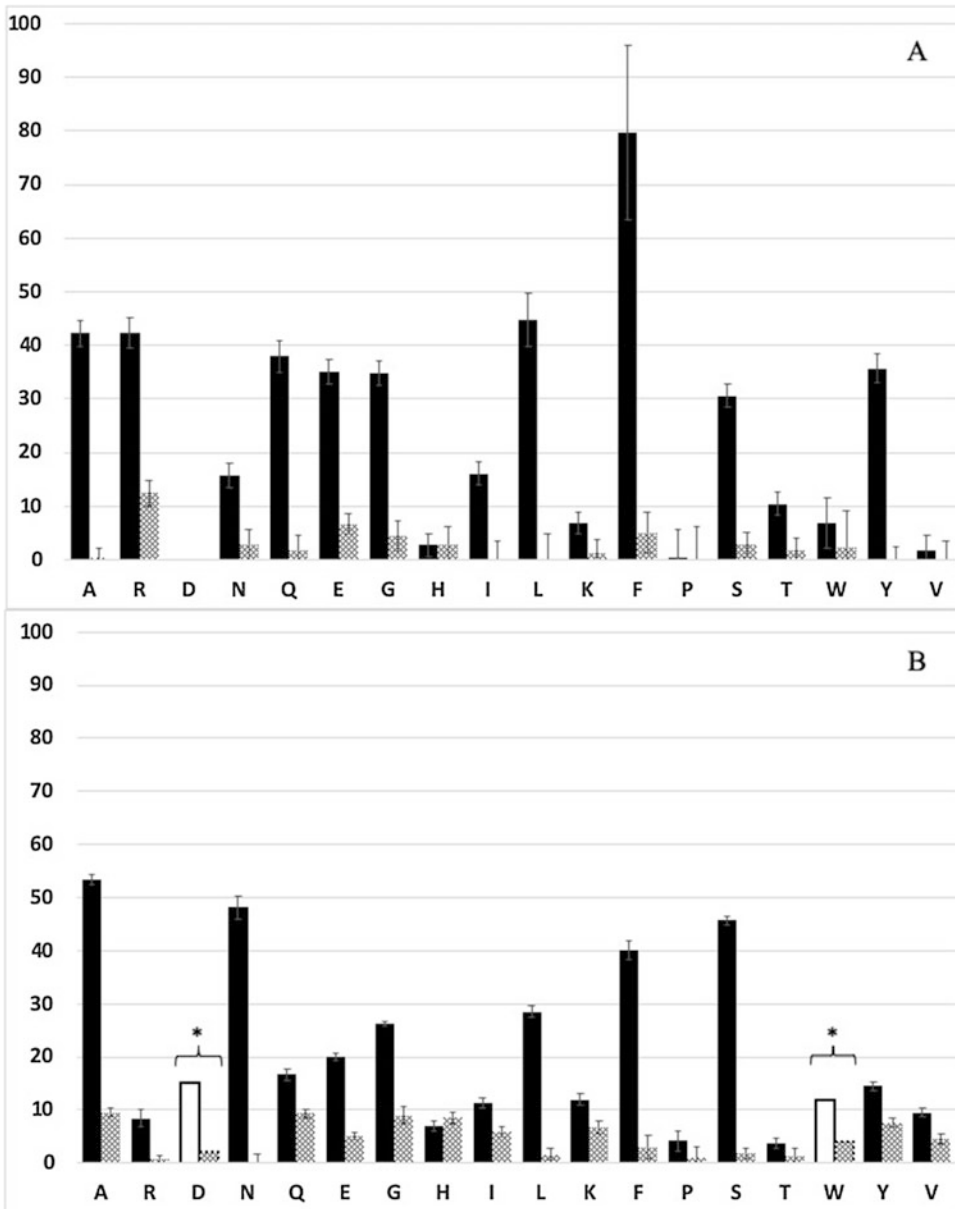


Fig. 2 Residual protonation levels (%) detected in C α -position of proteins expressed in vitro in $^1\text{H}_2\text{O}$ using perdeuterated amino mix. For quantification purpose the Nucleotide-Binding Domain of the P1B-type ATPase HMA8 from *Arabidopsis thaliana* (16.7 kDa, 155 amino acids [33]), was expressed as a model protein. Expression in vitro was undertaken in $^1\text{H}_2\text{O}$ following the protocols presented in Subheading 3 to produce perdeuterated U- ^2H , ^{13}C , ^{15}N] protein using the batch mode (a) (see Note 17) or the CECF mode (b) without inhibitors (black) or with inhibitors (AOA, D-malate and NaBH_4 S30 extract treatment) (gray). In (a) hydrolyzed U- ^2H , ^{13}C , ^{15}N] algal extract (Celtone[®] see Note 16) was complemented with unlabeled tryptophan. In (b) cell-free mixture containing 20 U- ^2H , ^{13}C , ^{15}N] amino acids (Merck) was used. Quantification of residual protonation was obtained from 3D HNCA experiments [34] using the integration module of the NMRPipe software [35]. To distinguish both isotopomers, the deuterium was decoupled during ^{13}C edition, while the ^1H

residual protonation is reduced compared with the corresponding one observed *in vivo* using $^1\text{H}_2\text{O}$ and perdeuterated amino acids (30–80% of residual protonation is observed on C_α [10, 19]), it remains substantial at an average level of ca. 25% with values varying for the different amino acids from a few percent up to ca. 50% (*see* Fig. 2). Such residual protonation on aliphatic protons on backbone atoms introduces heterogeneities in the samples and deteriorates the quality of most NMR spectra. To limit such residual protonation, the transaminases can be inactivated by the addition of inhibitors (AOA, D-malate) and NaBH_4 reduction of PLP cofactors with the following protocol [31, 32]. During its preparation, the S30 extract can be treated with NaBH_4 after the first dialysis step (*see* Subheading 3.1, **step 3**) and then stored similarly as the untreated extract. Residual protonation on C_α site in proteins produced *in vitro* using this treated S30 extract is reduced, with a maximum residual protonation of 10% for few amino acids, and an overall mean of less than 5% residual protonation (*see* Fig. 2). The NaBH_4 treatment and the addition of AOA and D-malate in the S30 are compatible with the cell-free protein production but usually result in a decrease of protein synthesis yield by a factor of ca. 2.

3.4.1 Inhibition of Transaminases from S30 Extracts

1. Solubilize the NaBH_4 powder at 100 mM in dimethylformamide (DMF) to keep its reductive power (*see* **Note 18**).
2. Treat the S30 extract by addition of NaBH_4 (at a final concentration of 20 mM) in a large container under gentle shaking at 4 °C and incubate the reaction for 10 min.
3. NaBH_4 and DMF are removed from the S30 extract by three dialysis steps against 100 volumes of cold S30 buffer using 12 kDa cutoff dialysis tubes.
4. Centrifuge the S30 extract in 50 mL tubes at $5000 \times g$ and 4 °C. Prepare 2 mL safe-lock tubes to aliquot the supernatant in 1 mL fractions and freeze the tubes immediately in liquid nitrogen. These NaBH_4 treated S30 extract aliquots can be stored for several years at –80 °C.

Fig. 2 (continued) decoupling was omitted. Percentage of ^1H or ^2H isotopes on C_α positions was deduced from the volume ratio of the correlations corresponding to the ^{13}C – ^1H and the ^{13}C – ^2H pairs. For each residue type, the results displayed are the percentage of residual protonation calculated from the mean of 2 to 7 amino acids (exceptions are for histidine in both production modes, phenylalanine and tryptophan in batch mode and proline in CECF mode, where only one amino acid was quantifiable). Quantification was not undertaken for methionine and aspartic acid because correlations were overlapping. Labeled tryptophan is absent in the sample produced in CECF mode (**b**) because it was added to the algal extract in an unlabeled form. The data for those amino acids, denoted by an asterisk (*), were therefore taken from Otting and coll [31]. There are no quantifications for cysteine as this protein does not have any in its sequence

3.4.2 *In Vitro* Synthesis of Perdeuterated Protein

In order to produce perdeuterated proteins with minimal residual protonation on aliphatic hydrogen sites in CECF mode, the following modifications need to be implemented in the protocol described in Subheading 3.3. The three unlabeled amino acids mixtures (**step 1**) should be replaced by an adequate mix of U-(^{15}N , ^2H) or U-(^{15}N , ^{13}C , ^2H) amino acids (*see Note 16*). The total volume of amino acids should be less than 825 μL (**step 3**), and the average concentration of each amino acid should be 15 mM (or 1 mM in final reaction or feeding mix). In order to avoid incorporation of unlabeled glutamate, the 208 mM of potassium glutamate (**step 5**) should be substituted by 150 mM of ammonium acetate. NaBH_4 treated S30 extract should be used at **step 6**, and AOA and DM should be added in S30 extracts at a concentration of 20 mM each following the protocol described above.

3.5 Application to the Production of Large Perdeuterated Proteins for NMR Investigations

The introduction of advanced isotopic labeling protocols, based on perdeuteration and selective protonation of few sites in the target protein, has enabled NMR investigations of very large protein assemblies [15, 36–38]. Recently, the near complete assignment of backbone heavy atoms of the 468 kDa homododecameric protein TET2 [39, 40] was obtained by solid state NMR [41, 42]. To accomplish this tour de force, Schanda and colleagues had to use several uniformly and specifically labeled samples, including perdeuterated samples. Surprisingly, despite that most frequencies of the backbone heavy atoms were assigned, less than half of the backbone $^1\text{H}_\text{N}$ nuclei frequencies could be identified. One of the factors limiting the assignment of amide proton frequencies is the high stability of this protein. Indeed, as the used samples were produced in $\text{M9}/^2\text{H}_2\text{O}$ *E. coli* cultures (*see Fig. 1a*), all hydrogen positions are thus deuterated and many of those located in the core of such stable protein are protected from the exchange with the buffer solvent, hampering back-protonation when TET2 is dialyzed against $^1\text{H}_2\text{O}$. The complete TET2 denaturation followed by its refolding in $^1\text{H}_2\text{O}$ solvent could represent a solution [37]. However, TET2 refolding is a tedious process, characterized by a very low protein recovery yield, precluding cost-effective production of such sample.

In order to protonate all the amide sites, we decided to produce perdeuterated TET2 protein directly in a $^1\text{H}_2\text{O}$ buffer using the cell-free synthesis protocol described above. The corresponding 2D-(^1H , ^{15}N) solid state NMR spectrum (*see Fig. 3*) displays well dispersed signals characteristic of folded proteins. In order to investigate the effects of incomplete back-protonation, a second sample has been produced using the same protocol but using $^2\text{H}_2\text{O}$ as a solvent during *in vitro* protein synthesis. The Fig. 4a, b presents the same zoom of a 2D-(^1H , ^{15}N)-CRINEPT-HMQC-TROSY [12] acquired in solution using both samples. A lot of amide proton signals observed when the sample is produced in $^1\text{H}_2\text{O}$ solvent (*see*

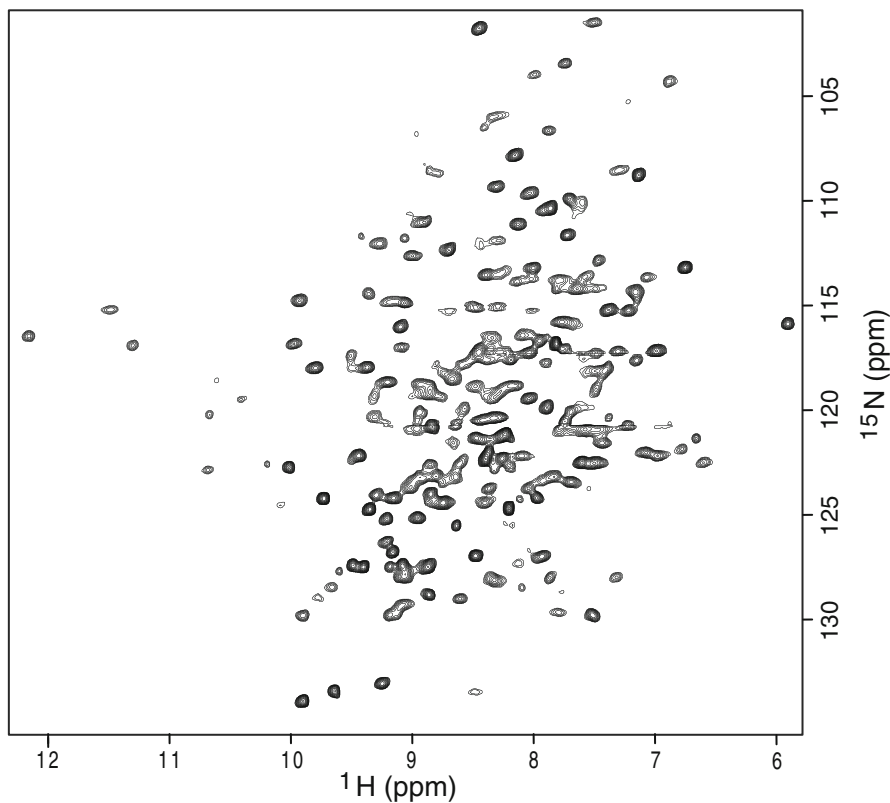


Fig. 3 2D- $(^{15}\text{N}, ^1\text{H})$ solid-state NMR spectra of protein TET2 acquired using a 600 MHz Bruker Avance III HD spectrometer equipped with a 1.3 mm triple-resonance MAS probe. U- $[^2\text{H}, ^{15}\text{N}]$ TET2 sample used to acquire the spectrum was produced *in vitro* (CECF mode), using $^1\text{H}_2\text{O}$ as a solvent during cell-free protein synthesis and amino acids mix from hydrolyzed U- $[^{15}\text{N}, ^2\text{H}]$ Celtone[®] complemented with unlabeled W and C amino acids. Sample was sedimented overnight at $65,000 \times g$ in a 1.3 mm rotor [43]. Heteronuclear transfers were performed using 750 ms cross polarization sequence (^1H : 15 kHz, ^{15}N : 40 kHz), the MAS frequency was set to 55 kHz and the effective temperature was 28 °C [42]

Fig. 4a) cannot be observed anymore when the sample is produced in $^2\text{H}_2\text{O}$ (see Fig. 4b), although the sample was previously incubated in $^1\text{H}_2\text{O}$ for 10 days to promote exchange of amide deuterons with solvent protons. Similarly, the sample used to acquire spectrum displayed on Fig. 4a was dialyzed in $^2\text{H}_2\text{O}$ for 10 days, before acquiring a new 2D $-(^1\text{H}, ^{15}\text{N})$ spectrum (see Fig. 4c). Most of the missing signals for the sample produced in $^2\text{H}_2\text{O}$ (see Fig. 4b) can still be observed with the sample produced in $^1\text{H}_2\text{O}$ but extensively dialyzed in $^2\text{H}_2\text{O}$ (see Fig. 4c). These experiments confirm that incomplete back-protonation of amide protons in large perdeuterated proteins is a major drawback for their investigations using NMR spectroscopy. As shown in this chapter, cell-free protein synthesis offers an efficient alternative to produce fully perdeuterated proteins with all solvent-exchangeable hydrogen sites fully occupied with ^1H spins, and without suffering from protonation artifacts on other sites. S30 extracts can be prepared in large scale, aliquoted and stored at $-80\text{ }^\circ\text{C}$, before being used to produce in 24 h perdeuterated proteins in milligram quantities. After

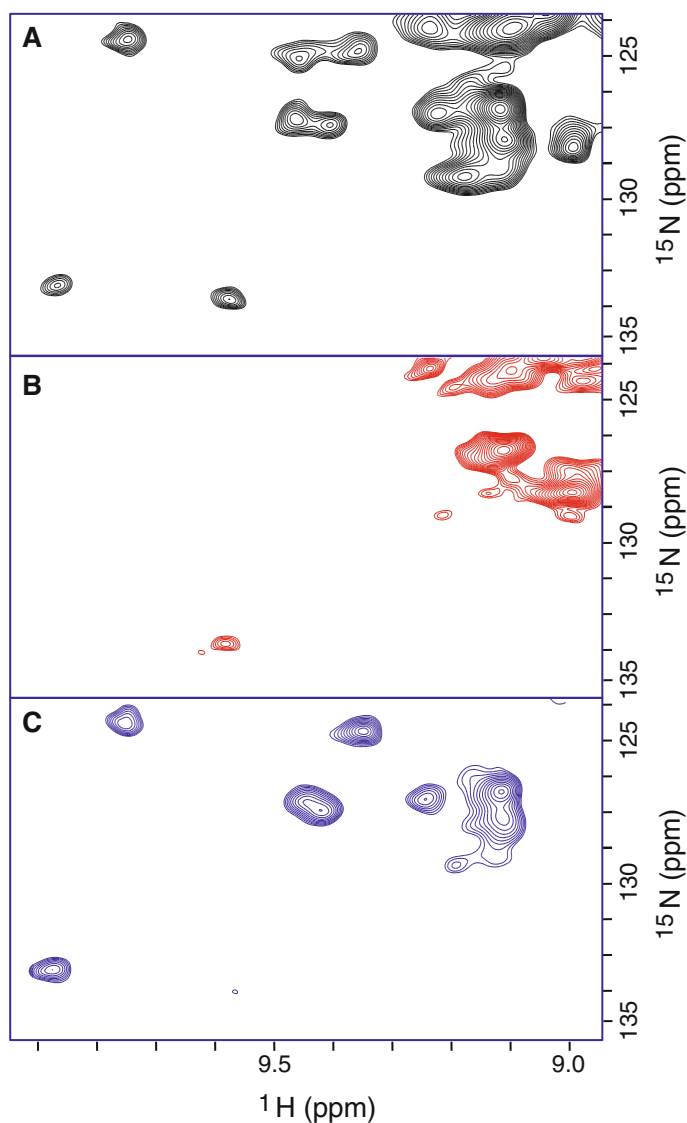


Fig. 4 2D-(^{15}N , ^1H) extracts from solution NMR spectra of the 468 kDa TET2 protein assembly using CRINEPT-HMQC-TROSY experiment [12] acquired on a 950 MHz Bruker Avance III HD spectrometer equipped with a 5 mm cryogenically cooled pulsed-field-gradient triple-resonance probe at a temperature of 50 °C. U- $[\text{}^2\text{H}$, $^{15}\text{N}]$ TET2 assemblies were produced in vitro (CECF mode), concentrated (200 μL at 10 μM) in 20 mM TRIS buffer (pH 7.4) containing 20 mM NaCl, and loaded in 4 mm Shigemi tubes. In (a) the sample used to acquire spectra presented was produced using $^1\text{H}_2\text{O}$ as a solvent during Cell-free protein synthesis and amino mix from hydrolyzed U- $[\text{}^{15}\text{N}$, $^2\text{H}]$ Celtone[®] complemented with unlabeled W and C amino acids. Sample (b) was produced using the same amino acids source but $^2\text{H}_2\text{O}$ as a solvent during cell-free synthesis, and the purified protein was then dialyzed in $^1\text{H}_2\text{O}$ buffer and stored at room temperature for 10 days to allow ^2H - \rightarrow ^1H substitution at exchangeable positions. Sample (c) was produced as described for sample (a), but the purified protein was dialyzed in $^2\text{H}_2\text{O}$ buffer and stored at room temperature for 10 days to allow ^1H - \rightarrow ^2H substitution in order to detect signals of solvent protected amide positions only

purification, such samples can be used directly to collect NMR data, without requiring tedious refolding of the perdeuterated proteins in $^1\text{H}_2\text{O}$ to ensure protonation of all NMR observable amide sites.

4 Notes

1. Alternatively, the 12 L culture could be performed in flasks up to an $\text{OD}_{600} = 1$, enabling the preparation of 50 mL of S30 extract.
2. Different *E. coli* strains can be used for S30 extract preparation, such as A19 [27], BL21 DE3 star [44], Rosetta pRARE [45], and codon-plus RIL [46].
3. In this protocol, we use a Techfors-S fermenter (INFORS HT) equipped with a steam generator allowing for in situ sterilization. Note that approximately 1 L of water is lost by evaporation during the autoclave of the culture medium. If the fermenter is not equipped with a steam generator, use instead sterile water (8 L) and 3 L of $4\times$ Z-media autoclaved separately.
4. Wait for 20 min before calibrating the 100% pO_2 at 37 °C, 550 rpm and at maximum airflow.
5. If pO_2 starts to increase, bacteria will not be in the exponential phase anymore and ribosomes will have a lower activity. For BL21(DE3), final OD_{600} should not exceed 3.2 (and culture should not last for more than 4 h to stay in the exponential growth phase). These values must be adapted for the different *E. coli* strains.
6. The complete cooling needs to be achieved in less than 20 min. Use chilled water flow at a temperature of less than 16 °C to cool down the double-walled fermentation vessels.
7. The pellet can easily be resuspended using a cell homogenizer such as Tissue Master 125 (OMNI International).
8. The total mass of wet cells for a 12 L culture is expected to be around 150 g. We recommend a short storage of cell pellets at 4 °C rather than freezing the cells.
9. Prepare the dialysis buffer in advance without DTT and store it at 4 °C. Add the DTT extemporaneously.
10. If the T7 RNAPol has a His-tag, it will coelute with the His-Tagged target protein during affinity purification. It has to be taken into account when the purification strategy is defined.
11. Precipitation of nucleic acids can be monitored by measurement of absorbance at 260 nm before and after addition of PEI.

12. The chromatography device needs to be RNase free. Flush extensively all the system with 0.5 to 1 M NaOH, RNase-free water, 2 M NaCl solution, and finally RNase-free water.
13. Analysis of T7RNAPol (MW: 98800 Da, molar ϵ : $140,000 \text{ M}^{-1} \cdot \text{cm}^{-1}$) is usually performed on 8% SDS-PAGE gel and by absorbance measurement.
14. Cutoff should be 12 kDa or less depending on the size of synthesized protein.
15. If protein is degraded or precipitated during cell-free expression, optimization of the duration and temperature of the synthesis should be performed to determine the optimal conditions.
16. The sources of labeled amino acids (AA) with different labeling schemes can be found from many providers (CIL, Merck, Cortecnet, ...). They are available as individual amino acids, 16 AA mix, or 20 AA mix ready-to-use for cell-free expression. The main advantage of individual amino acids is that users can adjust the concentration of each amino acid in accordance with the protein target sequence [47]. Algal extract (Isogro[®] or Celtone[®]) can also be used as a cheap source of 16 labeled amino acids. C, W, N, and E amino acids are degraded during Algal extract preparation and need to be added in the amino acid mix for cell free expression. Algal extracts contain ca. 60% of these 16 AA at different concentrations, in the form of free amino acids or small peptides. These small peptides can be enzymatically hydrolyzed by a protease (Pronase from *Streptomyces griseus*) in order to be incorporated during in vitro protein synthesis. To hydrolyze 1 g of algal extract in a volume of 20 mL (20 mM Tris-HCl pH 7.5), add 33 mg of this protease and incubate under stirring at 37 °C overnight. The next morning, proceed to a heat shock (95 °C for 30 min) to inactivate the protease. Centrifuge ($5000 \times g$, 15 min) and recover the supernatant containing the free amino acids. Hydrolyzed algal extract should be used at a final concentration of 3–5 mg/mL in the feeding and reaction mixtures.
17. In order to reduce the amount of expensive labeled amino acids, the cell-free reaction can be performed in Batch mode [27]. Synthesis in Batch mode is performed using only the reaction mix described in Subheading 3.3. Cell-free reaction by-products inhibit the protein synthesis and therefore the reaction is usually limited to 2–3 h. This protocol reduces quantities of labeled amino acids required by a factor of 10 but also decreases the protein synthesis yield by a factor of 3–5.
18. As NaBH₄ is a strong reducing agent activated by water, the stock solution must be prepared in DMF. NaBH₄ attacks the

aldehyde group of PLP and reduces the Schiff base occurring between PLP and NH group of transaminases. H₂ release is accompanied by the formation of foam. Work under a fume-hood.

Acknowledgments

The authors thank Prof. Eva Pebay-Peyroula for providing the clone of the HMA8 ATPases nucleotide-binding domain from *Arabidopsis thaliana*. This work used the high field NMR and Cell-Free facilities at the Grenoble Instruct-ERIC Center (ISBG; UMS 3518 CNRS-CEA-UGA-EMBL) within the Grenoble Partnership for Structural Biology (PSB). Platform access was supported by FRISBI (ANR-10-INBS-05-02) and GRAL, a project of the University Grenoble Alpes graduate school (Ecoles Universitaires de Recherche) CBH-EUR-GS (ANR-17-EURE-0003). IBS acknowledges integration into the Interdisciplinary Research Institute of Grenoble (IRIG, CEA). This work was supported by grants from CEA/NMR-Bio (research program C24990) and the Agence Nationale de la Recherche (ANR-17-CE29-0010 CH2-PROBE).

References

1. Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York
2. Bax A (1989) Two-dimensional NMR and protein structure. *Annu Rev Biochem* 58:223–256
3. Ikura M, Kay LE, Bax A (1990) A novel approach for sequential assignment of ¹H, ¹³C, and ¹⁵N spectra of larger proteins: heteronuclear triple-resonance NMR spectroscopy. Application to calmodulin. *Biochemistry* 29:4659–4667
4. Kay LE, Ikura M, Tschudin R, Bax A (1990) Three-dimensional triple resonance NMR spectroscopy of isotopically enriched proteins. *J Magn Reson* 213:442–445
5. Bax A (2011) Triple resonance three-dimensional protein NMR: before it became a black box. *J Magn Reson* 89:496–514
6. Ikura M, Clore GM, Gronenborn AM, Zhu G, Klee CB, Bax A (1992) Solution structure of a calmodulin-target peptide complex by multidimensional NMR. *Science* 256:632–638
7. Yamazaki T, Lee W, Arrowsmith CH, Muhandiram DR, Kay LE (1994) A suite of triple resonance NMR experiments for the backbone assignment of ¹⁵N, ¹³C, ²H labeled proteins with high sensitivity. *J Am Chem Soc* 116:11655–11666. (Article)
8. Venters RA, Huang C-C, Farmer BT II, Trolard R, Spicer LD, Fierke CA (1995) High-level ²H/¹³C/¹⁵N labeling of proteins for NMR studies. *J Biomol NMR* 5:339–244
9. Muchmore SW, Sattler M, Liang H, Meadows RP, Harlan JE, Yoon HS, Nettlesheim D, Chang BS, Thompson CB, Wong SL, Ng SL, Fesik SW (1996) X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. *Nature* 381:335–341
10. Gardner KH, Kay LE (1998) The use of ²H, ¹³C, ¹⁵N multidimensional NMR to study the structure and dynamics of proteins. *Annu Rev Biophys Biomol Struct* 27:357–406
11. 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
12. Riek R, Wider G, Pervushin K, Wüthrich K (1999) Polarization transfer by cross-correlated relaxation in solution NMR with very large molecules. *Proc Natl Acad Sci U S A* 96:4918–4923

13. Tugarinov V, Muhandiram R, Ayed A, Kay LE (2002) Four-dimensional NMR spectroscopy of a 723-residue protein: chemical shift assignments and secondary structure of malate synthase g. *J Am Chem Soc* 124:10025–10035
14. Wider G, Wüthrich K (1999) NMR spectroscopy of large molecules and multimolecular assemblies in solution. *Curr Opin Struct Biol* 9:594–601
15. Fiaux J, Bertelsen EB, Horwich AL, Wüthrich K (2002) NMR analysis of a 900K GroEL GroES complex. *Nature* 418:207–211
16. Sounier R, Blanchard L, Wu Z, Boisbouvier J (2007) High-accuracy distance measurement between remote Methyls in specifically protonated proteins. *J Am Chem Soc* 129:472–473
17. Fiaux J, Bertelsen EB, Horwich AL, Wüthrich K (2004) Uniform and residue-specific ¹⁵N-labeling of proteins on a highly deuterated background. *J Biomol NMR* 29:289–297
18. Rasia R, Noirclerc-Savoie M, Gallet B, Bologna N, Plevin M, Blanchard L, Palatnik J, Brutscher B, Vernet T, Boisbouvier J (2009) Parallel screening and optimization of protein constructs for structural studies. *Protein Sci* 18:434–439
19. Smith BO, Ito Y, Raine A, Teichmann S, Ben-Tovim L et al (1996) An approach to structure determination using limited NMR data from larger proteins selectively protonated at specific residue types. *J Biomol NMR* 8:360–368
20. Zubay G (1973) In vitro synthesis of protein in microbial systems. *Ann Rev Genet* 7:267–287
21. Pratt C (1980) Kinetics and regulation of cell-free alkaline phosphatase synthesis. *J Bacteriol* 143(3):1265–1274
22. Roberts BE, Paterson BM (1973) Efficient translation of tobacco mosaic virus RNA and rabbit globin 9S RNA in a cell-free system from commercial wheat germ. *PNAS* 70:2330–2334
23. Klammt C, Löhr F, Schafer B, Haase W, Dötsch V, Ruterjans H, Glaubitz C, Bernhard F (2004) High level cell-free expression and specific labeling of integral membrane proteins. *Eur J Biochem* 271:568–580. <https://doi.org/10.1111/j.1432-1033.2003.03959.x>
24. Davanloo P, Rosenberg AH, Dunn JJ, Studier FW (1984) Cloning and expression of the gene for bacteriophage T7 RNA polymerase. *Proc Natl Acad Sci U S A* 81:2035–2039
25. Zawadzki V, Gross HJ (1991) Rapid and simple purification of T7 RNA polymerase. *Nucleic Acids Res* 25:1948
26. Spirin AS, Baranov VI, Ryabova LA, Ovodov SY, Alakhov YB (1988) A continuous cell-free translation system capable of producing polypeptides in high yield. *Science* 242:1162–1164
27. Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T, Yokoyama S (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett* 442:15–19
28. Martin GA, Kawaguchi R, Lam Y, DeGiovanni A, Fukushima M, Mutter W (2001) High-yield, in vitro protein expression using a continuous-exchange, coupled transcription/translation system. *BioTechniques* 31:948–953
29. Haberstock S, Roos C, Hoevels Y, Dötsch V, Schnapp G, Pautsch A, Bernhard F (2012) A systematic approach to increase the efficiency of membrane protein production in cell-free expression systems. *Protein Expr Purif* 82:308–316
30. Etezady-Esfarjani T, Hiller S, Villalba C, Wüthrich K (2007) Cell-free protein synthesis of perdeuterated proteins for NMR studies. *J Biomol NMR* 39:229–238
31. Su XC, Loh CT, Qi R, Otting G (2011) Suppression of isotope scrambling in cell-free protein synthesis by broadband inhibition of PLP enzymes for selective ¹⁵N-labelling and production of perdeuterated proteins in H₂O. *J Biomol NMR* 50:35–42
32. Yokoyama J, Matsuda T, Koshiba S, Tochio N, Kigawa T (2011) A practical method for cell-free protein synthesis to avoid stable isotope scrambling and dilution. *Anal Biochem* 411:223–229. <https://doi.org/10.1016/j.ab.2011.01.017>
33. Mayerhofer H, Sautron E, Rolland N, Catty P, Seigneurin-Berny D, Pebay-Peyroula E, Ravaud S (2016) Structural insights into the nucleotide-binding domains of the P1B-type ATPases HMA6 and HMA8 from *Arabidopsis thaliana*. *PLoS One* 11(11):e0165666. <https://doi.org/10.1371/journal.pone.0165666>
34. Lescop E, Schanda P, Brutscher B (2007) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. *J Magn Reson* 187:163–169
35. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6:277–293
36. Sprangers R, Kay LE (2007) Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature* 445:618–622
37. Macek P, Kerfah R, Boeri Erba E, Crublet E, Moriscot C, Schoehn G, Amero C, Boisbouvier J (2017) Unraveling self-assembly pathways of

- the 468 kDa Proteolytic machine TET2. *Sci Adv* 3(4):e1601601. <https://doi.org/10.1126/sciadv.1601601>
38. Mas G, Guan J-Y, Crublet E, Colas Debled E, Moriscot C, Gans P, Schoehn G, Macek P, Schanda P, Boisbouvier J (2018) Structural investigation of a chaperonin in action reveals how nucleotide binding regulates the functional cycle. *Sci Adv* 4(9):eaau4196. <https://doi.org/10.1126/sciadv.aau4196>
39. Franzetti B, Schoehn G, Hernandez JF, Jaquinod M, Ruigrok RW, Zaccari G (2002) Tetrahedral aminopeptidase: a novel large protease complex from archaea. *EMBO J* 21:2132–2138
40. Borissenko L, Groll M (2005) Crystal structure of TET protease reveals complementary protein degradation pathways in prokaryotes. *J Mol Biol* 346:1207–1219
41. Fraga H, Arnaud CA, Gauto DF, Audin MJC, Kurauskas V, Macek P, Krichel C, Guan JY, Boisbouvier J, Sprangers R, Breyton C, Schanda P (2017) Solid-state NMR H-N-(C)-H and H-N-C-C 3D/4D correlation experiments for resonance assignment of large proteins. *Chem Phys Chem* 18:2697–2703
42. Gauto D, Estrozi L, Schwieters C, Effantin G, Macek P, Sounier R, Sivertsen AC, Schmidt E, Kerfah R, Mas G, Colletier JP, Güntert P, Favier A, Schoehn G, Schanda P, Boisbouvier J (2019) Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. *Nat Commun* 10(1):1234567890. <https://doi.org/10.1038/s41467-019-10490-9>
43. Bertini I, Luchinat C, Parigi G, Ravera E, Reif B, Turano P (2011) Solid-state NMR of proteins sedimented by ultracentrifugation. *Proc Natl Acad Sci U S A* 108:10396–10399. <https://doi.org/10.1073/pnas.1103854108>
44. Torizawa T, Shimizu M, Taoka M, Miyano H, Kainosho M (2004) Efficient production of isotopically labeled proteins by cell-free synthesis: a practical protocol. *J Biomol NMR* 30:311–325
45. Apponyi MA, Ozawa K, Dixon NE, Otting G (2008) Cell-free protein synthesis for analysis by NMR spectroscopy. *Methods Mol Biol* 426:257–268. https://doi.org/10.1007/978-1-60327-058-8_16
46. Kigawa T, Yabuki T, Matsuda N, Matsuda T, Nakajima R, Tanaka A, Yokoyama S (2004) Preparation of *Escherichia coli* cell extract for highly productive cell-free protein expression. *J Struct Funct Genom* 5:63–68
47. Pedersen A, Hellberg K, Enberg J, Karlson G (2011) Rational improvement of cell-free protein synthesis. *New Biotechnol* 28:218–224. <https://doi.org/10.1016/j.nbt.2010.06.015>



Minimizing Heterogeneity of Protein Samples for Metal Transporter Proteins Using SAXS and Metal Radioisotopes

Shah Kamranur Rahman

Abstract

The scattering profiles at small angles, obtained after an X-ray beam is incident on biological samples (protein), are nowadays successfully used to obtain important structural information. Small angle X-ray scattering (SAXS) is now helpful in providing information about shape, conformation, and assembly state of molecules, besides macromolecular folding–unfolding, aggregation, and extended conformations. The article discusses here a protocol to identify those fractions of heterogeneous proteins that are rich in homogeneous samples, testified by proper conformation and protein activity. The protocol in reference to a class of proteins known as metal binding (transporter) proteins or ion channels is discussed using applications of SAXS and metal radioisotopes. With requisite modifications, the protocol can be adapted to other classes of proteins.

Key words SAXS, Gel filtration, Metal binding protein, Radioactive isotope

1 Introduction

When the purified protein sample is eventually subjected to gel filtration, it resolves the otherwise heterogeneous populations into different fractions of elution. We were able to exploit this trait of gel filtration [1] to find out the best suitable homogeneous population for any purified protein sample. Here, we are discussing this protocol with reference to the metal binding protein using SAXS and metal binding activity. The elution fraction with maximum metal binding and activity can yield the best possible homogeneous population for further analysis by single-particle cryo-EM or crystallography. Even if the peak is single still each fraction has a sample with protein molecules in a particular orientation. Even if there is a single peak for a sample it actually contains a mixture of populations of proteins probably in different conformations, contributing to conformational heterogeneity. For metal binding proteins or channels, only a small population (a subfraction) of this single peak will have the ion channels properly formed and

showing the corresponding optimal activity. Each drop of the filtrate from gel filtration can be analyzed by SAXS [2] to check which subfraction has the size corresponding to properly folded ion channel. The protein population with the correct size can be tested for radioisotope metal activity. Taken together, the fraction that shows correct size corresponding to a properly folded ion channel with maximum expected activity is the most homogeneous (and ideal) sample that can be analyzed downstream for single-particle analysis under cryo-EM or for crystallography.

2 Materials

1. Size Exclusion Column Superdex 200 10/300.
2. FPLC AKTA instrument.
3. Buffer: *PBS* (phosphate buffer saline).
4. Disposable, sterile syringes, needles, and 0.22 μM filter.
5. Radioactive setup: Pyrex glass sheets and boxes to shield sample away from the user. Incubation buffer (30 mM Tris-Cl, 150 mM NaCl, pH 7.4). Radioactive $^{45}\text{CaCl}_2$ salt.
6. Small angle X-ray scattering instrumental setup, either in-house or access to the beamline. The gel filtration attachment setup that connects the gel filtration assembly with the SAXS instrument.

3 Methods

3.1 Gel Filtration

1. Fix the Size Exclusion Column Superdex 200 10/300 GL in a vertical position to the stand near FPLC AKTA instrument.
2. Open the top knob, pour some drops of buffer onto it, then the bottom cap of the column avoiding any entry of air into the resin.
3. Equilibration of column: pass two column volumes (CV) of buffer at a flow rate of 0.2 mL/min through the resin with PBS to equilibrate the resin.
4. Inject around 500 μL protein sample carefully to the column without disturbing the column surface.
5. Elution of sample: Elute the column with 1.5 CV of buffer at a flow rate of 0.2 mL/min.
6. Collect the fractions of the single peak (Fractions 1–9), each 0.5 mL in volume, for further analysis by SAXS (Fig. 1).

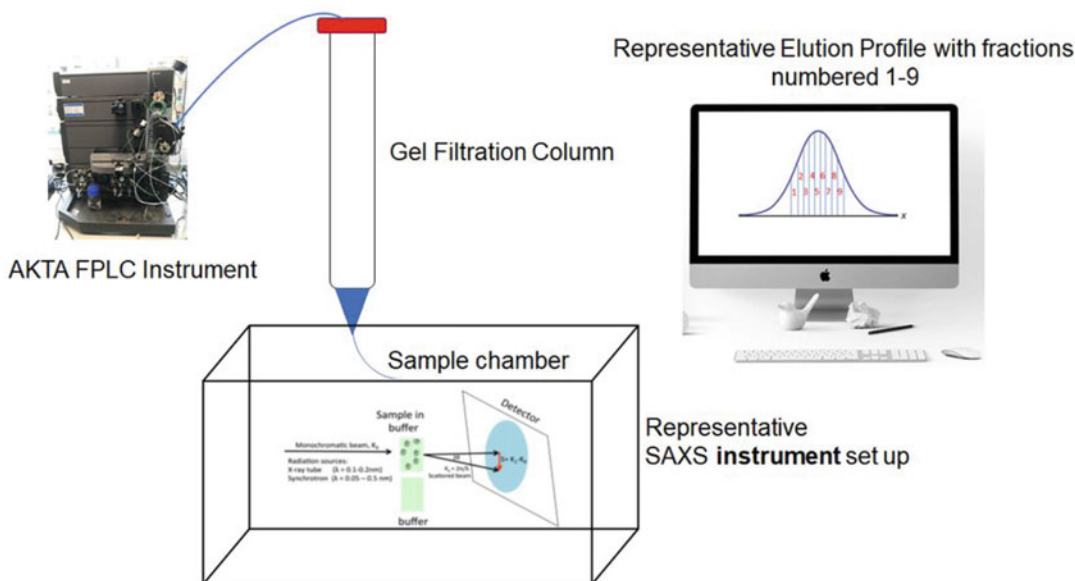


Fig. 1 The figure shows protein fractions (1–9) that can be individually tested by SAXS for correct size. Of these nine fractions, the one with the correct size and best metal binding activity is ideal and homogeneous

3.2 SAXS

1. Centrifuge the protein samples in PBS first at 13,000 rpm ($15871 \times g$) for 30 min at 6 °C using microfuge to remove very large aggregates, if they settling down the 1.5 mL centrifuge tube. Purify the supernatant obtained after centrifugation, using Vivaspin[®] (MWCO 1,000,000 Da, Sartorius).
2. Wash Vivaspin tube once with water and twice with PBS before adding protein sample to it. Place the samples in the upper chamber of the Vivaspin 500 tubes and centrifuge at 10,000 rpm ($12208 \times g$) for 5 min at 6 °C. The fraction that comes down is expected to be free of aggregates of the order of 1 MDa that is left out in the upper chamber. Take the purified protein sample from the lower chamber for X-ray exposure.
3. Check any observed radiation damage during X-ray exposure by comparing 10 successive time frames with 15-s exposures. Analyze data sets at different concentrations to detect possible interparticle interactions to find out any tendency of protein sample to aggregate.
4. Determine Rg value of each fraction of the elution peak to check for any sudden increase in Rg value, thus confirming that protein samples are free from aggregates [3].
5. Also, observe SAXS profile of each frame to confirm absence of aggregation or radiation damage of samples. Process the data using ATSAS program package [4].

6. Normalize the data to the intensity of transmitted beam. Subtract from the scattering profile of each protein sample fraction, the scattering of buffer using SCATTER software.
7. Average the subtracted SAXS profiles that came as output files using PRIMUS [5] and SCATTER [6].
8. After averaging of subtracted profile calculate the forward scattering $I(0)$ and the radius of gyration, R_g , using Guinier approximation [7] (*see Note 1*). Additionally, employ a Kratky plot (*see Note 1*) to qualitatively compare overall conformational state of each fraction of the elution profile [8, 9].
9. To obtain entire scattering pattern run GNOM program [10], which provides maximum particle dimensions D_{max} and distance distribution functions $P(r)$ (*see Note 2*). Estimate the apparent molecular masses of each fraction by comparing forward scattering of samples with reference solutions of bovine serum albumin (molecular mass 66 kDa). Combine the SAXS data of each sample taken in different fractions using SCATTER.

3.3 Radioactive Labeling of each Fraction

1. Purify protein sample and incubate with Chelex[®] 100 molecular biology grade resin to remove any bound metal in metalloproteins.
2. Run a parallel gel filtration and collect each fraction similarly as done with the samples meant for SAXS analysis.
3. Incubate sample of each fraction with radioactive incubation buffer (30 mM Tris-HCl, 150 mM NaCl, pH 7.4) supplemented with radioactive ^{45}Ca as 10 mM $^{45}\text{CaCl}_2$ in 10 μL volume for 30 min at 25–30 °C.
4. Run the sample on 10% SDS-PAGE gel, for 4 h at 90 V. Dry the gel and place in the cassette with X-ray film or exposure sheet overnight. Analyze the radioactive exposure on film under analyzer.
5. Compare the intensity ^{45}Ca of each fraction with the corresponding envelope model developed from SAXS data for the protein sample.
6. Identify precisely the ideal sample that shows optimum binding and conformation; in other words, a particular fraction that is better than others is chosen and sent out for single-particle cryo-EM analysis.

4 Notes

1. The most straightforward way for deducing forward scattering $I(0)$ and the radius of gyration, R_g , is Guinier method [6]. The fundamental to which is Guinier equation defined as:

$$I(q) = I(0) \frac{1}{3} R_g^2 q^2$$

Alternatively, the equation in linear ($y = mx + c$) form is given as follows:

$$\ln I(q) = -\frac{R_g^2}{3} \cdot q^2 + \ln I(0).$$

where q is vector subtraction after diffraction. The representative plot based on linear equation is first plotted. The equation suggests that Guinier approximation is based on power law expansion ($\ln [I(q)]$ Vs q^2); therefore, a Guinier plot ($\ln [I(q)]$ versus q^2) can give $I(0)$ and R_g value from y -axis intercept and slope of linear region. However, this linear region is obtained after fitting of data points at low q -range (between q_{\min} or q_0 and q_{\max}) such that the curve is kept linear. The rationale behind determination of q -range is shape of protein. The low q -range is decided such that $q_{\max} \times R_g = 1.3$ (globular proteins) or $q_{\max} \times R_g = 0.8$ (Collagen like proteins). In principle, there are mainly two methods to obtain R_g value, one is automated method and the other a manual fitting method. The difference between two methods is that automated method fits data points without trimming bad data points (with significant errors) in the linear curve and q_{\min} is q_0 . This can be done by using AUTORG [11, 12] tool of ATSAS package or by SCATTER [6]. However, in manual method data points with significant deviation are trimmed off by adjusting q -range such that q_{\min} is not q_0 and q_{\max} is selected such that $q_{\max} \times R_g < = 1.3$ or 0.8 .

2. Likewise, the distance distribution function $P(r)$ is obtained by the following formula:

$$p(r) = \frac{r}{2\pi^2} \int_0^\infty q^2 I(q) \frac{\text{sinsin}(qr)}{qr} dq$$

The next step in data processing is to calculate molecular mass (MM) of sample, which is estimated by comparing forward scattering intensity of protein sample with scattering intensity of a known standard, say, bovine serum albumin or lysozyme.

References

1. Andrews P (1965) The gel-filtration behaviour of proteins related to their molecular weights over a wide range. *Biochem J* 96:595–606
2. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc* 129:5656–5664. <https://doi.org/10.1021/ja069124n>. Epub 2007 Apr 6
3. Debye P (1947) Molecular-weight determination by light scattering. *J Phys Colloid Chem* 51(1):18–32
4. Konarev PV (2006) ATSAS 2.1, a program package for small-angle scattering data analysis. *J Appl Crystallogr* 39:277–286
5. Roessle MW et al (2007) Upgrade of the small-angle X-ray scattering beamline X33 at the European molecular biology laboratory Hamburg. *J Appl Crystallogr* 40:s190–s194
6. <http://www.bioisis.net/>. Accessed 30 June 2016
7. Guinier A (1939) *Ann Phys* 12:161
8. Glatter O, Kratky O (1982) *Small angle x-ray scattering*. Academic Press, London
9. Glatter O (1977) A new method for the evaluation of small-angle scattering data. *J Appl Crystallogr* 10:415–421
10. Chacon P et al (2000) Reconstruction of protein form with X-ray solution scattering and a genetic algorithm. *J Mol Biol* 299:1289–1302
11. Petoukhov MV et al (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Crystallogr* 45 (Pt 2):342–350
12. Petoukhov MV et al (2007) ATSAS 2.1 - towards automated and web supported small-angle scattering data analysis. *J Appl Crystallogr* 40:s223–s228

Part II

Structural Analyses and Data Management



Hydrogen–Deuterium Exchange Mass Spectrometry for Probing Changes in Conformation and Dynamics of Proteins

Pui-Kin So

Abstract

Hydrogen–deuterium exchange mass spectrometry (HDX-MS) is, nowadays, an increasingly important technique in studying protein conformation and dynamics. This technique possesses the advantages of low sample consumption, less limitation in protein size, and relatively simple experimental workflow. An HDX-MS experiment typically includes the steps of sample preparation, HDX reaction, quenching of HDX reaction, protease digestion, and LC-MS analysis. Although HDX-MS has been an established technique and automatic sample handling devices are commercially available nowadays, proper experimental conditions of each step are crucial for a successful HDX-MS experiment. This chapter is to provide a general guideline for each step in the HDX-MS workflow and highlight some precautions needed to be taken in order to acquire useful conformational and dynamic information.

Key words Hydrogen–deuterium exchange, Protein conformation, Protein dynamics, Mass spectrometry, Protein–ligand interaction

1 Introduction

Hydrogen–deuterium exchange mass spectrometry (HDX-MS) is, nowadays, an important technique for studying protein conformation and dynamics. This technique is widely applied not only to fundamental researches, including studies on structural and dynamic aspects of protein–drug interactions and protein–protein interactions, and effect of point mutations on protein functions, but also to industrial applications, such as determination of structural integrity of biopharmaceutical agents (e.g., antibodies) [1–3].

The prevalence of HDX-MS could be due to its several distinct advantages compared to other biophysical techniques. For example, this technique is featured by significantly low consumption of protein samples, typically femtomole to picomole level, because of the high sensitivity of modern mass spectrometers [1–3]. In addition, this technique has less limitation in protein size; therefore,

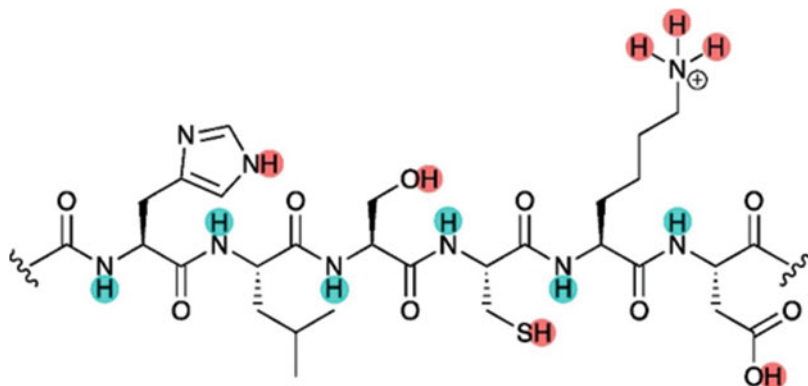


Fig. 1 Exchangeable hydrogens in a polypeptide chain. Hydrogens on amide bonds and side chains are highlighted with blue and red, respectively. (Reprinted from ref. 3 with permission from Elsevier)

study of large proteins and protein complexes is feasible [1–3]. Furthermore, the sample preparation involved is relatively simple and does not involve procedures requiring extensive optimization of conditions, such as production of quality protein crystals [1–3]. The commercial availability of online automatic sample handling devices in recent years has also made HDX-MS more easily accessible.

In HDX-MS, a protein is incubated in a deuterium solution, from which hydrogens on the protein, including hydrogens on amide bonds, amino acid side chains, and C- and N-termini, undergo exchange with deuterium in the solution (Fig. 1) [1–3]. However, only exchange of amide hydrogens is monitored, because the exchange of hydrogens on side chains and N- and C-termini is too rapid to be monitored and deuteriums incorporated on these regions often are back-exchanged with hydrogens when exposed to protic solvents during analysis [1–3]. As exchanging a hydrogen with a deuterium results in a mass increase of ~1 Da, the degree of HDX can be readily monitored by mass spectrometry. Typically, amide hydrogens at more buried (less solvent accessible) regions or regions with extensive hydrogen bond network exchange slower than those at exposed regions (more solvent accessible) or regions with less degree of hydrogen bond interactions [1–3]. Therefore, changes in conformation and dynamics of proteins can be reflected by changes in HDX behavior of amide hydrogens. In most cases, a pair or a couple of protein samples at different conditions (e.g., wild type vs mutant and unbound vs ligand-bound) are subjected to HDX-MS analysis in parallel under identical experimental conditions and their HDX properties are compared for characterization of conformational and dynamic changes.

Determination of conformational and dynamic change of proteins by HDX is typically performed with a continuous labeling

approach [1–3]. Another approach, pulsed labeling [4], for study of protein folding kinetics, is not included in the scope of this chapter.

In the continuous labeling approach, a protein is first incubated in a deuterium buffer to initiate HDX. At different exchange time intervals, the exchange reaction is quenched by addition of a chilled quenching buffer to lower the pH of the solution to ~2.5, at which the rate of HDX is minimum [1–3]. Typically, at least 5–8 time points are obtained in order to acquire a comprehensive HDX time profile for comparison. To ensure changes in HDX behavior observed are statistically significant, at least three independent replicates should be obtained for each HDX time point. To obtain global conformational and dynamic information, exchange-quenched samples are directly analyzed by liquid chromatography–mass spectrometry (LC-MS) for determination of global deuterium incorporation levels from measured molecular mass of intact proteins. For investigating conformation and dynamics of local protein regions, exchange-quenched samples are digested by an acid-stable protease (e.g., pepsin) and subsequently the digested samples are analyzed by LC-MS for determining the deuterium uptake levels of various peptides.

An HDX-MS experiment involves multiple steps and proper experimental conditions in each step are critical for the successfulness of an experiment. Particularly, special precautions must be taken to minimize the problem of back-exchange, which refers to the fact that deuteriums incorporated on a protein exchange back to hydrogens when exposed to protic solutions during the HDX-MS workflow [1–3]. Although the experimental procedure and conditions could substantially vary depending on the purposes of experiments and hardware setup, this chapter is to provide a general guideline for each experimental step in an HDX-MS experiment and describe the important measures that must be taken in order to obtain useful information on protein conformation and dynamics.

2 Materials

2.1 *Centrifugal Ultrafiltration Devices for Buffer Exchange*

Centrifugal ultrafiltration devices with a wide range of molecular weight cutoffs (e.g., 1–1000 kDa) are commercially available.

2.2 *HDX Buffer*

50–500 mM buffer commonly used in protein chemistry, such as ammonium acetate, phosphate, and Tris, but prepared in deuterium oxide (D₂O) instead of water (H₂O). The pH is adjusted to physiological pH (i.e., ~7; pH reading is 0.4 lower than pD) of the protein system under investigation.

- 2.3 Quenching Buffer** A buffer compositionally the same as the HDX buffer adjusted to acidic pH, such that the pH of a 1:1 mixture of the quenching buffer and HDX buffer is 2–3. Store the quenching buffer at 0 °C to minimize back-exchange.
- 2.4 Proteases** Common commercially available acid-stable proteases include pepsin, protease type XIII (aspergillopepsin), and protease type XVIII (rhizopuspepsin) [1–3, 5]. Immobilized columns of these proteases and their mixtures are also commercially available. Other proteases, such as plasmepsins [6], *Aspergillus niger* prolyl endoprotease [7], aspartic protease nepenthesin-1 [8], and aspartic protease from rice field eel [9], have also demonstrated to deliver desired digestion efficiency under acidic conditions, yet these proteases are not commercially available; thus, additional efforts and materials are required for production and purification.
- 2.5 LC Trap Column** C4 columns for intact proteins, C18 columns for peptides.
- 2.6 LC Analytical Column** C4 columns for intact proteins, C18 columns for peptides.
- 2.7 LC Solvents** Trapping solvent: 95–100% HPLC grade or milliQ Water: 0–5% HPLC grade acetonitrile (ACN) with 0.1% formic acid.
Aqueous mobile phase: HPLC grade or milliQ Water with 0.1–1% formic acid.
Organic mobile phase: HPLC grade acetonitrile (ACN) with 0.1–1% formic acid.
- 2.8 Instrumentation**
1. Liquid chromatography–high resolution mass spectrometer (e.g., quadrupole time-of-flight, Orbitrap, and Fourier-transform ion cyclotron resonance mass spectrometer).
 2. Automatic sample handling device for HDX and quenching (optional).
- 2.9 MS-Compatible, Nonionic Detergent** For membrane proteins only. n-Dodecyl- β -D-maltopyranoside (DDM), n-Octyl- β -D-glucopyranoside (OG), or Triton X-100 (TX100) at or slightly higher than the critical micelle concentration (CMC) (*see step 2* of Subheading 3.1).

3 Methods

The workflow for a continuous labeling HDX-MS experiment typically includes sample preparation, hydrogen–deuterium exchange (HDX) and quenching, protease digestion, LC-MS analysis, and data analysis for calculation of deuterium uptake (Fig. 2). At least three independent replicates should be obtained for each



Fig. 2 A general experimental workflow for continuous labeling HDX-MS. (Reprinted from ref. 3 with permission from Elsevier)

HDX time point in order to obtain statistically significant results. Details of individual step are described as follows:

3.1 Sample Preparation

1. Prepare medium to high micromolar range, typically 10–100 μM , of a protein in a suitable buffer (undeuterated). If needed, perform buffer exchange using centrifugal ultrafiltration devices with suitable molecule weight cutoff (refer to Subheading 2.1). Note that the use of samples with overly high concentration could potentially lead to severe carryover problem, that is, signals from residual samples in the previous run occur in the current run. (*see Notes 1-3* for the details of carryover problem and the suitability of protein concentration and buffer system).
2. For membrane proteins, add a MS-compatible nonionic detergent to form lipid micelles for maintaining the stability and solubility of proteins [10, 11]. The concentration of detergents is typically at or slightly higher than the critical micelle concentration (CMC) [10]. n-Dodecyl- β -D-maltopyranoside (DDM) has been one of the most widely used MS-compatible nonionic detergent, while many other nonionic detergents, such as n-octyl- β -D-glucopyranoside (OG) and Triton X-100 (TX100), have also been successfully utilized in MS applications. The CMC of DDM, OG, and TX100 are 0.0087%, 0.53%, and 0.015%, respectively (common nonionic detergents and their CMC can be found in ref. 10) [10, 12] (*see Notes 4 and 5* for precautions and alternative methods in handling membrane proteins).
3. For study of change in conformation and dynamics upon binding with a ligand, that is, drug or protein, add the binding partner to the protein solution to initiate protein binding. Refer to the dissociation constant (K_d) of the binding reaction, if available, to access the amount of binding molecule to be added. For weak binding, addition of excessive amount of binding partner may be required to achieve a high degree of binding.

3.2 Hydrogen–Deuterium Exchange (HDX) and Quenching

HDX and quenching can be performed automatically with automatic sample handling devices, which are commonly online with LC-MS systems, or manually by hand mixing.

1. Diluted the protein 10–20 fold with the HDX buffer (refer to Subheading 2.2 about HDX buffer) for initiating HDX reaction.
2. At different reaction time points, typically from tens of seconds to hours depending on protein systems, mix the sample under exchange with equal volume of chilled quenching buffer (refer to Subheading 2.3 about quenching buffer). At least five time points should be made over the entire HDX time profile.
3. For the use of automatic sample handling devices, exchange-quenched samples are typically digested and analyzed by LC-MS immediately to minimize the period allowed for back-exchange. For manual operation, exchange-quenched samples should be frozen with liquid nitrogen immediately and stored under $-80\text{ }^{\circ}\text{C}$ until further processing or analysis.

3.3 Protease Digestion

Protease digestion is an important step determining the spatial resolution that can be achieved in an HDX-MS study. This process can be performed online with commercially available immobilized protease columns or offline through solution mixing. For online digestion, the immobilized protease column applied is one part of the LC flow path, typically in front of the trap column (Fig. 3) (also refer to step 1, i.e., Trapping, in Subheading 3.4). Digestion takes place when the exchange-quenched sample is passed through the protease column during the trapping process (also refer to step 1, i.e., Trapping, in Subheading 3.4). Because the digestion process is part of the automatic sample injection cycle and less sample handling is involved, online digestion with immobilized columns is a convenient approach and the mainstream in HDX-MS nowadays (*see* Notes 6–9 for optimization of conditions for protein digestion to obtain desired sequence coverage).

3.3.1 Online Digestion

1. Connect the immobilized protease column between the sample loop and trap column (*see* Note 10 for precautions in storage and evaluation of efficiency of immobilized protease columns).
2. Ensure the instrument is operated in trapping mode.
3. Rinse the immobilized protease column with trapping solvent (refer to Subheading 2.6) for at least 15 min to equilibrate the column.
4. Inject exchange-quenched samples for online digestion.

3.3.2 Offline Digestion

1. Prepare a solution of active digestive enzyme by dissolving the protease (pepsin, protease type XIII (aspergillopepsin) and

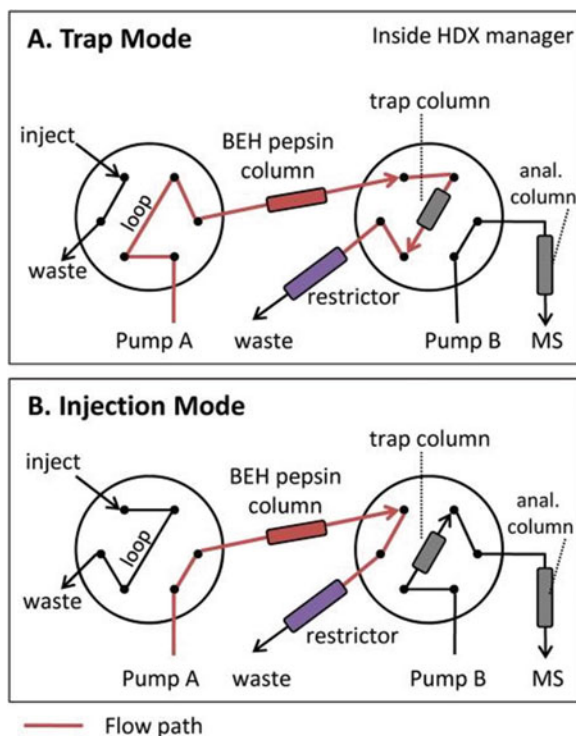


Fig. 3 A schematic diagram showing the LC flow path during (A) trapping and (B) LC separation. (Reprinted from ref. 13 with permission from American Chemical Society)

protease type XVIII (rhizopuspepsin)), in the quenching buffer prechilled to 0 °C in an ice bath.

- Spin down the protease slurry completely (*see Note 11*).
- Mix the exchange-quenched sample with the clear supernatant of the protease solution in protease–protein ratio (w/w) of typically 1:1 for pepsin and up to ~20:1 for other proteases, such as protease type XIII (aspergillopepsin) and protease type XVIII (rhizopuspepsin). Incubate the mixture in an ice bath for 1–5 min for protease digestion. To minimize back-exchange, the digestion period must be optimized to determine the shortest period that can achieve the highest sequence coverage possible.

3.4 Liquid Chromatography–Mass Spectrometry (LC-MS) Analysis

LC-MS analysis is typically performed with a high resolution mass spectrometric instrument coupled online with a LC system (refer to Subheading 2.8). It is highly important to realize that LC separation is a major cause of back-exchange because of prolonged exposure of deuterium-incorporated proteins/peptides to flowing protic LC mobile phase solvents. Nowadays, the use of ultra-high performance liquid chromatography (UPLC or UHPLC) that

allows rapid LC separation with desired separation efficiency is the mainstream in HDX-MS because fast LC separation leads to less back-exchange. Various steps in a typical LC-MS run are as follows:

1. Trapping

The first step of a typical LC-MS injection cycle is trapping, which plays a role in focusing of samples, desalting, and removal of polar contaminants. Set the LC system to “trap mode,” in which the sample loop is connected with the reverse phase trap column and the outlet of the trap column is connected to a waste line (Fig. 3). Proteins or peptides from the sample loop are delivered to the trap column by a continuous flow of low percentage of ACN (usually less than 5% ACN). As only a weak solvent, that is, low percentage of ACN, is applied during the trapping process, the peptides/proteins tend to be retained and “trapped” in the reverse phase trap column, leading to accumulation and focusing of the samples. At the same time, salts and polar contaminants (e.g., urea) having low retention on the reverse phase trap column are washed away and leave the system through the waste line (Fig. 3). This trapping step is particularly important if high concentration of salts or denaturants (e.g., guanidine hydrochloride or urea) are present in samples. For online digestion with immobilized protease columns, the protein is first passed through the protease column for digestion and then to the trap column during the trapping process (also refer to Subheading 3.3) (Fig. 3). If the LC system is not equipped with trapping configuration, desalting and sample cleanup can be performed by running a high aqueous content of solvent for several minutes at the beginning of the LC gradient. During this desalting and cleanup period, it is important that the switching valve for controlling the flowing pathway of LC eluent should be turned to waste position to avoid contamination of the mass spectrometer.

2. LC separation

After trapping, the LC system is then switched to “injection mode,” such that the trap column becomes connected with the analytical column for chromatographic separation (Fig. 3). Proteins and peptides are typically eluted and separated with reverse phase LC columns, commonly C4 for intact proteins and C18 for peptides. LC mobile phases for separation are usually H₂O and acetonitrile (ACN) with 0.1–1% formic acid. Lower percentage (e.g., 0.1–0.2%) of formic acid is typically applied to analysis of peptides and higher percentage, that is, up to 0.5–1%, of which is usually required for intact proteins to achieve higher ionization efficiency during electrospray ionization in MS. The elution gradient typically begins with low percentage of ACN (less than 5%) and subsequently the ACN

content is increased gradually with time for elution and separation of proteins/peptides. To minimize back-exchange, the LC separation gradient should be optimized to determine the shortest separation gradient that can achieve nonoverlapping mass peaks. In general, LC separation gradients for HDX-MS are in the range of 5–15 min, depending on the molecular size of proteins (larger proteins typically generate more peptides upon protease digestion and require longer separation gradients to avoid mass peak overlapping) (*see Note 12* for the importance of retention stability).

Apart from determination of a short and sufficiently effective solvent gradient, another measure for minimizing back-exchange is that LC separation should be performed at low temperature, typically close to 0 °C. For automatic sample handling devices specialized for HDX-MS, the LC flow path including the analytical column is housed in a chilled compartment with temperature control. For conventional LC-MS without specialized HDX-MS device, the analytical column can be embedded in an ice bath to achieve the low separation temperature (*see Note 13* for the potential of overpressure in LC system).

3. MS acquisition

MS analysis can be performed in full m/z range acquisition mode with typical electrospray ionization conditions. For time-of-flight based instruments, reference mass correction with internal standard (an analyte with known mass) must be enabled to ensure high mass accuracy, as mass shift could affect calculation of deuterium uptake. If available, ion-mobility separation function, which allows separation of ions based on their size and shape, could be enabled to improve ion separation capacity [14]. However, we found that activation of ion-mobility separation function on traveling wave ion mobility-based instruments could lead to higher susceptibility to ion saturation problem.

Although, in many cases, study on peptide level is able to deliver sufficiently detailed information on protein conformation and dynamics, tandem mass spectrometric analysis (MS/MS) can be performed to further fragment the peptides into sequence-specific ions for acquiring residue level information. Yet it is important to realize that conventional fragmentation techniques (e.g., collision-induced dissociation) are known to cause significant deuterium scrambling, that is, random delocalization of deuterium on peptide chains, resulting in loss of useful residue level information [3, 15–17]. This problem could be alleviated by the use of alternative fragmentation approaches, that is, electron capture dissociation (ECD) and electron transfer dissociation (ETD) [15–17], if available.

4. Calculation of deuterium uptake of proteins/peptides

After acquisition of LC-MS spectral data, deuterium uptake is then calculated with various commercial or open source software, such as DynamX (Waters), HDX Workbench (<http://hdxworkbench.com/>), Mass Spec Studio (<https://www.msstudio.ca/>), and HX Express (<http://www.hxms.com/HXExpress/>). Although the algorithms of different software might have their particular features, calculation of deuterium incorporation level of proteins (global deuterium uptake) and peptides (local deuterium uptake) is fundamentally based on the following equation:

$$\% \text{of deuterium incorporation} = \frac{(M_t - M_o) / (M_{100\%} - M_o)}{\times 100\%}$$

where M_t is the centroid molecular mass of a protein/peptide after exchanging for a period of time t , M_o is the centroid average molecular mass of a protein/peptide without deuterium uptake and $M_{100\%}$ is the centroid average molecular mass of a protein/peptide with maximum deuterium uptake. Typically, the data of *percentage of deuterium incorporation vs exchange time* for a pair of or a couple of samples (e.g., wild type vs mutant and unbound vs ligand-bound proteins) are plotted together for comparison.

4 Notes

1. Carryover is a commonly encountered problem in LC-MS, particularly when samples with high concentration are injected. This problem is particularly significant in HDX-MS, because the low column temperature leads to slower interactions between the analyte and stationary phase of the column [18]. Carryover can be resulted from accumulation of sample residue in various parts of the LC system (e.g., pepsin column, trap column, analytical column, valves). To ensure analysis is not interfered by carryover, it is necessary to incorporate one or more blank injection(s) between each sample run and make sure that signals from proteins/peptides are not present in significant levels in the blank injection(s). If carryover problem is significant, a number of washing protocols were demonstrated to effectively wash away residual proteins/peptides [18, 19]. For example, it was shown that injecting 10% formic acid, 50% trifluoroethanol, 80% methanol, and 80% ACN in sequence into the LC system was effective to remove residual peptides [18]. However, it should be noted that washing protocol could be protein system dependent; therefore, it should be optimized for different systems.

2. The appropriate concentration of protein samples in an HDX-MS experiment is dependent on the sensitivity of the mass spectrometer utilized. Higher concentration (typically high micromolar range) of protein samples might generate mass peaks with higher intensity, yet could lead to more significant carryover problem and more effort would be needed to remove residual proteins/peptides. Therefore, it is desirable to determine the lowest protein concentration able to produce mass peaks with acceptable intensity.
3. The buffer system must be optimized such that the protein is stable at room temperature for at least the period of the exchange endpoint, which is commonly in the range of tens of minutes to hours depending on protein systems. Protein precipitation could cause blockage of valves and tubings of LC systems, which would require extensive effort for troubleshooting. It is desirable to test the protein stability over time with various means (e.g., observation for protein precipitation, activity assay, measurement of protein concentration) before carrying out an HDX-MS experiment.
4. Membrane proteins are more susceptible to protein precipitation problem. Detergents are usually required to add not only to the protein sample but also to the HDX buffer and quenching buffer in order to maintain the concentration of detergents throughout the whole HDX-MS workflow.
5. Apart from addition of detergents for micelle formation, other methods involving construction of natively like lipid bilayers, such as bicelles [20], liposomes [21, 22], and nanodisk [11, 23–25], have also been applied to maintain the stability and native structure of membrane proteins in HDX-MS experiments.
6. At the beginning of an HDX-MS experiment, optimization of protease digestion conditions for achieving the highest sequence coverage possible is usually performed with an undeuterated sample under exchange-quenched conditions. The digested sample is then analyzed by LC-tandem mass spectrometry (LC-MS/MS). Assignment of peptide peaks can be executed with peptide identification software such as *ExpASY Findpept* or other software from mass spectrometer suppliers. Acid-stable proteases are generally of relatively low digestion specificity and optimization for achieving high sequence coverage is usually done on trial-and-error basis. Among various acid-stable proteases, pepsin is the most well-characterized candidate and could be a desired starting point of optimization. However, different proteases might produce different peptide fragments; therefore the use of multiple proteases, either separately or in mixtures, could potentially

increase sequence coverage [26–28]. Besides, addition of denaturants (e.g., 1–4 M of urea or guanidine hydrochloride) and reducing agents (for disulfide bond-containing proteins; e.g., high millimolar TCEP) to the sample through the quenching buffer were also shown to improve digestion efficiency [3, 9, 12, 22, 26, 28]. For online digestion with immobilized protease columns, increasing the column pressure up to 10,000 psi, which can be achieved by connecting a flow restrictor to the waste line beyond the trap column (Fig. 3), was also demonstrated to allow for more effective digestion [13].

7. Although online digestion is more convenient, we found that this approach is not necessarily effective for all proteins. If online digestion cannot deliver desired digestion efficiency, offline digestion in solution should be explored.
8. During optimization of protease digestion conditions for achieving high sequence coverage, it should be taken into account that peptide mass peaks that can be well resolved without deuterium incorporation are not necessarily resolved after deuterium uptake, because deuterium incorporation will cause significant broadening of isotopic distribution.
9. Protein digestion with acid-stable proteases often produces peptides with overlapped sequences and a considerable number of which might only differ by one amino acid at N- or C-terminus. Peptides with overlapped sequences could act as a mean for cross-validation of HDX-MS data, that is, similar extent of change in HDX behavior is usually observed for peptides with a large proportion of overlapped sequences.
10. If the immobilized protease column is not in use, it is a good practice to disassemble the column from the LC system and store the column at 4 °C in order to maintain the lifetime of the immobilized protease. Deterioration of digestion column could be indicated by detection of abnormally long peptides after digestion.
11. For offline in-solution digestion, it is highly important to spin down the protease slurry completely and collect the clear supernatant for analysis. Injection of slurry mixtures could potentially lead to blockage of tubings and valves of the LC system.
12. Retention time stability is significantly important to generate reproducible HDX-MS results, as the extent of back-exchange is closely related to retention time, that is, more back-exchange for longer retention on column. One should be aware of the problem of retention time shift and perform troubleshooting if happened. Retention time shift could be due to many reasons, such as leaking, partial blockage of tubings and valves, hardware problems, and deterioration of columns.

13. As LC separation is executed under low temperature, that is, 0 °C, to minimize back-exchange, the potential of overpressure is much higher than typical LC experiments as viscosity of mobile phase solvents is higher at lower temperature. It should be ensured that overpressure does not occur over the entire solvent gradient.
14. It is important to realize that back-exchange begins to occur after the addition of quenching buffer; therefore, all downstream processes, that is, protease digestion and LC-MS analysis, must be performed as soon as and as quick as possible. For automatic sample handling devices for HDX-MS, exchange-quenched samples are digested and analyzed as soon as possible for minimizing back-exchange under program control. However, if HDX-MS is performed manually, exchange-quenched samples must be thawed, digested (no matter performed online or offline), and analyzed as soon as possible. Note that samples should be thawed, digested, and analyzed one by one. Avoid thawing and digesting all samples simultaneously and storing all samples in the LC autosampler for sequential LC-MS analysis, as prolonged storing exchange-quenched samples in the autosampler could lead to significant back-exchange.

Acknowledgments

This work is supported by the mass spectrometry division of University Research Facility in Life Sciences and University Research Facility in Chemical and Environmental Analysis of The Hong Kong Polytechnic University.

References

1. Wei H, Mo JJ, Tao L, Russell RJ, Tymiak AA, Chen GD, Iacob RE, Engen JR (2014) Hydrogen/deuterium exchange mass spectrometry for probing higher order structure of protein therapeutics: methodology and applications. *Drug Discov Today* 19(1):95–102. <https://doi.org/10.1016/j.drudis.2013.07.019>
2. Bou-Assaf GM, Marshall AG (2015) Biophysical mass spectrometry for biopharmaceutical process development: focus on hydrogen/deuterium exchange. In: Houde DJ, Berkowitz SA (eds) *Biophysical characterization of proteins in developing biopharmaceuticals*. Elsevier, Amsterdam, Chapter 12. <https://doi.org/10.1016/B1978-1010-1444-59573-59577.00012-59579>
3. Oganessian I, Lento C, Wilson DJ (2018) Contemporary hydrogen deuterium exchange mass spectrometry. *Methods* 144:27–42. <https://doi.org/10.1016/j.ymeth.2018.04.023>
4. Konermann L, Simmons DA (2003) Protein-folding kinetics and mechanisms studied by pulse-labeling and mass spectrometry. *Mass Spectrom Rev* 22(1):1–26. <https://doi.org/10.1002/mas.10044>
5. Cravello L, Lascoux D, Forest E (2003) Use of different proteases working in acidic conditions to improve sequence coverage and resolution in hydrogen/deuterium exchange of large proteins. *Rapid Commun Mass Sp* 17(21):2387–2393. <https://doi.org/10.1002/rcm.1207>
6. Marcoux J, Thierry E, Vives C, Signor L, Fieschi F, Forest E (2010) Investigating alternative acidic proteases for H/D exchange coupled to mass spectrometry: plasmepsin 2 but

- not Plasmepsin 4 is active under quenching conditions. *J Am Soc Mass Spectrom* 21 (1):76–79. <https://doi.org/10.1016/j.jasms.2009.09.005>
7. Tsiatsiani L, Akeroyd M, Olsthoorn M, Heck AJR (2017) Aspergillus Niger Prolyl Endoprotease for hydrogen-deuterium exchange mass spectrometry and protein structural studies. *Anal Chem* 89(15):7966–7973. <https://doi.org/10.1021/acs.analchem.7b01161>
 8. Kadek A, Mrazek H, Halada P, Rey M, Schriemer DC, Man P (2014) Aspartic protease nepenthesin-I as a tool for digestion in hydrogen/deuterium exchange mass spectrometry. *Anal Chem* 86(9):4287–4294. <https://doi.org/10.1021/ac404076j>
 9. Ahn J, Cao MJ, Yu YQ, Engen JR (2013) Accessing the reproducibility and specificity of pepsin and other aspartic proteases. *Biochim Biophys Acta* 1834(6):1222–1229. <https://doi.org/10.1016/j.bbapap.2012.10.003>
 10. Laganowsky A, Reading E, Hopper JTS, Robinson CV (2013) Mass spectrometry of intact membrane protein complexes. *Nat Protoc* 8(4):639–651. <https://doi.org/10.1038/nprot.2013.024>
 11. Li MJ, Guttman M, Atkins WM (2018) Conformational dynamics of P-glycoprotein in lipid nanodiscs and detergent micelles reveal complex motions on a wide time scale. *J Biol Chem* 293(17):6297–6307. <https://doi.org/10.1074/jbc.RA118.002190>
 12. Eisinger ML, Dorrbaum AR, Michel H, Padan E, Langer JD (2017) Ligand-induced conformational dynamics of the &ITEscherichia&IT &ITcoli &ITNa⁺/H⁺ antiporter NhaA revealed by hydrogen/deuterium exchange mass spectrometry. *Proc Natl Acad Sci U S A* 114(44):11691–11696. <https://doi.org/10.1073/pnas.1703422114>
 13. Ahn J, Jung MC, Wyndham K, Yu YQ, Engen JR (2012) Pepsin immobilized on high-strength hybrid particles for continuous flow online digestion at 10 000 psi. *Anal Chem* 84 (16):7256–7262. <https://doi.org/10.1021/ac301749h>
 14. Cryar A, Groves K, Quaglia M (2017) Online hydrogen-deuterium exchange traveling wave ion mobility mass spectrometry (HDX-IM-MS): a systematic evaluation. *J Am Soc Mass Spectrom* 28(6):1192–1202. <https://doi.org/10.1007/s13361-017-1633-z>
 15. Rand KD, Pringle SD, Morris M, Engen JR, Brown JM (2011) ETD in a traveling wave ion guide at tuned Z-spray ion source conditions allows for site-specific hydrogen/deuterium exchange measurements. *J Am Soc Mass Spectrom* 22(10):1784–1793. <https://doi.org/10.1007/s13361-011-0196-7>
 16. Landgraf RR, Chalmers MJ, Griffin PR (2012) Automated hydrogen/deuterium exchange electron transfer dissociation high resolution mass spectrometry measured at single-amide resolution. *J Am Soc Mass Spectrom* 23 (2):301–309. <https://doi.org/10.1007/s13361-011-0298-2>
 17. Masson GR, Maslen SL, Williams RL (2017) Analysis of phosphoinositide 3-kinase inhibitors by bottom-up electron-transfer dissociation hydrogen/deuterium exchange mass spectrometry. *Biochem J* 474 (11):1867–1877. <https://doi.org/10.1042/Bcj20170127>
 18. Fang J, Rand KD, Beuning PJ, Engen JR (2011) False EX1 signatures caused by sample carryover during HX MS analyses. *Int J Mass Spectrom* 302(1–3):19–25. <https://doi.org/10.1016/j.ijms.2010.06.039>
 19. Majumdar R, Manikwar P, Hickey JM, Arora J, Middaugh CR, Volkin DB, Weis DD (2012) Minimizing carry-over in an online pepsin digestion system used for the H/D exchange mass spectrometric analysis of an IgG1 monoclonal antibody. *J Am Soc Mass Spectrom* 23 (12):2140–2148. <https://doi.org/10.1007/s13361-012-0485-9>
 20. Duc NM, Du Y, Thorsen TS, Lee SY, Zhang C, Kato H, Kobilka BK, Chung KY (2015) Effective application of Bicelles for conformational analysis of G protein-coupled receptors by hydrogen/deuterium exchange mass spectrometry. *J Am Soc Mass Spectrom* 26 (5):808–817. <https://doi.org/10.1007/s13361-015-1083-4>
 21. Rigaud JL, Levy D (2003) Reconstitution of membrane proteins into liposomes. *Methods Enzymol* 372:65–86
 22. Anderson KW, Gallagher ES, Hudgens JW (2018) Automated removal of phospholipids from membrane proteins for H/D exchange mass spectrometry workflows. *Anal Chem* 90 (11):6409–6412. <https://doi.org/10.1021/acs.analchem.8b00429>
 23. Hebling CM, Morgan CR, Stafford DW, Jorgenson JW, Rand KD, Engen JR (2010) Conformational analysis of membrane proteins in phospholipid bilayer Nanodiscs by hydrogen exchange mass spectrometry. *Anal Chem* 82 (13):5415–5419. <https://doi.org/10.1021/ac100962c>
 24. Droege KD, Keithly ME, Sanders CR, Armstrong RN, Thompson MK (2017) Structural dynamics of 15-Lipoxygenase-2 via hydrogen-deuterium exchange. *Biochemistry* 56

- (38):5065–5074. <https://doi.org/10.1021/acs.biochem.7b00559>
25. Martens C, Shekhar M, Borysik AJ, Lau AM, Reading E, Tajkhorshid E, Booth PJ, Politis A (2018) Direct protein-lipid interactions shape the conformational landscape of secondary transporters. *Nat Commun* 9:4151. <https://doi.org/10.1038/s41467-018-06704-1>
26. Zhang HM, McLoughlin SM, Frausto SD, Tang HL, Emmett MR, Marshall AG (2010) Simultaneous reduction and digestion of proteins with disulfide bonds for hydrogen/deuterium exchange monitored by mass spectrometry. *Anal Chem* 82(4):1450–1454. <https://doi.org/10.1021/ac902550n>
27. Mayne L, Kan ZY, Chetty PS, Ricciuti A, Walters BT, Englander SW (2011) Many overlapping peptides for protein hydrogen exchange experiments by the fragment separation-mass spectrometry method. *J Am Soc Mass Spectrom* 22(11):1898–1905. <https://doi.org/10.1007/s13361-011-0235-4>
28. Nirudodhi SN, Sperry JB, Rouse JC, Carroll JA (2017) Application of dual protease column for HDX-MS analysis of monoclonal antibodies. *J Pharm Sci* 106(2):530–536. <https://doi.org/10.1016/j.xphs.2016.10.023>



BeStSel: From Secondary Structure Analysis to Protein Fold Prediction by Circular Dichroism Spectroscopy

András Micsonai, Éva Bulyáki, and József Kardos

Abstract

Far-UV circular dichroism (CD) spectroscopy is a classical method for the study of the secondary structure of polypeptides in solution. It has been the general view that the α -helix content can be estimated accurately from the CD spectra. However, the technique was less reliable to estimate the β -sheet contents as a consequence of the structural variety of the β -sheets, which is reflected in a large spectral diversity of the CD spectra of proteins containing this secondary structure component. By taking into account the parallel or antiparallel orientation and the twist of the β -sheets, the Beta Structure Selection (BeStSel) method provides an improved β -structure determination and its performance is more accurate for any of the secondary structure types compared to previous CD spectrum analysis algorithms. Moreover, BeStSel provides extra information on the orientation and twist of the β -sheets which is sufficient for the prediction of the protein fold.

The advantage of CD spectroscopy is that it is a fast and inexpensive technique with easy data processing which can be used in a wide protein concentration range and under various buffer conditions. It is especially useful when the atomic resolution structure is not available, such as the case of protein aggregates, membrane proteins or natively disordered chains, for studying conformational transitions, testing the effect of the environmental conditions on the protein structure, for verifying the correct fold of recombinant proteins in every scientific fields working on proteins from basic protein science to biotechnology and pharmaceutical industry. Here, we provide a brief step-by-step guide to record the CD spectra of proteins and their analysis with the BeStSel method.

Key words Circular dichroism, Protein secondary structure, Protein fold, Amyloid, β -sheet

1 Introduction

Circular dichroism (CD) corresponds to the differential absorption between left and right circularly polarized light (Fig. 1). In the far-UV region between 170 and 250 nm, mostly the electronic transitions of the peptide bonds contribute to the CD spectrum of proteins [1, 2]. Depending on the local geometry, environment, and H-bond pattern of the peptide bonds, the polypeptide chains with different conformations can exhibit distinct, characteristic spectral profiles, which is manifested in the CD spectra of proteins

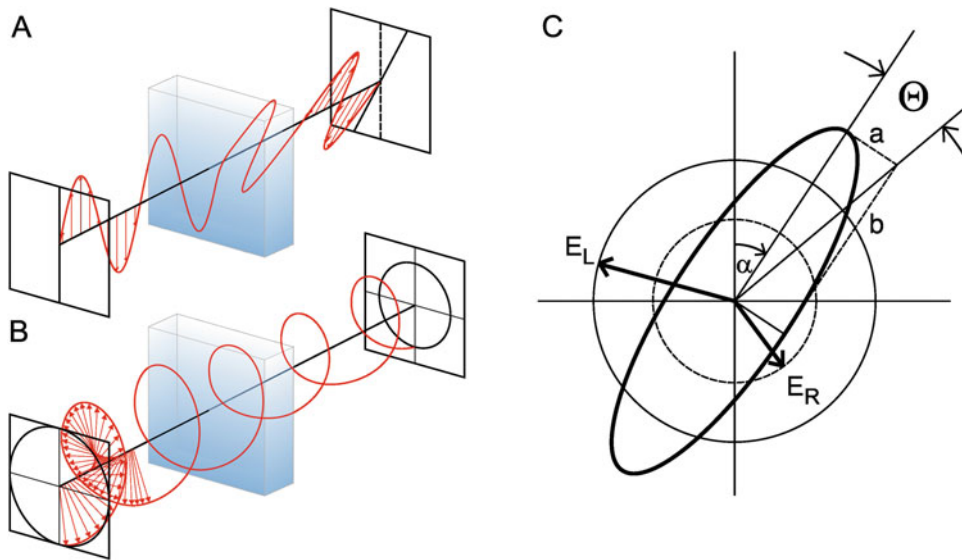


Fig. 1 The phenomenon of circular dichroism. Light is an electromagnetic wave which can be characterized by the electric and magnetic fields that are perpendicular to each other and the direction of the travel of the light. Linearly polarized light is characterized by the electric field vector oscillating in one plane (a), while the electric field vector of circularly polarized light is rotating around the axis of propagation by maintaining a constant amplitude (b). Looking into the light propagating toward the observer, electric field vector rotating counter-clockwise or clockwise depict the left and right circularly polarized lights, respectively. The summation of left and right circularly polarized light of equal amplitudes results in linearly polarized light while different amplitudes result elliptically polarized light (c). Optical active material (which should have chiral properties) interacts with light in a polarization dependent manner which can be manifested in optical rotation of the plane of polarization (a, and angle α in c) and in circular dichroism which is the differential absorption of the left and right circularly polarized light (b, c). For details of the theory of circular dichroism see [1]. At the practical level, the differential absorption of the left and right circularly polarized light can be expressed as the difference in the extinction coefficients, $\Delta\varepsilon = \varepsilon_L - \varepsilon_R$, or as the ellipticity of the summation of the left and right circularly polarized lights of different amplitudes, $\tan\theta = a/b = (E_R - E_L)/(E_R + E_L)$, where E_R and E_L are the amplitudes of the electric field vectors. θ will be negative if E_R is smaller than E_L . Measured ellipticity is usually given as θ in the unit of mdeg. When $\Delta\varepsilon$ is in $M^{-1}\cdot\text{cm}^{-1}$ units and θ is also normalized to the molar number of residues (more precisely, to the number of peptide bonds) and pathlength in cm, denoted as $[\theta]$ and given in the traditional unit of $\text{deg}\cdot\text{cm}^2\cdot\text{dmol}^{-1}$, the value of $\Delta\varepsilon$ is equal to $[\theta]/3298$ (we have to note, that for the correct equation, the factor of 3298 is not dimension-less)

of different structural classes (Fig. 2). This observation initiated the development of algorithms for the secondary structure estimation from the CD spectra. In the last 30 years, a dozen CD spectrum analysis algorithms made attempts to accurately estimate the secondary structure composition of the proteins. These methods use reference CD spectra of proteins with known structure to make an estimation of different types of secondary structure elements (most often helix, β -sheet, turn, and disordered). The mathematical background and performances of these methods are reviewed and compared [3, 4]. Generally, they predict the helix content more or less

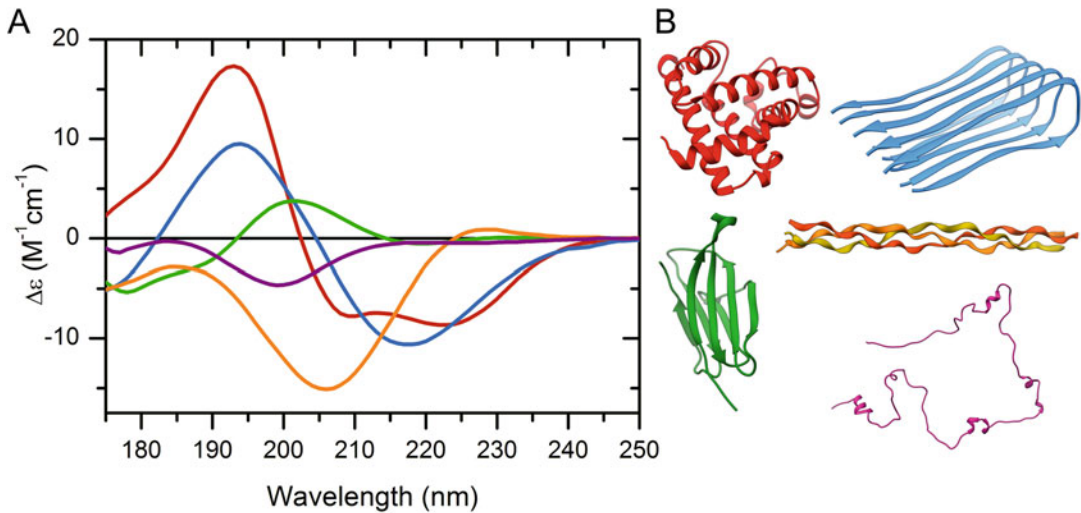


Fig. 2 Characteristic far-UV CD spectra of different protein architectures. Proteins of distinct secondary structures such as α -helix (red), parallel β -sheet (blue), antiparallel β -sheet (green), polyproline-helix (orange), and disordered chain (purple) exhibit characteristic spectral shapes indicating that CD spectroscopy can be useful for the determination of the secondary structure of proteins

accurately, while often fail to properly predict the β -sheet content due to the large spectral diversity of β -structured proteins (Fig. 3). In the background of this spectral diversity, there must be the variety of β -sheets in the orientation (parallel–antiparallel), the length and number of strands, and their twists, which made difficult to estimate this component from the CD spectrum and was believed to be an intrinsic limitation of the technique [5].

Recently, we have shown that the spectral contribution of β -sheets depends on the parallel–antiparallel orientation and the twist of the β -sheets [4]. Based on this observation, we have developed a new method named BeStSel (Beta Structure Selection) for the secondary structure estimation of proteins from the CD spectra that takes into account the orientation and twist of the β -sheets. The method defines eight structural components: regular and distorted α -helices, left-handed, relaxed (slightly right-hand twisted) and right-hand twisted antiparallel β -sheets, parallel β -sheet, turn, and “others” (Table 1, and for detailed definitions *see* Micsonai et al. [4]).

BeStSel provides an improved accuracy on a broad range of protein structures including β -sheet-rich proteins, membrane proteins, protein aggregates, and amyloid fibrils.

As a result of the detailed structural information gained from the CD spectrum, BeStSel is capable of predicting the protein fold down to the homology level using the CATH fold classification (Fig. 4) [9, 10].

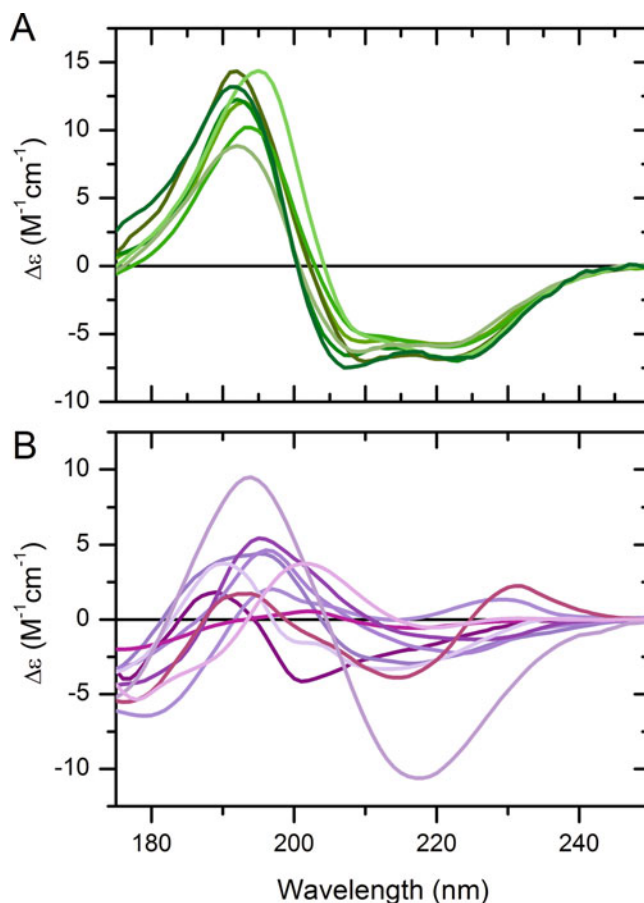


Fig. 3 The spectral diversity of β -structures. (a) α -helical proteins have uniform spectral shape as shown as demonstrated here by proteins having $\sim 50\%$ α -helix content. (b) Despite their similar ($\sim 50\%$) β -sheet content, β -structured proteins show a large spectral diversity making secondary structure estimation a difficult challenge

A web server was constructed at <http://bestsel.elte.hu> making the BeStSel method freely accessible for the scientific community.

In the Materials section completed with extended Notes we briefly describe the essential sample preparation steps for a reliable CD measurement that are necessary for an accurate secondary structure estimation. In the Methods section we give a step-by-step guide for the modules of the BeStSel webserver to analyze protein CD spectra.

2 Materials

A lot of buffer compounds and salts have high absorption in the far-UV region. Their use should be avoided or their concentration

Table 1
Structural components of BeStSel and their relation to the DSSP components [6]

Structural component	Description of the component	Related DSSP component
Helix1 ^a	Regular α -helix (middle part of α -helices)	H
Helix2 ^a	Distorted α -helix (2–2 residue at the end of α -helices)	
Anti1	Left-handed antiparallel β -sheet	E
Anti2	Relaxed (slightly right-hand twisted) antiparallel β -sheet	
Anti3	Right-hand twisted antiparallel β -sheet	
Parallel	Parallel β -sheet	
Turn	Turn, as defined by DSSP	T
Others	3_{10} -helix, π -helix, β -bridge, bend, loop/irregular and invisible regions of the structure	G,I,S,B,O

^aIt is important to note that most of the other algorithms such as SELCON [7], CONTIN, and CDSSTR [8] define mixed *Helix* components, instead of pure α -helix, as the sum of α - and 3_{10} -helices. This should be considered when comparing results of different methods

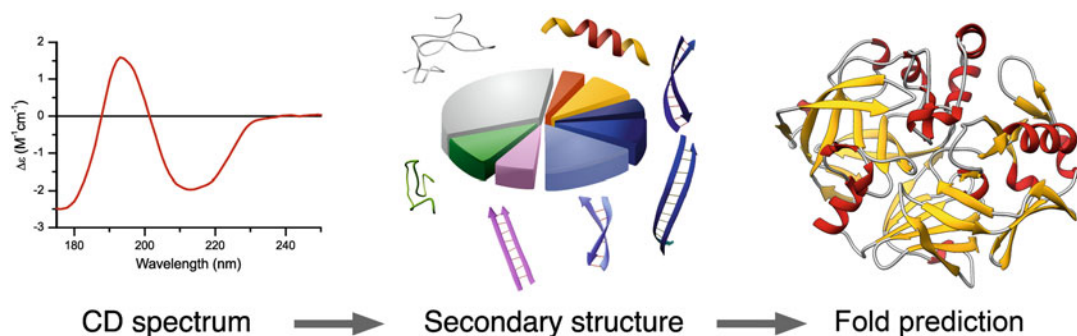


Fig. 4 The BeStSel method. Schematic representation of the secondary structure components of BeStSel (*see also* Table 1) and the pipeline of structure estimation. Obtaining the fractions of the eight components from the CD spectrum by BeStSel, the protein fold can be predicted

should be kept at the minimum that is acceptable for the protein. Phosphate buffer (not PBS) is suitable for CD spectroscopy with as low salt added as possible. However, it might be incompatible with other buffer components to be used, for example with calcium, or with the protein. High absorption of the buffer limits the usable wavelength range and can be avoided by choosing a shorter path-length cell which requires increased protein concentration (*see* Sub-heading 3.1 and Notes 1–3).

Depending on the instrument and the cell holder used, cylindrical or rectangular quartz cells can be used in the >180 nm wavelength region. Below 180 nm, or in the case of low sample volume, demountable calcium fluoride cells can be used.

3 Methods

3.1 Sample Preparation

1. The CD spectrum shows the average spectrum of the components having CD signal in the sample. It is important to have a pure, homogenous protein sample free of contaminations of other proteins or other chiral biomolecules such as nucleic acids. Check the purity of the sample by SDS-PAGE, mass spectrometry, absorption spectroscopy (for nucleic acid contamination), and other complementary methods. Take into consideration that the CD spectrum is also affected by expression tags often used in the case of recombinant proteins (*see Note 1*).
2. The inhomogeneity and light scattering also affect the CD signal causing shrinking of the amplitude and distorting the spectrum, which may have caused by protein aggregation and precipitation (*see Note 2*).
3. Transfer the sample into a buffer suitable for CD measurements. The best method for this is dialysis where the dialysis buffer can be used for baseline measurement. A lyophilized protein powder often contains contaminations, so it is advised not only to dissolve it in the proper buffer but dialyze it. An alternative method can be a transfer of the protein to the buffer of the measurement by using a filtration spinning tube or desalting column.
4. Determination of the accurate protein concentration is crucial for the correct normalization and quantitative analysis of the CD spectra. Select the pathlength of the cuvette depending on the concentration in a way that the product of the pathlength in mm and the concentration in mg/ml should be ~ 0.1 (it means that for a solution of 0.1 mg/ml concentration, a 1 mm cell is optimal for use). Selecting the appropriate buffer and pathlength, CD spectroscopy is capable of studying the protein structure in a wide concentration range of 0.05–20 mg/ml, which is a significant advantage over the other techniques used for protein structure determination, such as NMR, infrared spectroscopy, vibrational CD or RAMAN spectroscopy (*see Note 3* for concentration determination).
5. The instrumentation of CD spectroscopy is well-developed, the users routinely can measure the spectra in the 190–260 nm range with some considerations on the buffer and salt compositions of the sample. Choose shorter pathlengths (10–50 μm) and high protein concentrations (2–10 mg/ml) to record the spectra down to 180 nm on conventional instruments. Synchrotron radiation CD (SRCD) stations can collect spectra at even shorter wavelengths [11]. To collect high quality CD spectra suitable for quantitative

structural analysis, the instrumental parameters should be carefully chosen following the instrument manual. **Note 4** discusses the preferable measurement parameters. For quantitative measurements, calibrate the instrument occasionally for amplitude and wavelength accuracy (*see Note 5*).

3.2 Wavelength Range, Baseline Subtraction, and Data Normalization

1. CD spectroscopy is a type of absorption spectroscopy and the CD signal is measured above the overall absorption of the sample which should be kept at low for the good signal-to-noise ratio and linearity of the detector. The voltage (high tension, HT) of the detector is adjusted to this overall absorption and should not exceed a limit (e.g., ~600 V limit in the case of a detector having 900–1000 V maximum HT). Discard the data measured at HT values over this limit.
2. Correct the sample spectrum by subtracting the baseline measurement of the same buffer that is used for the protein. A moderate smoothing can be applied on the spectrum by taking care not to change significantly any sharp component or steep part of the spectrum.
3. Normalize the CD spectrum for the concentration, pathlength and number of peptide bonds. The mean residue ellipticity ($\text{deg}\cdot\text{cm}^2\cdot\text{dmol}^{-1}$) is defined as follows:

$$[\theta]_{MRE} = \theta / (10 \cdot c_r \cdot l)$$

where θ , the measured ellipticity, is in mdeg, c_r is the molar concentration per residue, and l is the pathlength in cm. The also commonly used extinction coefficient difference, $\Delta\epsilon = [\theta]/3298.2$, its unit is $\text{M}^{-1}\cdot\text{cm}^{-1}$. Although BeStSel can handle the baseline subtracted raw data, it is important to understand the normalization procedure because the output of BeStSel and the proper form of CD spectra for publication is the normalized data.

3.3 Single Spectrum Analysis

At the starting page of the BeStSel webserver, by default, data can be uploaded for single spectrum analysis in the form of a text file or can be copied into the window in two data columns, separator can be space, tab, comma or semicolon. Upload the data either as normalized in $\Delta\epsilon$ or $[\theta]_{MRE}$, or as measured, baseline subtracted data. In the latter case, you have to provide the concentration (μM), pathlength (cm), and the number of residues. The page is protected by a captcha against malicious use. In all cases, the program normalizes or converts the uploaded data to $\Delta\epsilon$, which can be verified in the next, *Data Examination* page. Note, that the numeric format uses dot as decimal point. If the spectrum in the *Data examination* page contains steps, probably the decimal sign is incorrect. Starting the calculation, the results will appear in a

graphical image with all the useful information provided: wavelength range, the estimated secondary structure content, the curve and error of the fitted spectrum, and user provided information. At first, data is analyzed in the possible widest wavelength range of the uploaded data. However, we strongly suggest to choose an appropriate wavelength range where the PMT voltage was below a limit (e.g., 600 volts) determined by the manufacturer upon the measurement (*see* Subheading 3.2). *See* **Notes 2** and **4** for buffer selection and experimental setup. Below the results, change the output format for your convenience. Results can be saved as a graphical image. For further data processing by the users, result can be shown in text format with the predicted secondary structure contents at the top and the experimental, fitted, and the residual data in columns below. Transfer the data by copying it to any data processing software to make your own plots, etc.

On the left side of the *Results* page, the wavelength range can be chosen and the analysis can be recalculated. Different wavelength ranges will provide slightly different results; however, in the case of using correct concentrations and normalization, the difference is within the estimation error. A scale factor can be chosen for recalculation, as well. The CD amplitude is multiplied with this factor. The “*Best factor*” function carries out a series of analysis by changing the current scaling factor automatically in the range of 0.5–2. The dependence of the individual secondary structure components on the CD amplitude is plotted. This can be informative in the case of uncertainties in the protein concentration or pathlength. In case of CD data in a wide wavelength range (down to at least 180 nm), the alteration of the factor with the lowest fitting NRMSD from 1 is a good indicator of incorrect concentration or pathlength values.

3.4 Fold Recognition

The eight secondary structure components of BeStSel bear sufficient information that is characteristic to the protein fold and makes possible its prediction. At first, twenty closest structures based on Euclidean distance are searched on the entire PDB. In case of single domain proteins, a fold prediction using the CATH protein fold classification [10, 12] can be done. The single domain PDB subset is a nonredundant collection of chains containing single CATH domains or homodomains filtered for $\leq 95\%$ sequence homology and resolution better than 3.0 Angströms. This dataset contains 55,350 single domains covering 4 classes, 41 architectures, and 1310 topologies and 5398 homologies [9]. The fold can be predicted by searching for the closest structures based on the Euclidean distance in the eight components. While this method does not take into account the possible error of the secondary structure estimation from CD, it can be used even if the secondary structural space is rarely populated by structures around the estimated result. Another method is surveying all the structures within the expected

error of the CD results and sort them by their fold and the frequency of that fold [4]. At the level of architecture and topology, the ten most populated groups are presented. The most sophisticated way of fold prediction is a weighted K-nearest neighbors search using the chain length as extra parameter. Fold prediction can be initiated from within the *Single Spectrum Analysis* after getting the secondary structure contents or from a separate block at the starting page by manually providing the Secondary structure contents and chain length [9].

Use the *Fold recognition* module to find structures in the PDB and fold domains in CATH that are similar to the experimentally investigated protein. This function can be especially useful to verify the correct fold of recombinant proteins or search for the fold of proteins having low sequence homology to the proteins in the PDB.

3.5 Multiple Spectra Analysis

In this module, upload a series of spectra in a text file or copy into the window from a worksheet to analyze the CD spectra as a function of temperature, ligand concentration, etc. In the uploaded data, the first row should contain the values of the variable as the function of which the spectra were recorded. Below, there are columns. The first column contains the wavelength values and the others columns contain the corresponding spectral data. Therefore, the total number of columns should be equal to the number of values in the first row plus one. Data separator can be either tab, comma, semicolon, or space. The units of the input data can be chosen similarly to *Single Spectrum Analysis*. After the checkup of the uploaded data as a series of spectra in $\Delta\epsilon$, starting the calculation, the estimated secondary structure contents will be shown on the *Result* page as the function of the given parameter (temperature, ligand concentration, etc.). The wavelength range can be changed or the results can be recalculated with using a scaling factor applied for all the spectra. The results can be saved as image or copied out as data text. We have to note that *Multiple Spectra Analysis* is developed for analysis of a series of related CD spectra with the same number of data points and wavelength ranges. Unrelated spectra should be evaluated separately in *Single Spectrum Analysis*.

3.6 Secondary Structure Composition from PDB Structures

In this module of BeStSel, provide the four letters codes of atomic resolution structures deposited in the PDB to list out their secondary structure contents. Besides the eight secondary structure components of BeStSel, the six components of SELCON/CONTIN/CDSSTR methods [8] and the eight components of DSSP [6] are also shown for the entire molecule or selected subunits. Upon selecting the chain, the protein fold classification is also provided using the CATH classification [10]. This module of the BeStSel server is useful to compare the secondary structure results to the available reference protein structures.

3.7 Limitations of the BeStSel Method

The eight secondary structure components of BeStSel do not account for some special secondary structure types. Polyproline-II helix, different type of turns, 3_{10} -helices are not distinguished by BeStSel and thus analysis for such structures is not adequate. BeStSel does not handle the aromatic contributions (other algorithms neither do) which gives some uncertainty when the number of aromatic residues is high in the protein. The spectra of highly disordered proteins somewhat remind the highly right-twisted antiparallel β -sheets (Anti3 component), and partly might be counted as Anti3 instead of “Others” [9].

4 Notes

1. Sample purity and preparation

The CD spectrum shows the average spectrum of the components having CD signal in the sample. Thus, it is important to have a pure, homogenous protein sample free of contaminations of other proteins or other chiral biomolecules such as nucleic acids. The purity of the sample should be checked by SDS-PAGE, mass spectrometry, absorption spectroscopy (for nucleic acid contamination) and other complementary methods. Recombinant proteins are often expressed using fused protein tags that provide higher expression or used for efficient purification (N-terminal extension of Met or more residues, His-, GST-, or other tags on either terminal) or stabilize the protein structure. These extensions or tags can affect the structure and stability of the proteins and contribute to the CD spectrum, as well. It is advised to have them removed from the protein. When removal of these extensions is not possible, it is important to take them into account in the analysis of the CD spectrum (number of residues, molecular weight, and presumed contribution to the estimated secondary structure contents).

CD spectroscopy is sensitive for light scattering effects which may have caused by protein aggregation and precipitation. To remove any precipitates, the sample should be spun down at least in a table top centrifuge at $>10,000 \times g$ force. To remove small oligomers of a protein, ultracentrifuge around $\sim 100,000 \times g$ could be used. In all cases the protein concentration should be determined after centrifugation.

In the case of measuring protein aggregates and amyloid fibrils, no centrifugation is applied or only a short centrifugation at low force can be used to remove the large aggregates which cause inhomogeneity and light scattering of the sample. Amyloid samples should be well homogenized by thorough pipetting or even using a slight ultrasonication.

2. Buffer selection

A lots of buffer compounds and salts have high absorption in the far-UV region. Their use should be avoided or their concentration should be kept at the minimum that is acceptable for the protein. Using shorter pathlengths (that needs higher protein concentrations) can decrease the buffer absorption. Table 2 shows the usable wavelength range for CD of the

Table 2
Absorption of different buffer compounds and salts in the far-UV^a

Compound	No absorption above	210 nm	200 nm	190 nm	180 nm
NaClO ₄	170 nm	0	0	0	0
NaF	170 nm	0	0	0	0
Boric acid	180 nm	0	0	0	0
NaCl	205 nm	0	0.02	>0.5	>0.5
Na ₂ HPO ₄	210 nm	0	0.05	0.3	>0.5
NaH ₂ PO ₄	195 nm	0	0	0.01	0.15
Na-acetate	220 nm	0.03	0.17	>0.5	>0.5
Glycine	220 nm	0.03	0.1	>0.5	>0.5
Diethylamine	240 nm	0.4	>0.5	>0.5	>0.5
NaOH	230 nm	>0.5	>2	>2	>2
Boric acid, NaOH	200 nm	0	0	0.09	0.3
Tricine	230 nm	0.22	0.44	>0.5	>0.5
TRIS	220 nm	0.02	0.13	0.24	>0.5
HEPES	230 nm	0.37	0.5	>0.5	>0.5
PIPES	230 nm	0.2	0.49	0.29	>0.5
MOPS	230 nm	0.1	0.34	0.28	>0.5
MES	230 nm	0.07	0.29	0.29	>0.5
Cacodylate	210 nm	0.01	0.01	0.22	>0.5
Citric acid ^b	240 nm	0.21	0.22	0.45	>2.5
Dithiothreitol ^b	255 nm	1.28	>3	>3	
Mercaptoethanol ^b	254 nm	0.71	2.35	2.02	
TCEP ^b	235 nm	0.24	0.64	2.78	
DMSO (0.1%) ^b	233 nm	1.8	>3	>3	
DMF (0.1%) ^b	243 nm	3.82	>3	>3	
GdnHCl (1 M) ^b	218 nm	0.36	>3	>3	
Urea (1 M) ^b	227 nm	0.29	>3	>3	

^aIf not specified differently, data is given for 10 mM solutions at 1 mm pathlength. Adapted from [13]

^bOwn measurement

different buffer compounds and salts. Denaturants such as GdnHCl and urea which are usually used at high concentrations have especially high absorptions which often make impossible the quantitative analysis of the CD spectrum in the lack of sufficient usable wavelength range. Instead of them dodine could be used [14], which denatures the protein at orders of magnitude lower concentrations. Sodium and reducing agents such as dithiothreitol or mercaptoethanol also have high absorption. These compounds should be dialyzed out from the sample prior to the measurement. Tris(2-carboxyethyl) phosphine (TCEP) is better as reducing agent for CD because of its lower effective concentration range and somewhat lower extinction coefficient. Short peptides or other organic chemicals are often dissolved in dimethyl sulfoxide (DMSO) which is noncompatible with CD spectroscopy even after ten thousand-fold dilution.

3. Concentration determination

An advantage of CD spectroscopy is the usable wide protein concentration range which starts at least an order of magnitude lower concentration than the minima for NMR, infrared, RAMAN and other spectroscopies used for the study of protein secondary structure. It can be as low as 0.05 mg/ml in a 2 mm cell and as high as 20 mg/ml in a 5 μ m cell. Thus, it is a complementary method for the other spectroscopy techniques to check whether at high concentration the protein still exhibits the same conformation as it does at low, more physiological concentrations. A lot of proteins aggregate at higher protein concentrations undermining the results of other, often expensive and time consuming methods. Using CD spectroscopy, the conformational state of the protein as a function of the concentration, pH and other parameters can be easily verified. At short pathlengths, CaF₂ cells are often used instead of quartz cells. Using very short pathlengths of few micrometers may result orientation of long molecules such as amyloid fibrils in the cell which should be taken into consideration.

The method considered to be the most accurate for concentration determination is quantitative amino acid analysis. In case the protein contains tryptophan and tyrosine residues, the concentration can be determined by measuring the absorbance at 280 nm. The extinction coefficient at 280 nm can be calculated from the primary sequence using the *ProtParam tool* (<https://web.expasy.org/protparam/>) [15]. In the absence of these amino acids, the concentration can be determined by the absorbance at 205 nm [16] or 214 nm [17]. An advantage of measuring at these two wavelengths is that, because of the high extinction coefficients, the CD samples can be directly measured. If the spectropolarimeter is capable of accurately

converting the HT values to absorbances, then the concentrations can be determined right from the CD measurements after subtracting the baseline absorptions. Extinction coefficients at 205 and 214 nm can be calculated from the amino acid sequence at the BeStSel homepage (<http://bestsel.elte.hu>).

4. Instrument settings

Although the CD spectra of the protein do not contain sharp peaks, the bandwidth should not be set to more than 2 nm, preferably, it is 1 nm. In case of continuous scanning mode when the wavelength is continuously changed at a scanning rate, the response/data integration time and the scanning rate should be harmonized in a way that during averaging of one data point, the wavelength should not be shifted more than the value of the bandwidth. It means that at a rate of 100 nm/min 0.5 or at most 1 s integration time should be used and these values are 1–2 s for 50 nm/min, 2–4 s for 20 nm/min and 4–8 s for 10 nm/min scanning rates. Depending on the amplitude and noise, several scans should be accumulated (averaged) at the convenience of the user. Usually a spectrum recording for 15 min overall time (~10 scans averaged at 50 nm/min scanning rate) is sufficient for an acceptable quality. To double the signal-to-noise ratio, four times more scans are needed. The baseline spectrum of the buffer should be collected with using the same parameters.

To collect as much information as possible, the CD spectra should be recorded in the widest usable wavelength range limited by the sample absorption at the low end, down to at least 200 nm but favorably to 190 or 180 nm. SRCD instruments can provide the CD spectra down to 175 nm. The recommended starting wavelength is 260 nm. In the 260–250 nm region (after baseline subtraction), a flat signal, close to zero, is an indication of a good baseline subtraction and the lack of light scattering effects and nucleic acid or other contaminations. Normally, the baseline CD spectrum of the buffer solution is recorded first and the usable wavelength range is estimated from the HT values which should not exceed the 50–60% of the maximum value. It is better to collect a fast protein sample spectrum first to determine the usable wavelength range and then carry out the high quality measurement only in the appropriate wavelength range to save time.

5. Instrument calibration

Conventional benchtop instruments are usually calibrated by the manufacturer and the calibration can be repeated occasionally following the instruction manual. In the case of SRCD beamlines, the spectra can be corrected by a reference measurement of 1S-(+)-10-camphorsulfonic acid (CSA) which provides a negative and a positive peak at 192.5 and 290.5 nm having $\Delta\epsilon$

values -4.72 and $2.36 \text{ M}^{-1}\cdot\text{cm}^{-1}$, respectively [18]. The concentration of the CSA can be determined at 280 nm using an extinction coefficient of $34.58 \pm 0.18 \text{ M}^{-1}\cdot\text{cm}^{-1}$ [19].

Acknowledgments

This work was supported by the National Research, Development and Innovation Fund of Hungary [K120391, KH125597, 2017-1.2.1-NKP-2017-00002, FIEK16-1-2016-0005, TÉT16-1-2016-0134, TÉT16-1-2016-0197, 2019-2.1.11-TÉT-2019-00079]; SOLEIL Synchrotron, France [proposals 20191810, 20181890, 20181896, 20180805, 20171582]; and CampusFrance [Balaton-Programme Hubert Curien, 38642YK]. A.M. is supported by the Bolyai János Scholarship of the Hungarian Academy of Sciences, and the New National Excellence Program (ÚNKP-18-4-ELTE-833, ÚNKP-19-4-ELTE-790).

References

1. Fasman GD (ed) (1996) Circular dichroism and the conformational analysis of biomolecules. Plenum Press, New York
2. Berova N, Nakanishi K, Woody RW (eds) (2000) Circular Dichroism: principles and applications, 2nd edn. Wiley, New York
3. Greenfield NJ (2006) Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* 1:2876–2890
4. Micsonai A, Wien F, Kernya L, Lee YH, Goto Y, Refregiers M, Kardos J (2015) Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc Natl Acad Sci U S A* 112: E3095–E3103
5. Khrapunov S (2009) Circular dichroism spectroscopy has intrinsic limitations for protein secondary structure analysis. *Anal Biochem* 389:174–176
6. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
7. Sreerama N, Venyaminov SY, Woody RW (1999) Estimation of the number of alpha-helical and beta-strand segments in proteins using circular dichroism spectroscopy. *Protein Sci* 8:370–380
8. Sreerama N, Woody RW (2000) Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal Biochem* 287:252–260
9. Micsonai A, Wien F, Bulyaki E, Kun J, Moussong E, Lee YH, Goto Y, Refregiers M, Kardos J (2018) BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res* 46:W315–W322
10. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH—a hierarchic classification of protein domain structures. *Structure* 5:1093–1108
11. Miles AJ, Wallace BA (2006) Synchrotron radiation circular dichroism spectroscopy of proteins and applications in structural and functional genomics. *Chem Soc Rev* 35:39–51
12. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG, Lehtinen S, Studer RA, Thornton J, Orengo CA (2015) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res* 43:D376–D381
13. Creighton TE (ed) (1989) Spectral methods of characterizing protein conformation and conformational changes in protein structure: a practical approach. Oxford University Press, Oxford
14. Guin D, Sye K, Dave K, Gruebele M (2016) Dodine as a transparent protein denaturant for circular dichroism and infrared studies. *Protein Sci* 25:1061–1068

15. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF (1999) Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol* 112:531–552
16. Anthis NJ, Clore GM (2013) Sequence-specific determination of protein and peptide concentrations by absorbance at 205 nm. *Protein Sci* 22:851–858
17. Kuipers BJ, Gruppen H (2007) Prediction of molar extinction coefficients of proteins and peptides using UV absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography-mass spectrometry analysis. *J Agric Food Chem* 55:5445–5451
18. Chen GC, Yang JT (1977) 2-point calibration of circular dichrometer with D-10-Camphor-sulfonic acid. *Anal Lett* 10:1195–1207
19. Miles AJ, Wien F, Wallace BA (2004) Redetermination of the extinction coefficient of camphor-10-sulfonic acid, a calibration standard for circular dichroism spectroscopy. *Anal Biochem* 335:338–339

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Navigating the Global Protein–Protein Interaction Landscape Using iRefWeb

Andrei L. Turinsky, Sam Dupont, Alexander Botzki, Sabry Razick, Brian Turner, Ian M. Donaldson, and Shoshana J. Wodak

Abstract

iRefWeb is a resource that provides web interface to a large collection of protein–protein interactions aggregated from major primary databases. The underlying data-consolidation process, called iRefIndex, implements a rigorous methodology of identifying redundant protein sequences and integrating disparate data records that reference the same peptide sequences, despite many potential differences in data identifiers across various source databases. iRefWeb offers a unified user interface to all interaction records and associated information collected by iRefIndex, in addition to a number of data filters and visual features that present the supporting evidence. Users of iRefWeb can explore the consolidated landscape of protein–protein interactions, establish the provenance and reliability of each data record, and compare annotations performed by different data curator teams. The iRefWeb portal is freely available at <http://wodaklab.org/iRefWeb>.

Key words Protein–protein interactions, Interaction networks, Proteomics, Literature curation, IMEx consortium, PSI-MI standards, Bioinformatics resources, iRefWeb, iRefIndex

1 Introduction

Protein–protein interactions (PPI) play a major role in biochemical processes across most of the cell types, as well as enabling intercellular signaling and other essential molecular activities in an organism [1–3]. These interactions often occur among groups of proteins, called protein complexes. The dynamic nature of such associations, the order in which different proteins interact with each other, the three-dimensional structure of proteins complexes—these are just several aspects that should be explored to gain a better understanding of the biological processes in the cell. Research in the area of protein interactions has been growing steadily, targeting a range of organisms, from bacteria and eukaryotes, to mammals including mouse, rat, and human [4–9]. These research efforts are enabled in large part by the advancements in the

interaction detection technologies, both on a small-scale and using high-throughput methods (*see* refs. 10, 11 for review).

In parallel with the primary studies aimed at the detection of protein interactions and networks, there has also been an impressive growth of data curation efforts, in which teams of experts examine the published articles and extract information on the protein interactions described therein. The goal of such efforts is to aggregate knowledge on PPIs from disparate studies into a unified collection of PPIs for each organism, and to enable access to the accumulated protein interactome for the research community worldwide. There is a number of PPI curation efforts and teams, some targeting general PPI networks while others focusing on specific organisms (e.g., mammalian), biological processes (e.g., extracellular matrix), or interaction types (e.g., protein complexes). This diversity has led to the proliferation of PPI databases, many of which overlap not only in terms of their general focus area but also their specific content [12].

In the early years of PPI curation efforts, different teams often curated the literature according to their internal protocols. As a result, PPI records generated by different teams curating the same publication often differed substantially. In a previous study, we reported that whenever two different databases curated the same PubMed article, on average, they fully agreed on just 42% of the reported PPIs [13]. The agreement on the identities of the proteins involved in these interactions was higher, about 62%. One of the common factors leading to differences is the representation of protein complexes: some databases describe a protein complex as a group of proteins while others break it into pairwise interactions. Another important factor has been the identity of the organism involved: for example, whenever the primary studies describe a mixture of mammalian protein constructs ambiguously (e.g., using mouse orthologs in place of human proteins in their experimental design) the curators have difficulties resolving these ambiguities, leading to different organisms being reported in the annotated records. Similarly, agreement was lacking on the description of the PPI itself, such as the exact terminology used to represent interaction types or detection methods used in the original publication.

Eventually the PPI community developed a set of PPI representation standards and common curation protocols, which have been led by the IMEx consortium [14–17] with key data standards made available through the Proteomics Standards Initiative (PSI). These efforts made significant contribution to the research fields by promoting two types of data formats for PPI representation: one is a tab-delimited textual format to represent molecular interactions (MI), called PSI-MITAB format, which presents a collection of PPIs as a human-readable table; the other is a PSI-MI XML format, which is more flexible but typically requires XML processing tools

and software to operate on [18]. In addition, the Molecular Interaction (MI) ontology is widely used as a controlled vocabulary for describing the type of interactions, the interaction detection methods used in the primary study, and various other annotation items related to the PPI [18]. E.g., the ontology term MI:0915 defines the interaction type “physical association” as follows: “*Interaction between molecules within the same physical complex. Often identified under conditions which suggest that the molecules are in close proximity but not necessarily in direct contact with each other.*” Such term would be found in the 12th column of the MITAB data format. The seventh MITAB column would contain interaction detection methods, such as MI:0676 “tandem affinity purification” as defined in the MI ontology. While different curators may not always agree on the supporting evidence found in the primary publication, the IMEx consortium provides guidelines for a common usage of the ontology terms and other curation practices.

An important later development was the introduction of the PSICQUIC web services [19], which maintain a repository of PPI data providers and allow their respective PPI databases to be searched in real time. There are a number of advantages to this automated approach: new data providers can join the PSICQUIC registry; the users are able to see which of the data repositories are currently online using PSICQUIC View from EBI (<http://www.ebi.ac.uk/Tools/webservices/psicquic/view/>); and the data records can be retrieved jointly from multiple databases using the same syntax, defined via the Molecular Interaction Query Language (<https://psicquic.github.io/MiqlDefinition.html>). The PSICQUIC View also provides a “clustering” tool, which attempts to resolve redundancies in the PPI records retrieved from multiple databases. However, this automatic framework has substantial drawbacks related to data aggregation: many of the redundancies between databases remain unresolved, especially whenever different source databases represent their data differently. For example, BioGRID [20] uses gene IDs as its main identifier system for interactors, whereas databases such as IntAct [21] and many others use protein IDs. This and other discrepancies create a substantial burden for the user, who needs to decide how to post-process the retrieved data records further, e.g., by using external ID mapping systems and/or writing custom software scripts. The required consolidation of PPI data beyond a simple matching of protein IDs is far from trivial. In fact, the user might not even be fully aware of the full extent of redundancies in the automatically merged dataset until the data are carefully examined.

To address these problems we have created the iRefWeb resource [22], which provides an easy access to PPI datasets consolidated from major public databases. Unlike the systems that integrate data automatically on-the-fly, iRefWeb offers its users a carefully constructed landscape of PPI data, with supporting

evidence and data provenance. The datasets are integrated using a well-defined and rigorous procedure based on peptide sequence matching, called iRefIndex [23], which is described in the next section. While an entire iRefIndex collection of consolidated PPIs can be downloaded at <https://irefindex.vib.be/>, iRefWeb provides a user-friendly web interface to the data, allowing users to invoke different types of filters to enable targeted searches and to examine the supporting evidence for different PPIs.

A key functionality of iRefWeb is its ability to compare annotations extracted by different curator teams: this comparative analysis helps the users to assess the overlaps between co-annotations of the same publication, identify the items present in some annotations but absent in others, and thus gauge the difficulties and ambiguities faced by the curators. Ultimately the users may choose to follow the link to the original publication and examine for themselves the strength of the supporting evidence for each PPI described therein.

A related feature is the PPI confidence score in iRefWeb, which reflects the available pieces of evidence in support of the PPI. For example, PPIs annotated in multiple studies have a higher confidence than those that were described only once. Interactions whose counterparts appear in other organisms, based on the orthology of the participating proteins, serve as additional validation. Certain interaction types (e.g., “direct interaction”) and detection methods used in the primary study also contribute to stronger confidence in the interaction record. iRefWeb quantifies these pieces of supporting evidence and combines them into a unified MI confidence score based on the methodology developed by the MINT database [24].

2 iRefIndex Consolidation

This section presents a summary of iRefIndex procedure for integrating PPI records from different source databases [23], which are made available for exploration through the iRefWeb portal. iRefIndex is one of the most rigorous procedures for consolidating redundant PPI records to date. iRefIndex V.16 released in 2019 provides a more recent version of the consolidated PPI data landscape than the current iRefWeb, which is based on an earlier release of iRefIndex (V.13) compiled in 2014. We also give a short description of the contents of iRefIndex V.16 and means to access the more recent version of the global PPI landscape it contains.

2.1 Summary of the iRefIndex Consolidation Procedure

For each interaction record retrieved from a source database, iRefIndex generates two sets of keys: one key for the interaction record and one for each participant protein. These keys are based solely on the sequence of the proteins, their taxonomy identifiers, and use the Secure Hash Algorithm as implemented in the SEGUID database [25]. Two interaction records will have identical keys if they

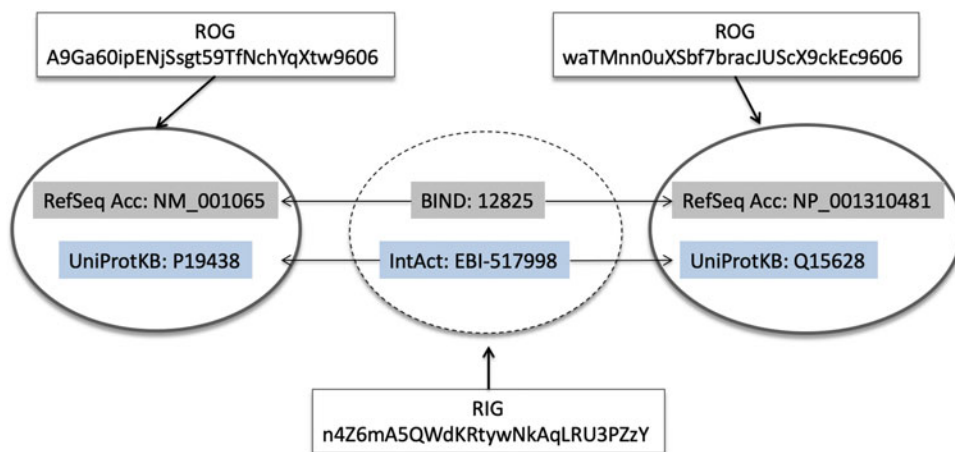


Fig. 1 Schematic illustration of how ROGs and RIGs are defined in iRefIndex. The two full line ellipsoids represent redundant object groups (ROG's). Each ROG contains a set of protein sequence accession numbers that point to records describing the exact same protein sequence from the same organism. The dotted ellipsoid represents a redundant interaction group (RIG). This RIG contains a set of protein interaction accession number that point to records describing interactions between the same two proteins (ROG's). Unique identifiers for ROG's and RIG's can be calculated independently using the primary sequences of the proteins, their taxonomy identifiers, and the SHA-1 algorithm (Secure Hash Algorithm digest)

refer to the same set of identical protein sequences and taxonomy identifiers.

iRefIndex assigns records with identical keys to a Redundant Object Group (ROG). Each interaction record involving only protein interactors is then assigned to a Redundant Interaction Group (RIG) on the basis of the ROG assignments (Fig. 1). A RIG identifier (RIGID) is then constructed by concatenating ROG identifiers and applying the SHA-1 algorithm (Secure Hash Algorithm digest) to the resulting string. The RIGID constitutes a unique and universal pointer to a set of interaction records (or records of a protein complex) that all involve the same proteins from the same organism. That being said, a given RIGID may involve proteins from different organisms, a rather common occurrence for interactions involving human and mouse proteins [26].

The crucial step of this consolidation procedure is to map protein database references found in the interaction record of the source database to ROGs. Most often, this reference consists of an external database identifier (e.g., UniProt) and an accession pointing to a record in that database (e.g., P31946), as well as the taxonomy identifier for the protein. But this is not always the case, as protein references in some interaction databases may be malformed, deprecated, ambiguous or outright missing, requiring several indirect steps to complete the mapping, or declare forfeit. The operations performed during the mapping process are therefore described by a mapping score, which reflects the ease with

which the protein reference provided by the source database could be matched to its sequence and taxon. This score should not be confused with a reliability score sometimes associated with a detected interactions [27, 28] as done in some recent publications [29].

The keys assigned to individual proteins that participate in a given interaction, or are part of a protein complex, enable the retrieval of all the information about a given interaction, independently of the protein references used in the original records of the source database.

2.2 iRefIndex Release V.16

The most recent release V.16 of the protein interaction index involved parsing PSI-MI files provided by the following 12 interaction databases, at the date of the build (March/April, 2019): BIND [30], BIND_TRANSLATION (a version of the BIND database recast in PSI-MI 2.5 XML format [<http://download.baderlab.org>]), BioGRID 3.5.171 [31], CORUM [32], DIP [33], HPRD [34], IntAct [21], IntAct Complex [35], InnateDB [36], MatrixDB [37], MPact [38], MPPI [39].

In addition the following sources were parsed starting from MITAB 2.5 formatted csv files extracted from the PSICQUIC [19] or the locally hosted web service: HPIDb [40], MINT [24], Reactome [41], UniProt [42], VirHostNet [43], as well as BHF-UCL, MPIDB, MBInfo, and QuickGO (for links to data downloaded from these four resources see <http://irefindex.vib.be/>).

In total, 3,169,715 records were retrieved from 21 different source databases, including 2,114,475 records where all described interacting elements (“interactors”) are proteins. As in previous iRefIndex and iRefWeb releases, records involving small molecules and proteins are not indexed. A summary of the interaction records retrieved from each of the 21 source databases, the protein-containing records, and the resulting consolidated interaction redundant groups (RIGs/RIGIDs) are presented in Table 1.

The indexed protein-only records and various associated data items from each record were parsed into a PostgreSQL database, which has been aggregated in the PSICQUIC View. Data of the V.16 release and details of the included information are also available in taxon-specific divisions via the PSI-MITAB 2.5 tab-delimited text format (at <http://irefindex.vib.be>).

The incremental change in the number of interaction records for major model organisms in iRefIndex V.16, relative to the V.13 release currently accessible in iRefWeb, is listed in Table 2. For human, mouse, and plants (*A. thaliana*) the number of consolidated interactions has more than doubled, whereas for the yeast *S. cerevisiae* and *E. coli*, the incremental change has been minor (<15%). The number of interactions, proteins, and curated articles (Pubmed IDs) for major model organisms and some less well

Table 1
Summary of mapping interaction records to RIGs (redundant interaction groups) in the iRefIndex V.16 build

Source DB	Total records	Protein-related PPI	PPI assigned RIGID	PPI assigned unique RIGID	% PPI assigned unique RIGID
BHF-UCL	2341	2328	2328	1515	65.08
BIND	157,736	153,063	73,206	54,161	73.98
BIND_TRANSLATION	192,923	84,138	82,228	60,872	74.03
BioGrid	1,653,530	778,945	775,480	568,254	73.28
CORUM	4274	4274	4270	4018	94.10
DIP	81,731	80,134	79,879	77,472	96.99
HPIDb	3007	2840	2840	1558	54.86
HPRD	83,022	83,022	82,983	40,542	48.86
InnateDB	18,408	18,300	17,807	12,728	71.48
IntAct	571,739	520,992	520,864	329,941	63.34
IntAct Complex	2536	2016	2016	1995	98.47
MatrixDB	36,945	36,867	36,867	22,374	60.69
MBInfo	542	522	522	331	63.41
MINT	81,305	80,746	80,731	44,969	55.70
MPact	16,504	16,504	16,373	13,398	81.83
MPIDB	1505	1504	1425	893	62.67
MPPI	1814	1758	1578	776	49.18
QuickGO	71,979	58,723	56,583	28,741	50.79
Reactome	141,996	141,996	141,844	130,128	91.74
UniProt	11,118	11,033	11,033	6239	56.55
VirHostNet	34,760	34,760	34,760	30,178	86.82
(All)	3,169,715	2,114,475	2,025,626	1,079,693	53.30

The source databases are listed in column 1. Columns 2 and 3 list, respectively, the total number of records, and protein-related records, retrieved from each source database. Columns 4 and 5 list the absolute number of protein-protein interactions (PPIs) assigned to all RIGs and to unique RIGs, respectively. Column 6 lists the fraction of unique RIGs, computed as the ratio of unique over all RIGs. RIGID constitutes a unique and universal pointer to a set of interaction or complex records that all involve the same proteins from the same organism. The last row lists the total number (or value) in each column

studied ones, available in the V.16 release of iRefIndex, is pictorially summarized in Fig. 2.

Presently, iRefIndex is updated by rebuilding the entire dataset. Releases are accompanied by a detailed README file listing the

Table 2
Incremental change in the number of interaction records for major model organisms in iRefIndex V.16 relative to V.13 currently available in iRefWeb

Taxonomy ID	Organism name	# PPI V.16	#PPI V.13	% Change
9606	<i>Homo sapiens</i>	662,814	222,098	198
559292	<i>S. cerevisiae</i> S288C	133,463	117,029	14
7227	<i>D. melanogaster</i>	74,582	44,906	66
10090	<i>Mus musculus</i>	65,514	30,137	117
3702	<i>A. thaliana</i>	57,809	21,454	169
6239	<i>C. elegans</i>	16,947	14,102	20
83333	<i>E. coli</i> K-12	16,706	15,269	9
192222	<i>C. jejuni</i> subsp. <i>jejuni</i>	11,930	11,973	~0
284812	<i>S. pombe</i> 972h-	10,179	8626	18

Columns 1–4 list, respectively, the NCBI taxonomy ID, the organism name, the number of consolidated interaction records in V.16 and the number of such records in V.13

release number, release date, a detailed description of the format and any change notices at <http://irefindex.vib.be/>. The code base of the current release V.16 is available at <https://github.com/abotzki/irefindex>.

Tools to access and analyze the data in iRefIndex are provided by PSICQUIC. These tools allow to perform various operations such as: selecting PPI data from specific source databases or publications (defined using PubMed IDs), or interactions detected by specific experimental methods. They also enable searching for specific proteins, separating binary interactions from complexes, computing general database statistics, and performing other types of specialized analyses. Alternatively, the iRefIndex V.16 dataset can also be queried via the Bioconductor package PSICQUIC (<https://www.bioconductor.org/packages/release/bioc/html/PSICQUIC.html>).

3 Materials

This section describes several pieces of information required to query the iRefWeb portal. These items are publicly available from third-party resources and may be used optionally, or in combinations, to refine the data searches and retrieval.

We will demonstrate an iRefWeb exploration using human histone deacetylase (HDAC) proteins, which are chromatin modifying factors involved in key epigenetic processes in the cell. In

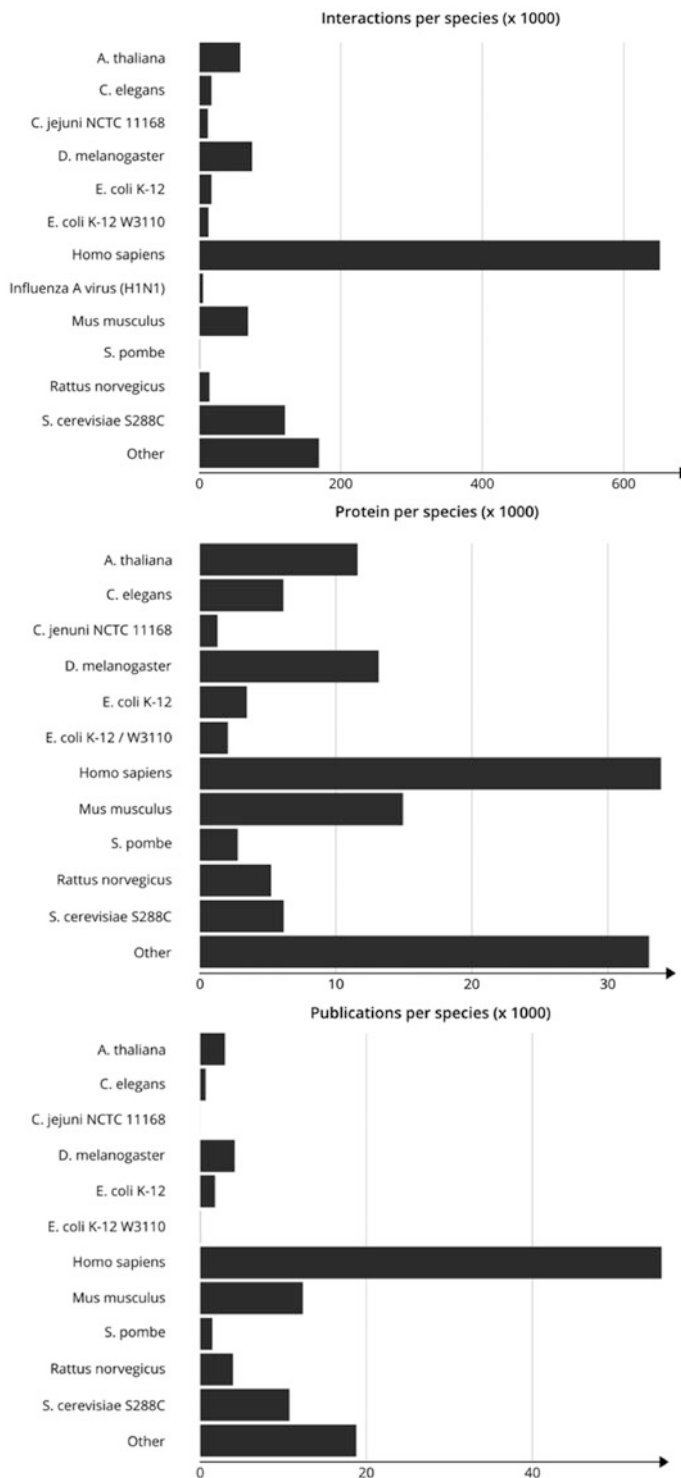


Fig. 2 Per organism statistics for iRefIndex V.16. Pictorial summary of the number of nonredundant interactions, proteins and publications for major model organisms consolidated from the 21 listed source databases and made available in iRefIndex release V.16. "Other" stands for other organisms

preparation for this scenario, the user may wish to examine several pieces of information outlined below.

1. Search the UniProtKB at <http://www.uniprot.org> [42] for “human histone deacetylase”. Observe that the histone deacetylase 1 protein has the UniProt identifier Q13547 and name HDAC1_HUMAN; whereas histone deacetylase 2 has the UniProt identifier Q92769 and name HDAC2_HUMAN (*see Note 1*).
2. Search NCBI Gene database at <http://www.ncbi.nlm.nih.gov/gene> [45] for “human histone deacetylase”. Observe that the histone deacetylase 1 gene has the Gene ID 3065 and histone deacetylase 2 has Gene ID 3066 (*see Note 2*).
3. Search PubMed at <https://www.ncbi.nlm.nih.gov/pubmed> for the publication ID 9150135. Examine the abstract of the paper by Zhang et al. [44, 45] and observe the names of the interactors involved in the human Sin3a protein complex, including HDAC1 and HDAC2.

4 Methods

In this section, we demonstrate how to search iRefWeb for PPIs involving human histone deacetylase proteins HDAC1 and HDAC2, and how to examine the supporting evidence for such interactions.

4.1 Search for PPIs Involving Specific Proteins

1. Start at the iRefWeb search page at <http://wodaklab.org/iRefWeb/search>.
2. Enter “hdac1” in the Left search box in the panel *Search Terms*. Once a popup panel with multiple matching interactors appear, select HDAC1_HUMAN using the checkbox, then click the button “Add your checked selections to your search query term” (*see Note 3*).
3. In the panel Search Filters, click the link *Expand All Filters*, which is updated automatically following each selection. Examine the PPI counts next to each item (Fig. 3).
4. Observe that there are 688 interactions involving HDAC1_HUMAN in the current iRefWeb version, of which the vast majority are single-organism interactions, i.e., those involving only human proteins (651 out of 688). However, some PPIs also involve proteins from mouse, rat, fruit fly, zebra fish, and several viral species, as could be seen in the filter.
5. Observe that the majority of the interaction records are supplied by the BioGrid source database: 447 out of 688, or 65% of the total collection (*see Note 4*).

Display filter counts and search results by: protein (0) interaction (688) publication (0)
 Expand All Filters — Collapse All Filters — Show Filter Help

Source Database

- bind (18)
- bind_translation (16)
- biogrid (447)
- corum (66)
- dip (25)
- hprd (165)
- innatedb (24)
- intact (133)
- matrixdb (0)
- mpact (0)
- mpidb (0)
- mppi (1)
- ophid (90)

Can match ANY of these

- seen by 1 DB (522)
- seen by 2 DBs (79)
- seen by 3 or more DBs (87)

Organism **

- single organism interaction (651)
- cross organism interaction (37)

- Homo sapiens (688)
- Saccharomyces cerevisiae S288c (0)
- Drosophila melanogaster (1)
- Mus musculus (10)
- Arabidopsis thaliana (0)
- Escherichia coli K-12 (0)
- Caenorhabditis elegans (0)
- Campylobacter jejuni subsp. jejuni NCTC 11168 (0)
- Schizosaccharomyces pombe 972h- (0)
- Rattus norvegicus (4)

More filters hidden

Can match ANY of these

Nature of Interaction

- unary (1)
- pairwise (511)
- multi-subunit (176)

- predicted (90)
- experimental (676)

- genetic (0)
- physical (688)

MI (MINT-Inspired) Score

Interaction Detection Method *

- decarboxylation assay (6)
- deglycosylase assay (90)
- fluorescence technology (16)
- mass spectrometry study of hydrogen/deuterium exchange (16)
- partial identification of protein sequence (2)
- phage display (44)
- static light scattering (411)
- unknown (42)

More filters hidden

Can match ANY of these

MI (MINT-Inspired) Organism Percentile

Interaction Type *

- aggregation (90)
- association (148)
- colocalization (45)
- dephosphorylation reaction (0)
- direct interaction (149)
- molecular interaction (203)
- phosphorylation reaction (1)
- physical association (417)

More filters hidden

Can match ANY of these

Number of Publications

- no valid PubMed ID (0)
- 1 or more publications (688)
- 2 or more publications (195)
- 3 or more publications (110)
- 4 or more publications (72)

Can match ANY of these

Fig. 3 iRefWeb filter panel. The filter panel presents the option to select specific types of PPI records, organisms of interests, or supporting evidence. The numbers next to each item represent the PPI counts in the corresponding category, and are dynamically updated once search terms are entered or filter items are selected. The currently shown counts represent all PPIs that involve HDAC1_HUMAN interactor

6. Click on the blue Search button to retrieve the table of results.
7. Click on the green Download Interactome button, then select MITAB option in the popup window, in order to save the results. The retrieved tab-delimited file has 54 columns (*see Note 5*).

4.2 Refine the Search Using Additional Proteins and Filters

1. Now enter “hdac2” in the Right search box in the panel *Search Terms* and select HDAC2_HUMAN in the popup panel, then click the button “Add your checked selections to your search query term” (*see Note 6*).
2. Observe that the PPI counts in the filters have changed, representing interactions that involve both HDAC1 and HDAC2

proteins. There are 90 such interactions in total. The filter subpanel Nature of Interaction shows that all but one of the PPIs are multi-subunit records, i.e., representing protein complexes. Only one result is a pairwise PPI.

3. Observe that now CORUM provides the largest number of results among all databases (37 out of 90 PPI records), which is perhaps not surprising given its focus on mammalian complexes. The second largest contribution (35 PPI records) comes from IntAct, which is a leading member of the IMEx consortium.
4. Select the check box next to the filter “pairwise” in the Nature of Interaction subpanel. Observe that this selection restricts the results to the single pairwise PPI between HDAC1 and HDAC2 proteins.
5. Observe that now the Source Database filters indicate that this PPI record is contributed by four different databases: BioGrid, HPRD, IntAct, and Ophid.
6. Click on the blue Search button to retrieve results, then click on the link “1026051” in the Interaction ID column to load the interaction record.
7. Examine the table of supporting information for the PPI extracted from the source databases: each row in the table represents a separate annotation including the original publication represented by the PubMed ID; the interaction type represented by the PSI-MI ontology term; and the interaction detection method represented by another PSI-MI term. This interaction has a very strong support from different source databases and from numerous publications (*see* **Notes 7 and 8**).
8. Observe the information on the MI Score near the top of the page, indicating 1.00—a very high degree of confidence. Click on the link “1.00” to retrieve a page with the MI score details.
9. Click “Show details” to review the many individual pieces of supporting evidence for this PPI, such as the list of 80 supporting PubMed publications, the list of many protein complexes in which the HDAC1-HDAC2 pair appears, and a list of interactions between orthologs of HDAC1 and HDAC2 in other organisms.

4.3 Examine the Differences in Annotations from Different Source Databases

1. Return to the PPI page for the interaction 1026051 between human HDAC1 and HDAC2 proteins. In the table of supporting evidence, observe three occurrences of the PubMed ID 9150135. These three table rows represent original records from BioGrid and IntAct: although both databases recorded the interaction type as “physical association” (MI:0915), their curation of the interaction detection methods is different: the BioGrid used the term “static light scattering” (MI:0104)

Interaction ID	Accession	Label	Organism	biogrid	corum	hprd	intact	ophid
745836	Q09028	RBBP4_HUMAN	H. sapiens	biogrid			intact	
745836	Q13547	HDAC1_HUMAN	H. sapiens	biogrid			intact	
873668	P24863-2	CCNC	H. sapiens	biogrid				
873668	P49336	CDK8_HUMAN	H. sapiens	biogrid				
915415	Q13547	HDAC1_HUMAN	H. sapiens			hprd	intact	ophid
915415	NM_005870	SAP18	H. sapiens			hprd	intact	ophid
977438	NM_005870	SAP18	H. sapiens		corum			
977438	Q09028	RBBP4_HUMAN	H. sapiens		corum			
977438	Q16576	RBBP7_HUMAN	H. sapiens		corum			
977438	Q13547	HDAC1_HUMAN	H. sapiens		corum			
977438	Q96ST3	SIN3A_HUMAN	H. sapiens		corum			
977438	Q92769	HDAC2_HUMAN	H. sapiens		corum			
977438	O75446	SAP30_HUMAN	H. sapiens		corum			
1012650	Q16576	RBBP7_HUMAN	H. sapiens	biogrid			intact	
1012650	Q13547	HDAC1_HUMAN	H. sapiens	biogrid			intact	
1026051	Q92769	HDAC2_HUMAN	H. sapiens	biogrid			intact	
1026051	Q13547	HDAC1_HUMAN	H. sapiens	biogrid			intact	
1078501	Q13547	HDAC1_HUMAN	H. sapiens				intact	
1081060	Q96ST3	SIN3A_HUMAN	H. sapiens			hprd		
1081060	Q09028	RBBP4_HUMAN	H. sapiens			hprd		
1647276	Q92769	HDAC2_HUMAN	H. sapiens			hprd		
1647276	NM_005870	SAP18	H. sapiens			hprd		
1647276	Q09028	RBBP4_HUMAN	H. sapiens			hprd		
1647276	Q13547	HDAC1_HUMAN	H. sapiens			hprd		
1647276	O75446	SAP30_HUMAN	H. sapiens			hprd		
1647276	icrogid:49542800	icrogid:49542800	H. sapiens			hprd		

Fig. 4 iRefWeb annotation for a specific publication. The annotation page summarizes all curation efforts from source databases for a given publication. The table also shows which PPIs or protein interactors were curated by each database, helping the user to examine the similarities and differences between individual annotations. The annotation table shown here corresponds to the PubMed ID 9150135

whereas IntAct recorder two different ontology terms, “bimolecular fluorescence complementation” (MI:0809) and “methyltransferase assay” (MI:0515), based on the same publication.

- To examine all available annotations for the PubMed ID 9150135, click on the link “9150135” in the Evidence table, which loads the annotation page for this primary publication (*see Note 9*).
- Observe that the article was curated by five different source databases, resulting in a total of nine distinct interactions involving ten distinct proteins (Fig. 4).

As can be seen from the annotation table, there are numerous differences between the curation records from the individual databases. Some of these relate to representing protein complexes as either a series of pairwise interactions or a single multi-subunit interaction. For example, the HDAC1-HDAC2 pair appears as a separate PPI record in BioGrid and IntAct, but as part of a larger multi-subunit complexes recorder by

CORUM and HPRD. However, the CORUM complex contains seven protein subunits whereas HPRD has only six. Observe that some of the pairwise PPIs are present in both BioGrid and IntAct while other are present in only one of these databases but not the other. For example, none of the PPIs from BioGrid include the histone deacetylase complex subunit SAP18, whereas all other databases included this interactor in their curation records. Only CORUM and HPRD included the protein SIN3A.

4. To assess the information in the original publication, expand the Abstract link near the top of the page. Both the abstract and the article title: “*Histone deacetylases and SAP18, a novel polypeptide, are components of a human Sin3 complex*” indicate that SAP18 and the Sin3a complex are the main subject of the study [44]. Interested users may then review the original publication and form their own opinion about the strength of the supporting evidence for each of the interactors and PPIs.

5 Notes

1. The UniProt search for “human histone deacetylase” returns thousands of records. Near the top of the list are not only several HDAC proteins but also their interactors that are part of the same protein complexes as the HDACs. These include histone deacetylase complex subunit SAP18 and SAP30, histone-binding proteins RBBP4 and RBBP7, and paired amphipathic helix protein SIN3A.
2. The NCBI search also shows various aliases of the two genes of interest: HDAC1 is also known as GON-10, HD1, KDAC1, RPD3, and RPD3L1; and HDAC2 is also known as HD2, KDAC2, RPD3, and YAF1. This demonstrates the need to use stable identifiers to represent interactors, and the potential ambiguities faced by the curators when they encounter gene names in the publications.
3. This action may also be performed by searching the Left box for the UniProt accession “Q13547” or the gene ID “3065.”
4. The fact that BioGrid provides the majority of the interactions is influenced in part by the representation of protein complexes: whereas many other databases may represent a complex as a single interaction record with multiple constituent proteins, BioGrid breaks it down into a series of separate pairwise interactions, thereby providing a larger number of individual PPI records.

5. A full description of each column of the PSI-MITAB file format is available at: https://irefindex.vib.be/wiki/index.php/README_MITAB2.6_for_iRefIndex_13.0.
6. As in the case of HDAC1, this action may also be performed by searching the Right box for the UniProt accession “Q92769” or the gene ID “3066.”
7. In this PPI record the majority of interaction types were curated using the terms “association” (MI:0914) and “physical association” (MI:0915), which indicates that the two proteins HDAC1 and HDAC2 were detected in the same protein complex but not necessarily in a direct physical contact.
8. The presence of a generic terms “molecular interaction” (MI:0000) indicates that the information on a specific interaction type was not provided in the original source-database record in the manner compliant with the PSI-MI ontology.
9. Alternatively, users may retrieve the same annotation table by searching for the 9150135 in the text box Pubmed IDs at the main iRefWeb search page <http://wodaklab.org/iRefWeb/search>.

Acknowledgements

This work was supported by the Canadian Institutes of Health Research (MOP#82940), the Ontario Research Fund, and the SickKids Foundation. S.J.W. and A.B. are grateful for help from Christof De Bo (VIB Bioinformatics Core Team, Ghent, Belgium). This chapter is dedicated to the memory of Brian Turner, the creator of the iRefWeb portal. He was an outstanding professional with many talents, following his own path. He will be thoroughly missed.

References

1. Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* 92:291–294
2. Kocher T, Superti-Furga G (2007) Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat Methods* 4:807–815
3. Chiu W, Baker ML, Almo SC (2006) Structural biology of cellular machines. *Trends Cell Biol* 16:144–150
4. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP et al (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440:637–643
5. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141–147
6. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N et al (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437:1173–1178
7. Guruharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O et al (2011) A protein complex

- network of *Drosophila melanogaster*. *Cell* 147:690–703
8. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S et al (2012) A census of human soluble protein complexes. *Cell* 150:1068–1081
 9. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N et al (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* 433:531–537
 10. Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput Biol* 3:e42
 11. Phizicky EM, Fields S (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* 59:94–123
 12. Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. *Nucleic Acids Res* 34:D504–D506
 13. Turinsky AL, Razick S, Turner B, Donaldson IM, Wodak SJ (2010) Literature curation of protein interactions: measuring agreement across major public databases. *Database* 2010:baq026
 14. Chaurasia G, Malhotra S, Russ J, Schnoegl S, Hanig C, Wanker EE, Futschik ME (2009) UniHI 4: new tools for query, analysis and visualization of the human protein-protein interactome. *Nucleic Acids Res* 37:D657–D660
 15. Orchard S, Binz PA, Borchers C, Gilson MK, Jones AR, Nicola G, Vizcaino JA, Deutsch EW, Hermjakob H (2012) Ten years of standardizing proteomic data: a report on the HUPO-PSI Spring Workshop: April 12–14th, 2012, San Diego, USA. *Proteomics* 12:2767–2772
 16. Orchard S, Kerrien S, Jones P, Ceol A, Chatr-Aryamontri A, Salwinski L, Neroth J, Hermjakob H (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics* 7(Suppl 1):28–34
 17. Orchard S, Hermjakob H, Apweiler R (2003) The proteomics standards initiative. *Proteomics* 3:1374–1376
 18. Kerrien S, Orchard S, Montecchi-Palazzi L, Aranda B, Quinn AF, Vinod N, Bader GD, Xenarios I, Wojcik J, Sherman D et al (2007) Broadening the horizon—level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol* 5:44
 19. del Toro N, Dumousseau M, Orchard S, Jimenez RC, Galeota E, Launay G, Goll J, Breuer K, Ono K, Salwinski L et al (2013) A new reference implementation of the PSICQUIC web service. *Nucleic Acids Res* 41:W601–W606
 20. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L et al (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res* 41:D816–D823
 21. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U et al (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 40:D841–D846
 22. Turner B, Razick S, Turinsky AL, Vlasblom J, Crowdy EK, Cho E, Morrison K, Donaldson IM, Wodak SJ (2010) iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* 2010:baq023
 23. Razick S, Magklaras G, Donaldson IM (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* 9:405
 24. Ceol A, Chatr-Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* 38:D532–D539
 25. Babnigg G, Giometti CS (2006) A database of unique protein sequence identifiers for proteome studies. *Proteomics* 6:4514–4522
 26. Turinsky AL, Razick S, Turner B, Donaldson IM, Wodak SJ (2011) Interaction databases on the same page. *Nat Biotechnol* 29:391–393
 27. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol Cell Proteom* 6:439–450
 28. Pu S, Vlasblom J, Turinsky A, Marcon E, Phanse S, Trimble SS, Olsen J, Greenblatt J, Emili A, Wodak SJ (2015) Extracting high confidence protein interactions from affinity purification data: at the crossroads. *J Proteome* 118:63–80
 29. Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkiewicz G, Workman CT, Rigina O, Rapacki K, Staerfeldt HH et al (2017) A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat Methods* 14:61–64
 30. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW (2001) BIND—the biomolecular interaction network database. *Nucleic Acids Res* 29:242–245

31. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A et al (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 45:D369–D379
32. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW (2010) CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res* 38:D497–D501
33. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res* 32:D449–D451
34. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A et al (2009) Human protein reference database—2009 update. *Nucleic Acids Res* 37: D767–D772
35. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del Toro N et al (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42:D358–D363
36. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock RE, Brinkman FS, Lynn DJ (2013) InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res* 41:D1228–D1233
37. Launay G, Salza R, Multedo D, Thierry-Mieg N, Ricard-Blum S (2015) MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. *Nucleic Acids Res* 43:D321–D327
38. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res* 34: D436–D441
39. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW et al (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics* 21:832–834
40. Ammari MG, Gresham CR, McCarthy FM, Nanduri B (2016) HPIDB 2.0: a curated database for host-pathogen interactions. *Database* 2016:baw103
41. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B et al (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 46:D649–D655
42. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45:D158–D169
43. Guirimand T, Delmotte S, Navratil V (2015) VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res* 43:D583–D587
44. Zhang Y, Iratni R, Erdjument-Bromage H, Tempst P, Reinberg D (1997) Histone deacetylases and SAP18, a novel polypeptide, are components of a human Sin3 complex. *Cell* 89:357–364
45. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S et al (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 40:D13–D25



Chapter 13

State-of-the-Art Data Management: Improving the Reproducibility, Consistency, and Traceability of Structural Biology and in Vitro Biochemical Experiments

David R. Cooper, Marek Grabowski, Matthew D. Zimmerman, Przemyslaw J. Porebski, Ivan G. Shabalin, Magdalena Woinska, Marcin J. Domagalski, Heping Zheng, Piotr Sroka, Marcin Cymborowski, Mateusz P. Czub, Ewa Niedzialkowska, Barat S. Venkataramany, Tomasz Osinski, Zbigniew Fratzczak, Jacek Bajor, Juliusz Gonera, Elizabeth MacLean, Kamila Wojciechowska, Krzysztof Konina, Wojciech Wajerowicz, Maksymilian Chruszcz, and Wlodek Minor

Abstract

Efficient and comprehensive data management is an indispensable component of modern scientific research and requires effective tools for all but the most trivial experiments. The LabDB system developed and used in our laboratory was originally designed to track the progress of a structure determination pipeline in several large National Institutes of Health (NIH) projects. While initially designed for structural biology experiments, its modular nature makes it easily applied in laboratories of various sizes in many experimental fields. Over many years, LabDB has transformed into a sophisticated system integrating a range of biochemical, biophysical, and crystallographic experimental data, which harvests data both directly from laboratory instruments and through human input via a web interface. The core module of the system handles many types of universal laboratory management data, such as laboratory personnel, chemical inventories, storage locations, and custom stock solutions. LabDB also tracks various biochemical experiments, including spectrophotometric and fluorescent assays, thermal shift assays, isothermal titration calorimetry experiments, and more. LabDB has been used to manage data for experiments that resulted in over 1200 deposits to the Protein Data Bank (PDB); the system is currently used by the Center for Structural Genomics of Infectious Diseases (CSGID) and several large laboratories. This chapter also provides examples of data mining analyses and warnings about incomplete and inconsistent experimental data. These features, together with its capabilities for detailed tracking, analysis, and auditing of experimental data, make the described system uniquely suited to inspect potential sources of irreproducibility in life sciences research.

Key words Structural biology, Databases, LIMS, Reproducibility

David R. Cooper and Marek Grabowski contributed equally to this work.

Yu Wai Chen and Chin-Pang Benu Yiu (eds.), *Structural Genomics: General Applications*, Methods in Molecular Biology, vol. 2199, https://doi.org/10.1007/978-1-0716-0892-0_13, © Springer Science+Business Media, LLC, part of Springer Nature 2021

1 Introduction

The problem of managing experimental data is as old as the research laboratory itself, and the efficient and comprehensive data management is an indispensable component of modern scientific research. The contemporary understanding of data management defines it as a “process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users” [1]. In recent years, data management has been increasingly recognized as one of the most vital factors affecting the reproducibility of research data. On the other hand, data management problems are quite often underestimated by both scientists and the general public. The widely publicized recent examples from the airline industry show that poor data management (as well as management in general) may have unpleasant consequences, like the public spectacle of dragging a passenger from a plane, which resulted in negative publicity and subsequent drop in stock value for one of the major airlines. Biomedical data management is generally much more sophisticated but still not perfect. Inconsistencies and errors in the public record are usually detected and corrected by the collective efforts of other scientists, but this is not an instantaneous process and is hampered by the difficulty of reporting negative results. The losses associated with lack of reproducibility are estimated to be on the order of many billions of dollars [2]. Identified inconsistencies are often followed by detailed analysis and in many cases by the correction of errors and frequently the sources of errors.

Traditionally, data management in research laboratories has been addressed by simple approaches such as paperbound lab notebooks and, since the 1980s, computerized spreadsheets. However, these approaches do not remove or even track inconsistencies and do not scale well to the requirements of modern biomedical research, especially in large-scale, high-throughput collaborative programs that generate vast amounts of experimental data in geographically distinct laboratories. These traditional approaches are often inadequate to assure the reproducibility of experiments, even when work is performed in a single laboratory. In the last 10 years, there has been an increasing awareness that the reproducibility of experimental research cannot be taken for granted [3, 4]. According to some estimates, about 50% of preclinical research (at the cost of around \$28 billion per year) may be irreproducible [2]. The concerns about reproducibility problems are motivating funding agencies worldwide to introduce new requirements for managing and sharing data generated from sponsored research [5].

Recent advances in information technology have led to the development of database-driven platforms to efficiently collect,

store, annotate, and analyze laboratory data. Electronic laboratory notebooks (ELNs) can be thought of as a digital replacement of a paperbound notebook. However, increasingly sophisticated laboratory information management systems (LIMSs) are slowly superseding the use of ELNs and spreadsheets [6]. Numerous electronic notebooks and LIMSs have been developed in academia and industry, not only for large-scale projects but also for traditional, small to mid-size laboratories. To address particular experimental problems, diverse specialized systems have been designed to track specific kinds of sequential, microarray, metabolomics, proteomics, chemical, pharmacological, structural, and functional data [7–12]. Manufacturers of laboratory equipment often provide proprietary tools for tracking the data collected by these systems, although these tools are usually limited in scope and tied to the manufacturer's equipment. Several commercial LIMSs have been released for specialized data management tasks [13–15], but while they have been widely adopted in clinical labs, most off-the-shelf systems are not versatile enough to capture the different types of experimental data that are generated by academic biomedical research [16]. Biological and biomedical data are highly interconnected, and effective data management systems must take into account the diversity of data and experimental methods. To our knowledge, no data management system can accommodate the breadth of information that is necessary to encompass the “big picture” for any substantial biomedical project. For that reason, none of the existing systems have so far reached widespread acceptance in academia. For big pharma, the data management systems are so valuable that they often do not disclose any information about them.

Structure-function research, which is a major focus of this chapter, requires the full characterization of proteins and other macromolecules, including 3-D structure determination. The structure determination pipeline includes cloning, protein production, method-specific sample preparation (such as crystallization, deuteration, or cryo-EM grid preparation, structure solution, and model refinement). Biochemical and biophysical experiments are usually performed in advance and are often critical during the structure determination and interpretation stage. Conversely, the 3-D macromolecular models that are experimentally determined frequently inspire subsequent functional experiments, which may include ligand binding experiments or mutational analysis to test models based on the structural interpretation. The final analysis usually requires the examination of the structural and functional information in the context of other similar macromolecular structures.

Integrating these diverse data presents a serious challenge for effective data management. In order to surmount this challenge, a number of LIMSs have been developed for structural biology, and many were designed by large-scale structural genomics

(SG) programs. These include Xtrack [12], SESAME [17], PiMS/xtalPiMS [18, 19], and HalX [20]. Some SG programs have relied on customized commercial LIMSs [21]. Modern crystallographic software suites, such as CCP4 [22, 23], Phenix [24, 25], and HKL-3000 [26, 27], organize the computational data they generate. These systems make it relatively easy for crystallographers to keep track of the parameters and results of calculations along the path from crystallographic data to a refined model that is ready for publication and deposition into the PDB.

Here, we describe our experiences managing structural biology data using a component-based data management system that we have developed for the acquisition, validation, storage, and analysis of biomedically oriented experimental data. The general description applies to any modern data management system; however, the examples and some details presented here reflect our experiences developing the LabDB modular LIMS. LabDB is composed of several separate components, each optimized to perform a particular task (Fig. 1). The reagents-tracking module, which tracks chemicals, laboratory supplies, and stock solutions, is a prerequisite for all other components. The components that are essential for a structural biology laboratory include protein production, crystallization, and structure determination modules.

The core of the system contains an underlying relational database and an associated web interface. Most of the components of the system are web-based, some use native interfaces, and some use a combination of both, according to the task involved. The database is directly interfaced with the *Xtaldb* application for designing, recording, and analyzing crystallization experiments [28–30] and the HKL-2000/-3000 crystallographic data processing and structure determination suite [26]. The integration of LabDB with the HKL suite allows scientists to automatically obtain information about the protein(s) and sample characterization during data collection, structure determination, and refinement. The initial structure determination results are directly transferred to LabDB, which allows others to design new biomedical experiments based on the structural information. The effect of the synergy of this integration cannot be overestimated.

LabDB focuses on minimizing human input by harvesting data directly from laboratory hardware whenever possible and uses several equipment-specific clients and modules for automated data acquisition. During the last 10 years, multiple instances of LabDB have been used to record experimental data for tens of thousands of protein targets in a number of large-scale high-throughput biomedical centers, including the Center for Structural Genomics of Infectious Diseases (CSGID), the Midwest Center for Structural Genomics (MCSG), New York Structural Genomics Research Consortium (NYSGRS), and the Enzyme Function Initiative (EFI).

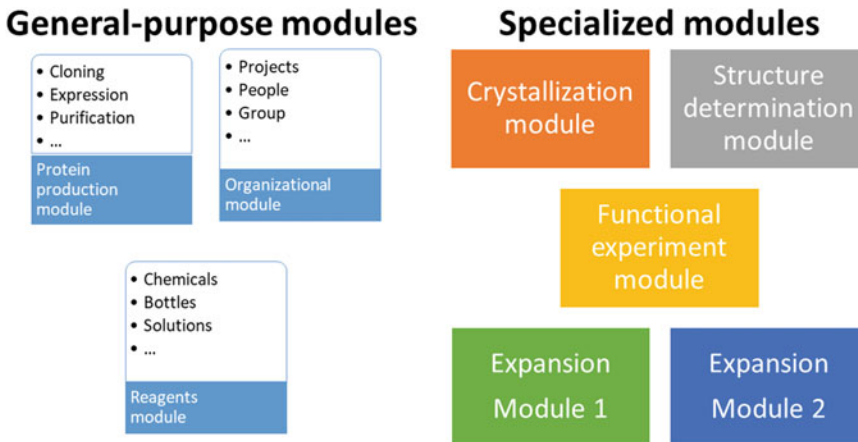


Fig. 1 The overall organization of LabDB, showing that expansion modules can be added to provide extra functionality

The system has played an important role in ensuring the reproducibility of experiments in the labs where it has been deployed.

2 Data Model, Data Acquisition, and Validation

2.1 Data Model

The data of interest in a structural biology laboratory is associated with entities representing different types of physical samples, experimental trials, calculations, etc. In turn, each of these data objects has attributes that describe specific details of these entities. Most existing LIMS implementations map the objects to a relational database with a schema comprised of the experimental components.

The LabDB relational database (currently LabDB uses the open-source PostgreSQL database) is arranged around “projects,” which correspond to individual proteins, macromolecules, or complexes. Mutants or other modifications of a protein (e.g., having different purification tags) are considered to be experimental variants within the same project. A given project usually encompasses multiple physical samples: clones, purified proteins, crystallization drops, harvested crystals, and others. Projects can optionally be aggregated into project groups. Data related to cloning, expression, purification, and biochemical characterization of the proteins are maintained by the “protein production” module, while data related to experimental crystallographic aspects are maintained by the crystallization component. The structure determination component, *bkldb*, keeps track of all computational crystallographic parameters, from data collection to deposition of the refined structural model into the PDB. *bkldb* is tightly integrated with the HKL-3000 structure determination suite that ensures automated metadata collection about the whole structure determination process.

LabDB was initially used to track all the steps of protein production, crystallization, and diffraction experiments for high-throughput structural biology and designed to require that a logical sequence of events would occur for each project. The origin of most experimental pipelines is a recombinant DNA clone, but a project can represent a complex of interacting proteins or a single protein purified from a natural source. In a typical workflow, clones are transformed into competent *E. coli* cells, and the encoded protein is expressed, purified, and crystallized for diffraction experiments. The purified proteins are often mixed with ligands, binding partners, or modifying enzymes before crystallization; in LabDB, such mixtures are referred to as “macromolecule preps” or just “macropreps.” During crystallization trials, the protein solutions are combined with a variety of components like buffers, salts, organic molecules, and/or ligands in an effort to grow crystals. When luck prevails, these usually fragile crystals are harvested and subjected to X-ray diffraction experiments [31]. Well-diffracting crystals can usually lead to interpretable electron density maps, which can be used to generate structural models of the protein. The final step is iterative refinement combined with model rebuilding and validation procedures. Each stage of this experimental pipeline has many parameters that must be recorded, as sometimes even slight changes can have dramatic effects on sample characterization. LabDB was designed to enforce recording a complete experimental provenance of any physical sample so that it can be traced back to the “source,” e.g., a clone, a purified protein, or a protein shipped from elsewhere.

2.2 Acquisition of Various Types of Laboratory Data

Acquisition of experimental data in a research laboratory can be performed in different ways depending on various factors, such as the type and complexity of the data, point of acquisition, available resources, and database/equipment compatibility. A major drawback of many existing LIMS is their overreliance on manual user input. While LIMS can provide benefits to experimenters, (e.g., the ability to easily share data with others, tools to analyze data, etc.), there can be drawbacks as well: namely, the additional time needed to enter and curate data. In practice, researchers are reluctant to adopt data management systems unless their benefits significantly outweigh the additional time outlay.

The underlying data acquisition concepts are presented in Fig. 2. In our experience, the integrity and completeness of the data are inversely proportional to the effort that is required from the user to input the data, especially in the case of “failed” experiments, where manual data entry is scarce and motivation to manually enter data may be particularly weak. If the required user effort is minimal and the data is automatically collected, the data will be more complete and generally have higher integrity (provided the equipment is working reliably). However, if the user is asked to

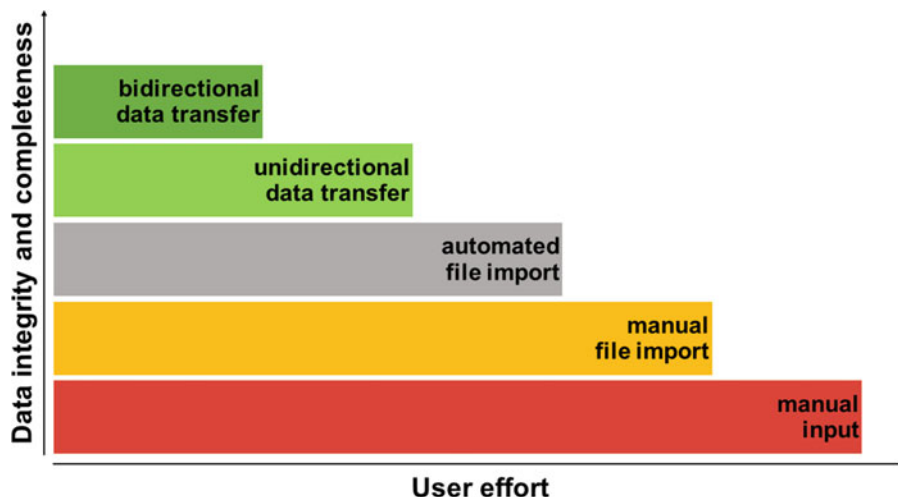


Fig. 2 Data integrity and completeness vs. the effort required from users to input data for various types of data entry

input data by hand, then the data will be limited and contain many errors and discrepancies. The difference is akin to tracking a person's location via phone GPS versus conducting a survey to determine where they were at a given time. Briefly, the methods of data acquisition can either be manual and automated, with further division into the complexity of input required by the user. The most time-consuming aspect of data acquisition is the manual input of data followed by manual upload of data files (e.g., importing data *en masse* from spreadsheets or comma-separated value (CSV) files or other output files produced from laboratory hardware) and manual entry of metadata. These last two steps are generally not required for the main experimental tasks but usually provide additional data; e.g., users may want to upload a gel associated with a protein purification (Fig. 3). Complex characterization steps such as thermal shift assays, isothermal titration calorimetry (ITC) experiments, and kinetic assays require a certain amount of metadata, which usually consists of an experimental protocol and the parameters of each experimental replicate.

Semiautomated data acquisition requires a user to either input or confirm some information that is either harvested automatically or is associated with an external event. LabDB can monitor computer folders or by running equipment-specific methods, with metadata provided by the user. More information can be extracted by uni- and bidirectional communication with the database/LIMS that are controlling the hardware and storing experimental results. Each direction of the communication improves data integrity. Bidirectional communication is the most complex to develop and requires cooperation between different vendors but provides the most benefits in terms of data integrity and completeness, as

View Gel - NewYork_Batch2_Gel1

[Edit gel information](#) | [Search gels](#)

Experiment information

Name	NewYork_Batch2_Gel1
Person	Niedzialkowska, Ewa
Experiment date	2013-05-15
Separation type	SDS-PAGE (protein)

Comment

Gel image

Attached files

NY_Oct2012_gel1_cut_48h.tif
gel_image_converted.png
gel_image_thumb.png
NY_Oct2012_gel1_cut_48h.scn

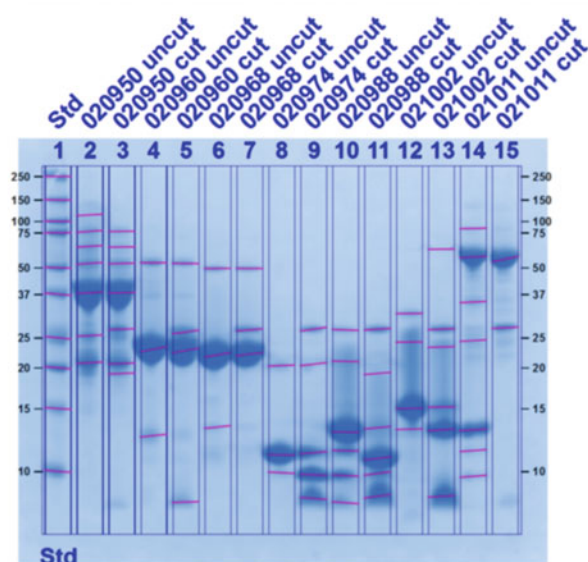


Fig. 3 Additional files, such as gel images can be associated with experiments in LabDB. LabDB can read additional data from image acquisition systems. The presented gel was acquired using a Bio-Rad Gel Doc™ EZ Gel Documentation System and processed and annotated using Image Lab™ Software

discussed below. However, this approach may be limited by the availability of a software development kit (SDK) or application programming interfaces (API) that would allow access to the data by external programs.

2.2.1 Reagents Module

The reagents component maintains an inventory of chemicals, stock solutions, and other reagents; their usage is tracked by other components. Chemicals are identified by CAS numbers and SMILES [32] representations, and pertinent information is downloaded from PubChem [33]. Bottles with chemicals or solutions

are labeled with barcodes. Solutions can be entered into LabDB through a web-based form. The form allows complex solutions to be created using a mixture of chemicals in their original chemical bottle and/or other solutions as their starting point. If a chemical bottle is used, then the chemical bottle's barcode can either be scanned or selected, and a final concentration is entered by the user. If a multicomponent solution is used as the starting point, all of the components of that solution are listed along with their stock concentration. Final concentrations are added by either entering the dilution factor or the final concentration of one of the components, thereby keeping the relative concentrations of solution components proportional. Multiple solutions or individual chemicals can be used to create a solution. Overall solution properties, such as name and volume, are required. The pH for the solution is not calculated from the individual components in the current version of LabDB. The default label will include a barcode, the solution's name, the creator's name, the date, and the solution components. Labels can be edited before they are printed on a network-connected label printer.

Storage System

The storage system contained within LabDB is a flexible, hierarchical system that allows storage containers to be placed within other containers. Each storage location can have one or more "children" locations, forming a set of hierarchical trees of ancestors and descendants, where the "root" storage locations are different rooms in the laboratory. Examples of storage containers in a room are the built-in shelves and large storage units such as freezers, refrigerators, and cabinets. The shelves of these large storage containers are themselves considered containers, as are any racks or boxes used to group similar reagents together. The hierarchical nature makes it possible to group items together and move them *en masse*. If one moves a freezer from one room to another, all the descendent storage locations (e.g., shelf 1, the blue box that is on shelf 2, etc.) and the individual items they contain are automatically moved. Each storage location can be assigned a barcode that can be used to quickly identify the location when assigning an item to a location (Fig. 4).

The storage location of individual items can be displayed in LabDB when viewing the item's details (or a list of items), and an itemized list of all the items contained within a storage location can be displayed in a tree view. The storage system has several quick entry methods. One can scan a storage location's barcode and then scan numerous items as they are placed in the location. In the inventory check feature, where one scans all of the items in a location, LabDB will indicate which items were missing or unexpectedly present (along with their currently assigned location). The inventory can be updated or changed on a per item basis. We are

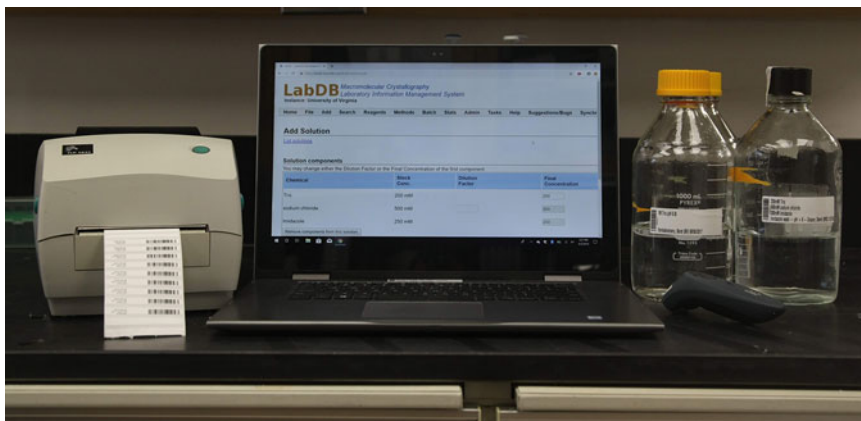


Fig. 4 Barcoding stock solutions in LabDB

currently adding safety information to the storage system, to alert users, or authorities in case of emergencies, about potentially harmful materials that are contained within a storage location.

2.2.2 Protein Production Module

The “Protein Production” component of LabDB stores information about protein cloning, expression, and purification. Experimental data are organized in a tree-like data structure: each experimental step must be connected to a preceding step (e.g., expression must be tied to a particular clone). Data can be entered manually by researchers using a web browser as a specific step is completed; batches of experiments performed *en masse* can be imported from spreadsheet files.

2.2.3 Crystallization Module

The crystallization module gathers data about crystallization trials: the setup of crystallization plates, the contents of individual wells and drops, the origin of harvested crystals, and any special conditions used during crystal harvesting (Fig. 5). The interface can be used to assign a crystallographic screen to the wells and drops of a crystallization plate. Most commercial screens are predefined and new commercial as well as custom crystallization screens are easily generated or can be downloaded from the Formulatrix web page [34]. Plate templates can be created with up to six different drops per well, making it possible to examine several parameters within each chamber (well) of the crystallization plate. Each drop of a chamber can contain a different macromolecular prep of a project, and the volumes of the macroprep and the screen used for each drop can be different. Each drop will have an associated screen, which does not have to be the same as the screen in the reservoir, unlike in the traditional style of crystallization plate. Thus, LabDB makes it possible to track alternative reservoir screening [35]. For example, sodium chloride can be in the reservoirs while different crystallization screens can be used for each drop position. This type

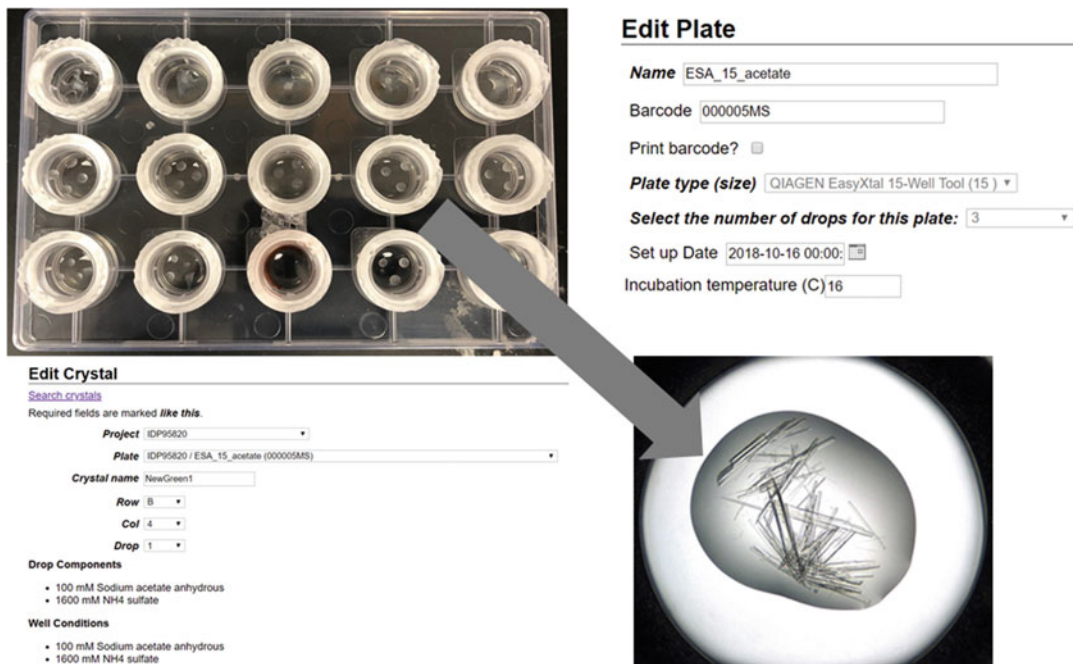


Fig. 5 Tracking of plate and crystal data in LabDB. Clockwise from left: 15-well plate, information about the plate, crystals from well B4 in the plate, and information about the crystal. Only a small portion of each page from LabDB is shown

of screening can make initial crystallization trials very efficient and has several other benefits [36].

The module can be configured to gather data directly from the Minstrel HT (Rigaku) and Rock Imager (Formulatrix) observation robots (data transferred periodically) or used in manual mode. The module can also interact with the *Xtaldb* application [28]: an expert system that provides tools to design crystallization plates, record drop images and annotations, and analyze the results.

2.2.4 Structure Determination Component

Information about structure determination, refinement, and validation are stored using the *hkldb* module, which is shared with the HKL-3000 suite [26]. *hkldb* allows a bi-directional transfer of data between HKL-3000 and LabDB. Crystals in LabDB can be selected for diffraction, data reduction, structure solution, and model refinement. The availability of complete information about sample production during any step of structure determination can be extremely helpful for the estimation of radiation decay or attempts to identify unassigned electron density. A report containing statistics of data collection and the refinement process can be generated (Fig. 6).

Data collection and refinement statistics for project **DJ-1** crystal **Ib-33**

Project description: DJ-1

sca /med/data2/diffraction/protein/2011-07-06-aps-21id-f/DJ-1/Ib-33/proc_2018/DJ-1_Ib-33_final.sca
 model /med/data2/diffraction/protein/2011-07-06-aps-21id-f/DJ-1/Ib-33/proc_2018/structure_mr/build_model_7/hkl_refine_45.pdb_tls

Data collection	
Resolution (Å)	50.00 - 1.83 (1.86 - 1.83)
Wavelength (Å)	0.97872
Space group	P6522
a, b, c (Å)	66.50, 66.50, 176.73
α, β, γ (°)	90, 90, 120
Completeness (%)	97.8 (79.3)
Observed reflections	423474
Unique reflections	20934
$\langle I \rangle / \langle \text{Sigma} I \rangle$	59.5 (2.2)
CC1/2 last shell	0.59
Redundancy	20.2 (15.8)
Rmerge	0.056 (1.258)
Wilson B factor (Å ²)	19.4
Refinement	
Rwork / Rfree	0.170 / 0.204
Bond lengths msd (Å)	0.007
Bond angles msd (°)	1.2
Mean B value (Å ²)	29
Number of protein atoms	1383




Fig. 6 An example of X-ray data collection and refinement statistics report produced by HKL-3000 using *hkldb*. This screenshot displays about half of the report

2.2.5 Biochemical Experiments

The value of structural data is magnified when the data are coupled with functional experiments such as enzymatic assays, binding studies, etc. LabDB can accommodate a variety of biochemical and biophysical experiments to ensure that all available experimental evidence is associated with projects. LabDB is not a substitute for external analysis programs that help interpret data coming from different instruments but rather is a mechanism that ensures the data is easily retrievable and accessible.

LabDB can import information about absorbance- and fluorescence-based enzyme kinetic assays. The system tracks both detailed studies of a particular enzyme-substrate reaction (e.g., Michaelis-Menten kinetics) and high-throughput screening plates involving many substrates or many enzymes. Individual experimental replicates can be recorded, but calculations of kinetic constants, such as K_M and k_{cat} , need to be performed outside of LabDB. The first step of entering assay experiments is to define the experimental protocol. The description should include a detailed description of the steps, instruments, and buffer composition. These protocols can be used for multiple projects, so defining an individual “kinetic assay” involves specifying the protein being used (the macromolecule prep) and the substrate being tested.

The alternate method of entering kinetic data allows multiple assays to be submitted at one time, varying either the enzyme or the substrate. This facilitates searching for a protein with a particular function or looking for the optimal substrate, respectively, and is compatible with many 96-well plate readers that can export data. Data can either be uploaded as a CSV file or manually entered in a web form. Compositions of experimental layouts can be saved as a plate design for reuse.

Thermal Shift Assays

Fluorescence-based thermal shift assays (FBTSAs) are based on the detection of fluorescence from a dye present in solution. As proteins in the solution denature in a heat dependent manner, the dye binds to exposed hydrophobic surfaces, which increases the dye's fluorescence. LabDB processes the results of FBSTA experiments by parsing the output generated by Bio-Rad CFX96 and Applied Biosystems 7900HT RT-PCR systems, allowing the melting curves to be visualized within the web interface. Three files can be uploaded for each experiment: the raw relative fluorescence units, the derivatives of the melting curves, and an experimental summary. This not only allows the data to be visualized within LabDB but allows the raw files to be accessible from anywhere. Individual curves are displayed below the aggregate figure (Fig. 7).

Isothermal Titration Calorimetry

Interactions of macromolecules with small molecules or other macromolecules can be detected using isothermal titration calorimetry (ITC). In an ITC experiment, small aliquots of a ligand or a macromolecule are injected into a sample cell, and the heat change caused by interaction is measured. ITC has become a staple of biochemical labs due to its ease of use and its ability to accurately characterize the binding affinity and stoichiometry of interactions. LabDB tracks the macropreps and solutions in the experimental cell and the injection syringe. The file generated from a MicroCal ITC system as well as an optional analysis file from Origin data analysis software can be uploaded and used to visualize the results.

Other Biochemical Assays

As the variety of experimental methods used by biomedical labs is broad, the LabDB system also permits entry of basic data about custom experimental assays. Each trial has to be coupled with a protocol that describes the method in more detail. The experiments can be flagged as to whether or not they are successful, and a list of people involved can be included. This entry is relatively generic and includes a text field with optional comments or notes, which allows the researcher to use a standard protocol, yet document any deviations or experimental conditions that are not explicitly described in the protocol.

Plate PA3944_2

[Return to plate layout](#)

Project	IDP52183 - CSGID
Macroprep	1 mg/ml
Chemical plate	PA3944_2
Person	Czub, Mateusz
Comment	A1-5 CoA 3 mM 6-10 CTRC 10 mM 11-12 blank (Tris pH 7.5 150 mM) B1-5 puromycin 12 mM 6-10 Aspartame 12 mM C1-5 Colistin 12 mM 6-10 a-D-glucosamine 12 mM D1-5 acetyl-D-glucosamine 12 mM 6-10 chloramphenicol 12 mM E1-5 glycine 12 mM 6-10 glycyglycine 12mM F1-5-glyglygly 12 mM G1-6 phenylglyoxal 6 mM

All curves combined

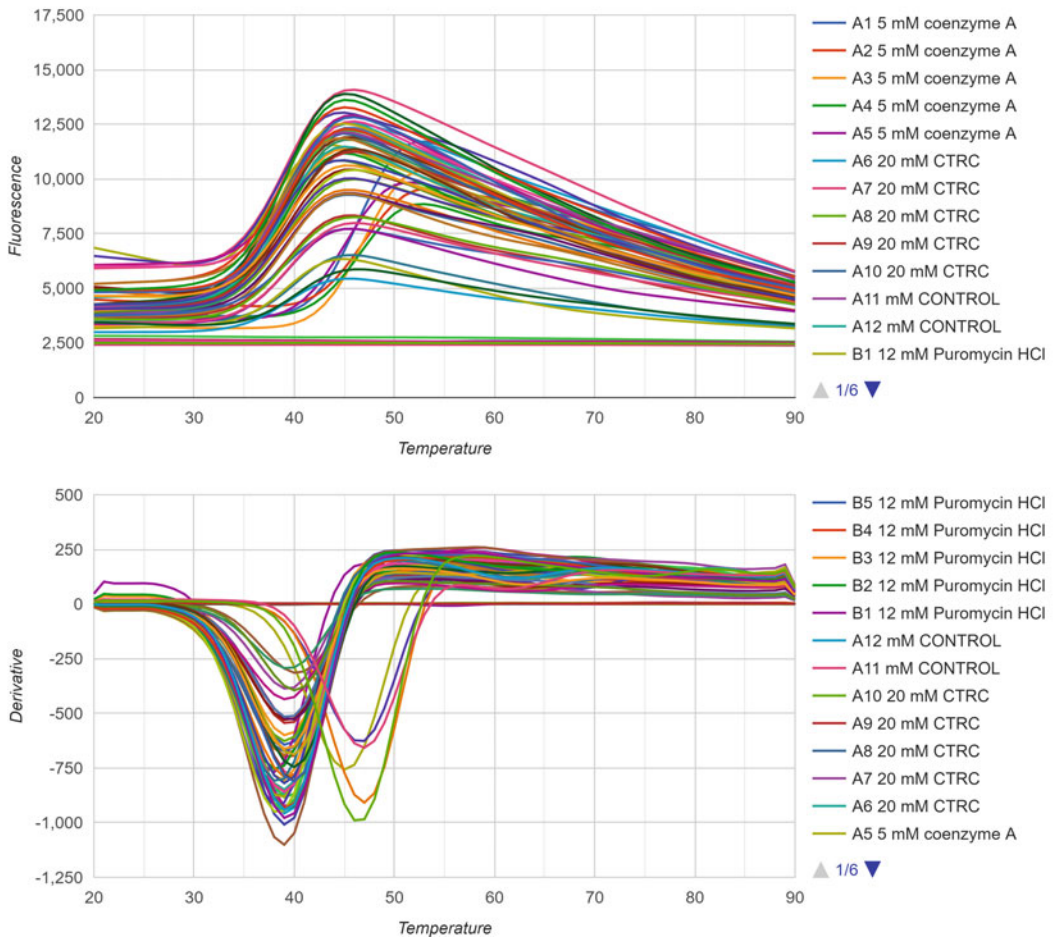


Fig. 7 A sample thermal shift assay with raw data obtained from the Bio-Rad CFX96 system and a derivative graph

2.3 Data Validation

Data integrity needs to be considered and verified at every step of the data management process. For example, constraints should be created during the database or web framework design to ensure that relationships between pieces of data are legitimate and that unique attributes (or combinations of attributes) are enforced. This will prevent entering data that are internally inconsistent or inconsistent with the data that are already in the database. For example, entering two plates with the same name or attempting to put a 96-well screen into a 24-well plate is not possible. Similarly, attempts to enter incomplete data are prevented by several mechanisms, including alerting the scientist responsible for specific instruments when incomplete data are automatically transferred from an instrument. There are several checks of data consistency. For example, plates cannot be transferred from the crystal storage system if the project is not already in the database because every plate must be associated with a project.

In contrast to experimental data, the integrity of computational data is generally not a challenge and is beyond the scope of this chapter.

3 Technical Implementation

Due to the collaborative nature of modern research, most existing LIMSs use one fundamental design—a central database and one or more clients. The clients are usually one of two kinds, either a web-based or desktop application, although some LIMSs use both to leverage their distinct capabilities.

Web-based applications are usually suitable for general user interfaces for data presentation and manual data input. They provide platform independent interfaces that can be accessed by any computer connected to the internet (or an internal network) using a standard browser. There is no need to develop and test the application on all possible operating system versions and configurations, which makes development and troubleshooting much more straightforward. Maintenance is easier, as the updates applied on the server are immediately accessible on the client computers. The interfaces can be adapted to portable devices such as tablets and smartphones, providing a consistent user experience across devices suitable for laboratory use. Mobile devices used at the bench provide the advantage that data can be input at the same time experiments are performed. However, web interfaces are limited in complexity by the constraints of the web programming environment, and it may be difficult to communicate with hardware attached to client machines. Nonetheless, rapidly evolving JavaScript technologies allow for more and richer interaction with web browsers, changing them from simple document viewers into fully featured extendible development platforms.

Desktop applications are suitable for clients that require communication with the scientific instruments and are often used for processing and analyzing substantial amounts of data. The development time and effort for such systems can be longer, but the resulting interfaces can be somewhat more sophisticated and have extensive communication with external hardware, such as direct control of the instrument or automated data gathering and processing. The major drawback is that desktop applications require development and maintenance of the standalone programs for several popular operating systems (e.g., Windows, macOS, Linux). Such programs are generally more difficult to install and to keep up to date since upgrades must be distributed to each client computer. Several Java-based clients have solved the problem of application distribution by distributing the code through a web browser but at the expense of requiring the user to maintain an up-to-date Java environment for each browser. In addition, the Java NPAPI plugin is no longer supported in most modern browsers due to security concerns. SESAME [17] is one LIMS that uses these types of applications as a general client. Recent advances in web technologies, such as cloud computing, software as a service, and semantic technologies, support the creation of sophisticated distributed systems. These advancements further blur the boundaries between different clients, most of which now serve only as a user interface, while all data is processed and stored in the cloud.

The current implementation of LabDB mainly uses the web-based approach. The web framework to generate the pages presented to the user incorporates the model-view-component (MVC) architecture (the current implementation uses CakePHP [37]). In this framework, individual tables in the database are represented by models, which are related to other models using associations that denote the relationship. For example, a purified protein may have many crystallization plates, but a particular crystallization drop “belongs to” only one crystallization plate. The CakePHP framework has been supplemented with jQuery and JavaScript, permitting forms or pages to be altered based on other choices in the forms. For example, when entering a crystallization plate, selecting the project will fetch the appropriate list of “macromolecular preps” of that protein. In many cases, forms will have a header column, which populates a whole collection of other form entries.

All four modules store the data they use in a central PostgreSQL database, and as a result, share common organizational information, such as projects, laboratories, user accounts, passwords, and identification barcodes. Information collected by one module is accessible within the others. For example, all purified proteins in the system are available to the crystallization module and can be used to prepare crystallization plate records. The overall system database is very large, containing about 250 tables and 30 views separated into a distinct schema for each component.

Each laboratory can have a separate instance of LabDB with its own database. All instances share the same schema but have different data. Some of the functionality of LabDB is provided by customized scripts that are run based on a schedule, which is controlled by the scheduling daemon cron. For example, a weekly progress report is emailed to the principal investigator (PI) once a week, and the external database for a crystallization “hotel” is queried every 30 min to keep the plates in each system synchronized. Most of the systems run by a daemon can also be triggered by a button within the interface, providing a means of generating instantaneous reports.

4 Data Analysis and Data Mining Tools

Each component provides a set of data analysis tools. LabDB provides several data mining tools to analyze the results of structure determination and biological assays. Virtually all types of data have detailed search tools. The basic search tool allows the user to search for a particular project name, description, responsible person, project status, etc. Each type of object (i.e., crystals, clones, thermal shift assays, etc.) will have object-specific search fields. For example, crystals can be filtered based on whether or not they have been tested for diffraction and kinetic assays can be filtered based on the specific protocol used. Applying a filter will return a paginated table containing the reduced set of objects. Each object will have a default set of columns that are displayed, but there is a “Select Columns” button that allows more (or fewer) columns to be displayed in the resulting list. All of the displayed columns can be used to sort the data.

In addition, there is a number of data analysis “dashboards” that provide a real-time overview of the status of the project in a pipeline. LabDB can display statistics for projects or researchers that summarize progress. The progress summary allows a LabDB user to specify a time period during which the experiments happened in the lab and has several predefined periods (e.g., for the past 1, 2, or 3 weeks, 1, 2, or 6 months, or 1 or 2 years). The result (Fig. 8) is a table reporting how many of each type of experiments were performed during the time period. Most table entries are links that will bring up the list of experiments associated with that number. LabDB also builds aggregate statistical reports for different groups of projects, which can be used to group projects that may be supported using various funding sources, similar projects that involve the same collaborator, or projects of related proteins.

In addition to being able to have progress summaries displayed within the interface, the weekly per person and project statistics are emailed to the lab members and the PI every week. This feature is not to check the performance of particular people but rather to

Statistics / Progress in Minor Lab LIMS by researcher

Last week

Person	Clones	Exprs	Purifs	Macro preps	Plates	Drops	Crystals	Datasets processed	Structure refs	Kinetic assays	Thermal shift assays	ITC
Cooper, David	0	0	0	0	0	0	63	0	0	0	0	0
Czub, Mateusz	0	0	0	0	0	0	13	7	1	0	0	0
Lipowska, Joanna	0	0	1	1	2	333	74	3	0	0	0	0
Shabalín, Ivan	0	0	0	0	0	0	54	7	7	0	0	0
Siuda, Monika	0	0	0	0	0	0	30	1	1	0	0	0
Steen, Ethan	0	0	0	0	1	288	0	0	0	0	0	0
Venkataramany, Barat	0	0	0	0	19	5199	0	0	0	0	0	0

Last month

Person	Clones	Exprs	Purifs	Macro preps	Plates	Drops	Crystals	Datasets processed	Structure refs	Kinetic assays	Thermal shift assays	ITC
Cooper, David	0	0	0	0	0	0	72	0	0	0	0	0
Czub, Mateusz	0	0	0	1	15	2202	13	7	2	0	0	0
Lipowska, Joanna	0	0	3	2	12	3213	74	3	1	3	0	0
Majorek, Karolina	0	0	0	0	0	0	0	0	2	0	0	0
Miks, Dylan	0	0	0	0	10	2880	0	0	0	0	0	0
Shabalín, Ivan	0	0	0	0	0	0	54	7	8	0	0	0
Siuda, Monika	0	0	0	0	4	960	30	1	1	0	0	0
Steen, Ethan	0	0	1	9	4	1152	0	0	0	0	0	0
Venkataramany, Barat	0	0	0	0	31	7281	0	0	0	0	0	0

See full statistics and details at the [Minor Lab LabDB page](#).

Fig. 8 An example of a weekly report automatically sent by LabDB to the PI and all current lab members

identify bottlenecks in the research and to identify experimental steps that need to be addressed.

Another major advantage of LIMS is the potential for quantitative data analysis. In our experience, both crystallization and cryoprotection protocols were significantly improved after simple analyses showed which approaches are more productive. Several researchers in the lab switched to the use of certain crystallization screens after large-scale analysis of the lab experiments tracked in LabDB showed that these screens were significantly more efficient than others for producing harvestable crystals. As another example, some of the projects were conducted with the “alternative reservoir” crystallization method [35]. Analysis of our crystallization trials showed that this approach produced more crystals per plate than the traditional approach. Switching to the new crystallization method significantly increased the lab’s productivity afterward. Similarly, an analysis of diffraction resolution vs. cryoprotectant used helped determine the best cryoprotectants for several projects.

Rigorous use of the database during experiments helps in preparation of publications, especially when the experimenter has left the laboratory, a frequent case in academia. Projects carried out using LabDB (all research centers) resulted in 156 publications within the last 7 years. Of those, six achieved a relative citation ratio (RCR) higher than 5, and two were classified as highly cited papers in the Essential Science Indicators database (i.e., they were in the top 1% of papers by field and publication year, according to

Table 1
Summary data from the Minor Lab instance of LabDB for projects carried out for three research centers and internal projects as of May 15, 2019

	CSGID	MCSG	NYSGRC	Minor Lab
Projects	162	129	118	796
Clones	85 (1.5)	18 (2.3)	16 (2.0)	139 (3.3)
Expressions	289 (3.4)	134 (5.6)	79 (3.8)	293 (5.7)
Purifications	390 (4.8)	121 (4.7)	81 (3.7)	326 (4.3)
Macropreps	857 (5.9)	384 (3.1)	398 (3.4)	1312 (1.7)
Plates	2626 (22.8)	1263 (30.8)	1001 (10.2)	2888 (23.9)
Crystallization Drops	415,630 (3614.2)	71,569 (1745.6)	208,878 (2131.4)	241,596 (1996.7)
Crystals	6844 (59.5)	2258 (19.5)	2370 (26.6)	6211 (8.5)
Diffraction Datasets	2193 (31.3)	1230 (12.1)	742 (13.3)	2952 (7.1)
Refinement Runs	16,590 (276.5)	6285 (133.7)	7139 (158.6)	9002 (77.6)

Results are given in the format: total number (average per project). The average number is calculated using the total number of experiments for the stage divided by the number of projects that had at least one experiment for that stage

Web of Science). The summary of the total amount of data stored in the Minor Lab instance of LabDB is available in Table 1.

5 User Experience and PI Perspective

To make the user experience less tedious and more intuitive, input in LabDB is designed to be easy, logical, and straightforward. This can be partially attributed to the involvement of people who perform experiments in the design of the interface. In particular, most relevant fields are populated based on previous experiments and information about the project. For example, when adding a new crystallization plate to the database, the experimenter can select both the type of plate and the crystallization screen used from a drop-down menu. The addition of crystals is greatly simplified with the “clone” button, which allows the crystal description with all associated information (crystal size, morphology, cryoprotectant used, source plate and well, crystallization condition) to be copied in one click, thereby requiring the experimenter to only change the name of the new crystal and adjust other details as necessary.

LabDB has incorporated many other features that add immediate convenience for its users. For example, the chemical storage module quickly indicates the availability and location of chemical bottles, which often saves researchers time when looking for a rarely used chemical or searching for what chemicals are currently available. Another feature is the ability to enter “free text” notes in some

fields. For example, the experimenter can make notes on cryoprotection and describe the crystal specifics (e.g., color, size, morphology, and added ligand) to keep a more detailed record of crystal harvesting. These text fields are searchable, allowing for later searches for crystals with the custom keyword or phrase entered by the experimenter. The use of these text fields makes LabDB more flexible but is at the expense of making data mining more difficult.

One of the major factors limiting the adoption of LIMSs in academic research is sociological in nature. Often, lab scientists are unenthusiastic to use LIMSs because they perceive inputting data as quite tedious and LIMSs to be less flexible and convenient compared to a lab notebook, while the benefits to the researcher performing the experiments are seen as minimal. A potential explanation of the low perceived benefit is that the experimenters tend to hope for “the best-case scenario” (no need to troubleshoot the experimental results, the experiment is published soon after it is performed, allowing the researcher to draw unrecorded details from the memory, etc.) and overlook the long-term benefits of LIMS. As a result, although many LIMSs are in principle available for usage in the laboratory, the often-tedious and inflexible input and minimal perceived benefits to the researcher performing experiments deter researchers from using them.

The extra effort that is occasionally required to enter information into a LIMS is sometimes resisted by researchers who cannot see the benefits of this effort. Indeed, even in some structural genomics laboratories that were required by the NIH to ensure all data was publicly available, some researchers have made comments such as “I don’t know why we even need a database,” and some users thought their shorthand description of crystallization plates was sufficient because “the code is scribbled on the wall over the microscope.” Tragically, the wall was painted when the experimenter was out of town. This type of complacency results from the shortsightedness of researchers who think that they are the only ones who will ever have to interpret their results. The reality is that most research projects, even if conducted in a single lab, rely on numerous researchers who come and go, often leaving the project’s PI with the difficult task of trying to locate notes and decrypt user’s codes and shorthand. Sometimes these notebooks get misplaced or irretrievably lost. The attitude that can defeat accurate data preservation is difficult to overcome and sometimes comes down to strict enforcement of data entry by the PI. LabDB’s weekly reports to the PI make it possible for the PI to see that the data has been preserved, and the PI can easily browse details and results of experiments. Most of our laboratory members come to appreciate the extra effort at some point, especially when writing papers about long-standing projects or performing experiments similar to ones performed years ago by researchers who have left the lab.

Another major factor that demotivates the adoption of LIMSs is the perception that they are difficult to install and maintain. PIs in smaller labs with few active experimental projects may perceive that there are few or no benefits to be gained by installing and using a LIMS and that any potential gains do not justify the time, effort, and money required for setup and maintenance. Labs of small to medium size may contend that usage of a laboratory notebook to record all experiments is much more feasible than attempting to supplant the notebook with software or hardware solutions. Although implementing a LIMS may be a challenge for smaller labs, a PI who does cutting-edge research may be asked for details of his/her experiments and face the issue of irreproducibility, which is not easy to handle without detailed records of experiments. Laboratories that perform very diverse or uncommon types of experiments and that do not find a LIMS that covers the breadth of techniques used may employ an ELN instead of a LIMS; while not perfect, this may be the only reasonable solution in such situations.

A viable solution for laboratories that do not want to install a LIMS is to utilize a LIMS that is provided as a Software-as-a-Service (SaaS), which avoids the complexity and cost of setting up and maintaining a secure server with database and web server capabilities. Many SaaS solutions (e.g., QBench, CloudLIMS, and webLIMS) are web-based, which is similar to the interface provided by LabDB. SaaS approaches provide all users with the latest version of the application because the code is maintained by the service provider. Typical SaaS applications will store a user's data either on their service or in the cloud, but it is possible to store data at the user's site. It is critical that potential SaaS users investigate the capabilities of exporting their data if they decide to discontinue the service.

6 Enhancing Reproducibility and Efficiency of Experiments: Case Studies

Management of experimental data is a critical factor in establishing a reproducible research workflow. LIMSs are especially helpful in ensuring continuity and reproducibility for large projects that involve multiple researchers and last for many years. In our experience, LabDB has proven itself to be much more durable than a paper notebook or series of spreadsheets. One example of a long-term project in our lab is the "albumin project," which aims to characterize interactions between albumins from various species and small molecules transported in the blood. Since the start of the project in 2008, several researchers, ranging in expertise from undergraduate students to research faculty, have participated in this project and performed numerous protein purifications and crystallization trials, which has led us to a collection of diffraction images

Filtered by: Project.name matches "IDP95820" AND Date is after 2010-03-10

Page 1 of 89, records 1-20 out of 1774 total

Project	Harvested by	Harvest date	Well	Drop	Drop solutions	Drop macropreps	Cryo. notes	Beam	Resol (Å)	Diffract. notes
IDP95820	Mateusz Czub	2017 Mar 06 - 01:51 PM	A2	1	1uL 1.9M Ammonium Sulfate, 0.2M Li sulfate, 0.1M Tris pH=7.4	1uL 35.42mg/mL IDP95820 Native ESA	Paratone + ZD method	21IDG	1.75	Amazing albumin with warfarin
IDP95820	Katarzyna Handing	2014 Nov 23 - 05:12 PM	A3	6	1uL 0.1Molal Tris, 2.4M NH4 phosphate pH=8.5	2uL 30mg/mL IDP95820 30_zn_opt	Paratone	21IDG	1.9	nice spots, paratone good cryo, ano sig 0.003
IDP95820	Mateusz Czub	2018 Jul 24 - 04:26 PM	C1	1	1uL 0.1M Sodium acetate anhydrous, 1.6M Ammonium Sulfate pH=4.6	1uL 37.6mg/mL IDP95820 ESA_37.6	paratone	19ID	2	dataset collected - really nice data
IDP95820	Ivan Shabalin	2015 Apr 19 - 07:36 PM	B3	1	1uL 0.1Molal Tris, 2.4M NH4 phosphate pH=9	1uL 34mg/mL IDP95820 ESA_34	paratone	21IDG	2.1	mos 0.2- 0.3
IDP95820	Katarzyna Handing	2016 Apr 16 - 01:09 PM	C3	1	1uL 2.2M Ammonium Sulfate, 0.2M Li sulfate, 0.1M Tris pH=7.4	1uL 34mg/mL IDP95820 ESA_34	paraton	23IDD	2.1	awesome helical dataset
IDP95820	Ivan Shabalin	2015 Apr 19 - 07:06 PM	A5	1	1uL 2M Ammonium Sulfate, 0.2M Li sulfate, 0.1M Tris HCl pH=6.5	1uL 34mg/mL IDP95820 ESA_34	paratone	21IDG	2.15	Mos 0.3; 1st dataset 2.34 Å. 2nd dataset fried 7 sec. 2.2 Å, 50 frames is enough.

Fig. 9 LabDB view of a list of crystals with selected experimental details for a particular project

for more than 1500 crystals (Fig. 9). LabDB has enabled the storage of all information about that data and experimental setups that is easily accessible by any lab member. Access to old but complete information/data has allowed projects to be completed using the most modern software and a state-of-the-art approach [38]. One such example comes from our recent structure of equine serum albumin (ESA) in complex with testosterone (PDB ID: 6MDQ) [39]. Crystals of this complex were obtained in 2011, but the project stalled for various reasons. The deposition of this structure in 2018 was possible because the details of the crystallization procedure were critical in the structure refinement process. Additional studies performed in 2018 led to publication in 2019 [39]. LabDB has allowed us to keep accurate records of experimental details over the course of the albumin project. Our ability to reproduce these experiments has allowed us to deposit 14 albumin structures in the PDB so far and publish five papers, one of which has garnered almost 300 citations [40].

Another case of LIMSS ensuring continuity and reproducibility of experiments stems from their capability to track chemicals and protein batches used in the experiment. Researchers are generally aware that variations among chemical batches (e.g., intended or unintended changes in the manufacturing process) may result in different outcomes of biomedical experiments [41]. In addition, for most projects, it is very important to use one purification protocol for all experiments to ensure that the protein was purified or modified in exactly the same way. A powerful example of such a

project is the “Gcn5-related N-acetyltransferases’ (GNAT) project,” during which we discovered and clearly demonstrated that buffers used during purification and the presence of 6×His-tag alter enzyme kinetics and cause discrepancies between findings based on a crystal structure and results of kinetic or binding studies [42]. Therefore, keeping track of all chemicals and their batches used in protein production and experiments thereafter, as well as other details such as the removal of the 6×His-tag, is crucial for ensuring reproducibility of these experiments. Using a LIMS to track experiments makes the task of keeping such records manageable. With the use of a LIMS, an experimenter can compare all chemical batches and procedures used in experiments and identify differences that may be the cause of irreproducibility.

7 Future Directions: Toward a Configurable LIMS Architecture

The requirements that shaped LabDB’s functionality were changing dynamically during the almost 15 years of its development. It comes as no surprise that in order to keep a software suite cutting-edge, it needs to be constantly maintained and extended with new functionalities. The dynamic development of scientific methodologies and software technologies are major but not the only limiting factors that affect the usability of a LIMS. Various laboratories have different data management needs that tend to change dramatically over time. Our ambition was to convert LabDB from a macromolecular crystallography LIMS into a versatile suite that would be flexible, customizable, and extendable and would allow for a user-driven evolution of database schema over time. For this purpose, we have made an effort to redesign the system architecture and simplify the underlying data model.

7.1 Data Model

The current implementation of LabDB was based on a relational database model, which enforces a strictly organized way of storing data and provides powerful query language but at the same time is difficult to change. The relational database schema imposes a data structure is defined up-front during system development and cannot be updated without changes in the source code. Another large issue is an object-relational impedance mismatch, i.e., set of difficulties happening when a relational database is served by an application program written in an object-oriented language.

An alternative for the relational model is NoSQL document databases, which do not require a predefined schema. To address those issues, we have designed a hybrid data model based on the PostgreSQL relational database engine and utilized this engine’s support for storage of JSON (JavaScript Object Notation) documents. This database structure is capable of storing experimental workflows represented as directed acyclic graphs. The graph nodes

are called “elements” and represent any physical or conceptual entity (e.g., chemical, sample, result) that is a subject or result of physical actions performed in the lab. The laboratory actions (e.g., experiments, analyses, shipments), called “processes,” are edges in the graphs. The complete workflows can be efficiently retraced using a recursive SQL query on a single table storing the graph’s adjacency lists.

Element and process records have only a few generic attributes, one of which is a JSON-type object that embeds all data specific to the particular entity. This structure allows any element and process data, structured with individual fields and structures, to be stored in the database. In contrast to NoSQL databases, the structure of objects is not completely schema-less. In our model, every element, process, and workflow object must hold a reference to a JSON Schema object defining the format of the data. The schema file serves as a consumer contract, i.e., it is applied to the incoming data to determine if the data conforms to the schema’s definition. This hybrid approach combines the consistency of relational databases with the flexibility of JSON data structures.

7.2 System Architecture

The original LabDB is a server-side web application based on the model-view-controller (MVC) architectural pattern and CakePHP framework. In this classic “thin client” design, all pages are generated by the server-side code and transferred as complete HTML documents to the browser. Over the past few years, the trends in web development have shifted to browser-based client functionalities. Such an approach gives more implementation flexibility, assures the ability to work in an offline mode, and lowers server requirements and infrastructure costs. The future architecture of LabDB will be based on the Representational State Transfer (REST) web API and independent JavaScript client applications. The prototype of the new LabDB API was implemented in a Python web framework, Django, with the use of the Django REST toolkit. The API decouples data storage from the client application, simplifying data sharing between different application programs. Thanks to the API, development of new front-end tools that access the data will not require changes within the LIMS itself. Additionally, the Django framework has a vertically split structure, which allows for the encapsulation of functionalities within so-called reusable apps. This gives the possibility for easy integration of LabDB’s API with existing scientific applications written in Django. The database abstraction layer was defined using Django’s object-relational mapping (ORM) module, providing a clean separation of concerns and easy refactoring possibilities.

7.3 Workflow Management

The main motivation for the redesign of LabDB was to make it adjustable to different workflows and the lab’s changing needs. In LabDB users would be able to define custom workflows using a business process graphical notation standard, called BPMN

(Business Process Modeling Notation) 2.0. The BPMN is a flow chart method that does not require any programming knowledge, thus bridging the gap between process intention and implementation. The BPMN flow charts are defined through the graphical editor using a set of predefined graphical elements that simplify business activities, flow, and processes. LabDB user would be able to predefine process steps, required reagents and samples, expected results, and assign people and instruments. The application will take care of translating the graphical workflow into a set of predicate schema definitions for incoming data. We believe that BPMN modeling can greatly improve the repeatability and reproducibility of biological workflows.

8 Conclusions

The ultimate bottleneck of modern biomedical research is the insufficient rate of conversion of vast experimental data into biomedical information. As we have argued in this chapter, using a well-designed LIMS for managing experimental data offers a number of benefits to a modern biomedical research lab.

1. A LIMS is the most convenient way of keeping track of an inventory of laboratory chemicals and specimens.
2. Using LIMSs helps to assure continuity of projects, as they create a persistent record of the performed experiments, which can be examined by researchers working in the lab regardless of whether the person who did the experiment continues to work there or not.
3. Supplemented by data mining tools, LIMSs may be used for optimizing experimental methods and protocols by identifying, e.g., the bottlenecks in the workflow, the best methods for conducting particular types of experiments, the optimal parameters for protocols, etc.
4. For the PIs, LIMSs provide a way of tracking lab activity across different projects, as well as the progress of individual projects and identification of factors that impede progress.
5. LIMSs can help researchers diagnose issues affecting the reproducibility of the experiments. As anybody who has ever worked in a biomedical laboratory knows well, repeating a past experiment does not always yield the same results. Even if the researcher does not make any errors in setting up an experiment, the result of this experiment may be affected by multiple factors, such as temperature, a different batch of reagents, etc., which are sometimes beyond the experimenter's control. Recording as many experimental details as possible in a LIMS may be helpful to identify what is different between the original experiment and its unsuccessful repetition. As was recently

stated in a Retraction Watch discussion, “in many cases, that act of looking something up in a database is enough to reveal a problem” [43].

6. Last but not least, the use of LIMSs as tools to easily share data with collaborators and the wider research community helps to facilitate open science. Unfortunately, modern data management systems will not be any better than the laboratory notebook if their data remains “siloes”—isolated, removed from the context of all other relevant data even when they located on a local hard disk or in the cloud. Some researchers are taking advantage of general-purpose repositories to upload their data into the “cloud” to ensure that it is not lost forever. Unfortunately, unindexed data without sufficient description (metadata) are impossible to locate and are as useful as the information that a lost diamond necklace is somewhere in the landfill.

Acknowledgments

We thank all the users of our data management programs who over many years provided us with numerous complaints, suggestions, and requests that gave us invaluable feedback to improve our tools. This work was supported by the National Institute of General Medical Sciences under Grants GM117080 and GM117325, National Institutes of Health BD2K program under grant HG008424, and the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under Contract No. HHSN272201700060C and HHSN272201200026C.

Disclosure statement: One of the authors (W.M.) notes that he has also been involved in the development of state-of-the-art software and data management and mining tools; some of them were commercialized by HKL Research, Inc. and are mentioned in the paper. W.M. is the co-founder of HKL Research, Inc. and a member of the board.

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

References

1. Data management. <http://www.businessdictionary.com/definition/data-management.html>. Accessed 6 May 2019
2. Freedman LP, Cockburn IM, Simcoe TS (2015) The economics of reproducibility in preclinical research. *PLoS Biol* 13(6): e1002165
3. Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 10(9):712–7c1

4. Begley CG, Ioannidis JP (2015) Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res* 116(1):116–126
5. Collins FS, Tabak LA (2014) Policy: NIH plans to enhance reproducibility. *Nature* 505(7485):612–613
6. McDowall RD, Pearce JC, Murkitt GS (1988) Laboratory information management systems—Part I. Concepts. *J Pharm Biomed Anal* 6(4):339–359
7. Hakkinen J, Levander F (2011) Laboratory data and sample management for proteomics. *Methods Mol Biol* 696:79–92
8. Hunter A, Dayalan S, De Souza D, Power B, Lorrimar R, Szabo T et al (2017) MASTR-MS: a web-based collaborative laboratory information management system (LIMS) for metabolomics. *Metabolomics* 13(2):14016–1142-2. Epub 2016 Dec 27
9. Lin K, Kools H, de Groot PJ, Gavai AK, Basnet RK, Cheng F et al (2011) MADMAX - management and analysis database for multiple -omics experiments. *J Integr Bioinform* 8(2):160;jib-2011-160
10. Stephan C, Kohl M, Turewicz M, Podwojski K, Meyer HE, Eisenacher M (2010) Using Laboratory Information Management Systems as central part of a proteomics data workflow. *Proteomics* 10(6):1230–1249
11. Venco F, Vaskin Y, Ceol A, Muller H (2014) SMITH: a LIMS for handling next-generation sequencing workflows. *BMC Bioinformatics* 15(Suppl 14):S3. Epub 2014 Nov 27
12. Harris M, Jones TA (2002) Xtrack - a web-based crystallographic notebook. *Acta Crystallogr D Biol Crystallogr* 58(Pt 10 Pt 2):1889–1891
13. Lab Information Management Systems (LIMS). <https://www.thermofisher.com/us/en/home/life-science/lab-data-management-analysis-software/enterprise-level-lab-informatics/lab-information-management-systems-lims.html>. Accessed 25 Apr 2019
14. Laboratory Information Management System (LIMS). <https://www.autoscribeinformatics.com/lims-laboratory-information-management-system>. Accessed 6 May 2019
15. Produce reliable results more quickly. <https://www.illumina.com/informatics/sample-experiment-management/lims.html>. Accessed 25 Apr 2019
16. St. Cyr K, Hill A, Warren P, Mounts D, Whitley M, Mounts W et al (2010) From project-to-peptides: customizing a commercial LIMS for LC-MS proteomics. *J Biomol Tech* 21(3):S9
17. Zolnai Z, Lee PT, Li J, Chapman MR, Newman CS, Phillips GN Jr et al (2003) Project management system for structural and functional proteomics: SESAME. *J Struct Funct Genom* 4(1):11–23
18. Morris C (2015) PiMS: a data management system for structural proteomics. *Methods Mol Biol* 1261:21–34
19. Daniel E, Lin B, Diprose JM, Griffiths SL, Morris C, Berry IM et al (2011) xtalPiMS: a PiMS-based web application for the management and monitoring of crystallization trials. *J Struct Biol* 175(2):230–235
20. Prilusky J, Oueillet E, Ulryck N, Pajon A, Bernauer J, Krimm I et al (2005) HalX: an open-source LIMS (Laboratory Information Management System) for small- to large-scale laboratories. *Acta Crystallogr D Biol Crystallogr* 61(Pt 6):671–678
21. Bonanno JB, Almo SC, Bresnick A, Chance MR, Fiser A, Swaminathan S et al (2005) New York-Structural GenomiX Research Consortium (NYSGXRC): a large scale center for the protein structure initiative. *J Struct Funct Genom* 6(2–3):225–232
22. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR et al (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D* 67(Pt 4):235–242
23. Potterton L, Agirre J, Ballard C, Cowtan K, Dodson E, Evans PR et al (2018) CCP4i2: the new graphical user interface to the CCP4 program suite. *Acta Crystallogr D Struct Biol* 74(Pt 2):68–84
24. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N et al (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D* 66(Pt 2):213–221
25. Echols N, Grosse-Kunstleve RW, Afonine PV, Bunkoczi G, Chen VB, Headd JJ et al (2012) Graphical tools for macromolecular crystallography in PHENIX. *J Appl Crystallogr* 45(Pt 3):581–586
26. Minor W, Cymborowski M, Otwinowski 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 D Biol Crystallogr* 62:859–866
27. Cymborowski M, Klimecka M, Chruszcz M, Zimmerman MD, Shumilin IA, Borek D et al (2010) To automate or not to automate: this is

- the question. *J Struct Funct Genom* 11 (3):211–221
28. Zimmerman MD, Grabowski M, Domagalski MJ, MacLean EM, Chruszcz M, Minor W (2014) Data management in the modern structural biology and biomedical research environment. *Methods Mol Biol* 1140:1–25
 29. Zimmerman MD, Chruszcz M, Koclega K, Otwinowski Z, Minor W (2005) The Xtaldb system for project salvaging in high-throughput crystallization. *Acta Crystallogr A* 61:c178–c179
 30. Zimmerman MD (2008) The crystallization expert system Xtaldb, and its application to the structure of the 5′- nucleotidase YfbR and other proteins [dissertation]. University of Virginia, Charlottesville
 31. Chruszcz M, Wlodawer A, Minor W (2008) Determination of protein structures—a series of fortunate events. *Biophys J* 95(1):1–9
 32. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28(1):31–36
 33. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A et al (2016) PubChem Substance and Compound databases. *Nucleic Acids Res* 44(D1):D1202–D1213
 34. Formulatrix. <https://formulatrix.com/>. Accessed 6 May 2019
 35. Newman J (2005) Expanding screening space through the use of alternative reservoirs in vapor-diffusion experiments. *Acta Crystallogr D Biol Crystallogr* 61(Pt 4):490–493
 36. Cooper DR, Boczek T, Grelewska K, Pinkowska M, Sikorska M, Zawadzki M et al (2007) Protein crystallization by surface entropy reduction: optimization of the SER strategy. *Acta Crystallogr D Biol Crystallogr* 63(Pt 5):636–645
 37. CakePHP. <https://cakephp.org/>. Accessed 6 May 2019
 38. Shabalin IG, Porebski PJ, Minor W (2018) Refining the macromolecular model - achieving the best agreement with the data from X-ray diffraction experiment. *Crystallogr Rev* 24(4):236–262
 39. Czub MP, Venkataramany BS, Majorek KA, Handing KB, Porebski PJ, Beeram SR et al (2018) Testosterone meets albumin - the molecular mechanism of sex hormone transport by serum albumins. *Chem Sci* 10 (6):1607–1618
 40. Majorek KA, Porebski PJ, Dayal A, Zimmerman MD, Jablonska K, Stewart AJ et al (2012) Structural and immunologic characterization of bovine, horse, and rabbit serum albumins. *Mol Immunol* 52(3–4):174–182
 41. Svare A, Nilsen TI, Asvold BO, Forsmo S, Schei B, Bjoro T et al (2013) Does thyroid function influence fracture risk? Prospective data from the HUNT2 study, Norway. *Eur J Endocrinol* 169(6):845–852
 42. Majorek KA, Kuhn ML, Chruszcz M, Anderson WF, Minor W (2014) Double trouble—buffer selection and His-tag presence may be responsible for nonreproducibility of biomedical experiments. *Protein Sci* 23 (10):1359–1368
 43. How a typo in a catalog number led to the correction of a scientific paper—and what we can learn from that. <https://retractionwatch.com/2018/10/18/how-a-typo-in-a-catalog-number-led-to-the-correction-of-a-scientific-paper-and-what-we-can-learn-from-that/>. Accessed 8 May 2019

Part III

Modeling, Simulation, and Visualization



Chapter 14

Protein Structure Modeling with MODELLER

Benjamin Webb and Andrej Sali

Abstract

Genome sequencing projects have resulted in a rapid increase in the number of known protein sequences. In contrast, only about one-hundredth of these sequences have been characterized at atomic resolution using experimental structure determination methods. Computational protein structure modeling techniques have the potential to bridge this sequence-structure gap. In the following chapter, we present an example that illustrates the use of MODELLER to construct a comparative model for a protein with unknown structure. Automation of a similar protocol has resulted in models of useful accuracy for domains in more than half of all known protein sequences.

Key words Comparative modeling, Fold assignment, Sequence-structure alignment, Model assessment, Multiple templates

1 Introduction

The function of a protein is determined by its sequence and its three-dimensional (3D) structure. Large-scale genome sequencing projects are providing researchers with millions of protein sequences, from various organisms, at an unprecedented pace [1]. However, the rate of experimental structural characterization of these sequences is limited by the cost, time, and experimental challenges inherent in the structural determination by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy [2].

In the absence of experimentally determined structures, computationally derived protein structure models are often valuable for generating testable hypotheses [3, 4]. Such models are generally produced using either comparative modeling methods, or free modeling techniques (also referred to as *ab initio* or *de novo* modeling) [5]. Comparative modeling relies on structural information from related proteins to guide the modeling procedure [6–8]. Free modeling does not require a related protein, but instead uses a variety of methods to combine physics with the known

behaviors of protein structures (for example by combining multiple short structural fragments extracted from known proteins) [9–11]; it is, however, extremely computationally expensive [5]. Comparative protein structure modeling, which this text focuses on, has been used to produce reliable structure models for at least one domain in more than half of all known sequences [12]. Hence, computational approaches can provide structural information for two orders of magnitude more sequences than experimental methods, and are expected to be increasingly relied upon as the gap between the number of known sequences and the number of experimentally determined structures continues to widen.

Comparative modeling consists of four main steps [6] (Fig. 1): (1) fold assignment that identifies overall similarity between the target sequence and at least one known structure (template); (2) alignment of the target sequence and the template(s); (3) building a model based on the alignment with the chosen template(s); and (4) predicting the accuracy of the model.

MODELLER is a computer program for comparative protein structure modeling [13, 14]. In the simplest case, the input is an alignment of a sequence to be modeled with the template structure (s), the atomic coordinates of the template(s), and a simple script file. MODELLER then automatically calculates a model containing all non-hydrogen atoms, without any user intervention and within seconds or minutes on a desktop computer. Apart from model building, MODELLER can perform auxiliary tasks such as fold assignment, alignment of two protein sequences or their profiles [15], multiple alignment of protein sequences and/or structures [16, 17], clustering of sequences and/or structures, and *ab initio* modeling of loops in protein structures [13].

MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints that include (1) homology-derived restraints on the distances and dihedral angles in the target sequence, extracted from its alignment with the template structures [14], (2) stereochemical restraints such as bond length and bond angle preferences, obtained from the CHARMM-22 molecular mechanics force-field [18], (3) statistical preferences for dihedral angles and nonbonded interatomic distances, obtained from a representative set of known protein structures [19, 20], and (4) optional manually curated restraints, such as those from NMR spectroscopy, rules of secondary structure packing, cross-linking experiments, fluorescence spectroscopy, image reconstruction from electron microscopy, site-directed mutagenesis, and intuition (Fig. 1). The spatial restraints, expressed as probability density functions, are combined into an objective function that is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing. This model building procedure is similar to structure determination by NMR spectroscopy.

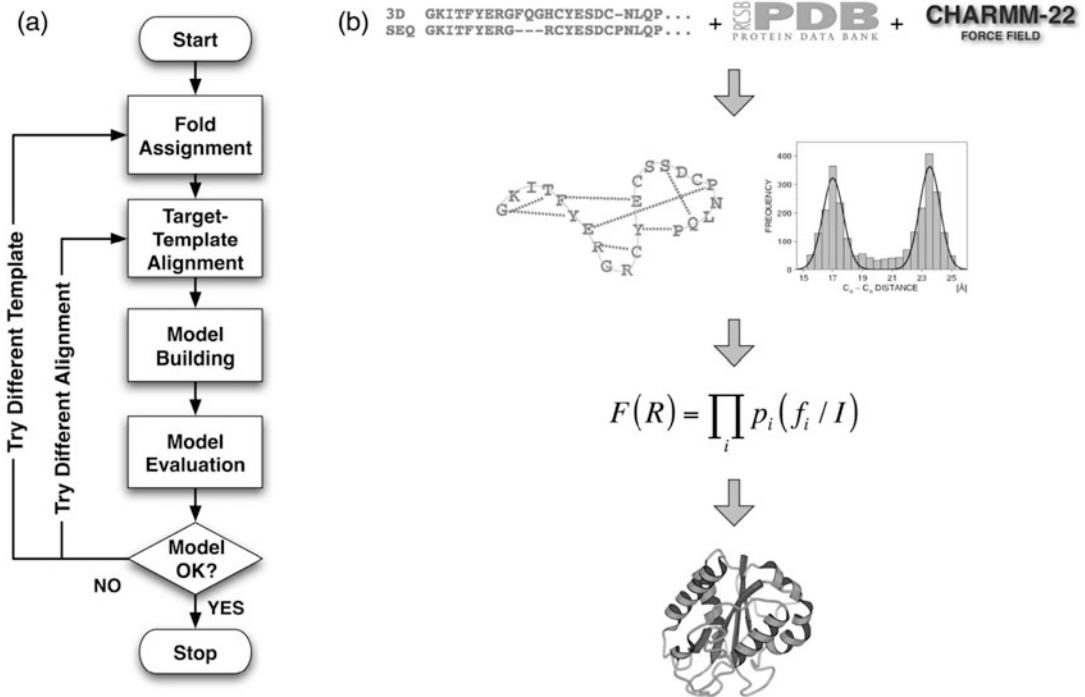


Fig. 1 Comparative protein structure modeling. (a) A flowchart illustrating the steps in the construction of a comparative model [6]. (b) Description of comparative modeling by extraction of spatial restraints as implemented in MODELLER [14]. By default, spatial restraints in MODELLER involve (1) homology-derived restraints from the aligned template structures, (2) statistical restraints derived from all known protein structures, and (3) stereochemical restraints from the CHARMM-22 molecular mechanics force-field. These restraints are combined into an objective function that is then optimized to calculate the final 3D model of the target sequence

In this chapter, we use a sequence with unknown structure to illustrate the use of various modules in MODELLER to perform the four steps of comparative modeling.

2 Materials

To follow the examples in this discussion, both the MODELLER software and a set of suitable input files are needed. The MODELLER software is free for academic use; it can be downloaded from <https://salilab.org/modeller/> and is available in binary form for most common machine types and operating systems (*see Note 1*). This text uses MODELLER 9.21, the most recent version at the time of writing, but the examples should also work with any newer version. The example input files can be downloaded from <https://salilab.org/modeller/tutorial/MMB19.zip>.

All MODELLER scripts are Python scripts. Python is pre-installed on most Linux and Mac machines; Windows users

can obtain it from <https://www.python.org/>. It is not necessary to install Python, or to have a detailed knowledge of its use, to use MODELLER, but it is helpful for creating and understanding the more advanced MODELLER scripts.

Note that `monospaced text` is used below for computer file and folder/directory names, command lines, file contents, and variable and class names.

3 Methods

The procedure for calculating a 3D model for a sequence with unknown structure will be illustrated using the following example: a novel gene for lactate dehydrogenase (LDH) was identified from the genomic sequence of *Trichomonas vaginalis* (TvLDH). The corresponding protein had higher sequence similarity to the malate dehydrogenase of the same species (TvMDH) than to any other LDH [21]. Comparative models were constructed for TvLDH and TvMDH to study the sequences in a structural context and to suggest site-directed mutagenesis experiments to elucidate changes in enzymatic specificity in this apparent case of convergent evolution. The native and mutated enzymes were subsequently expressed and their activities compared [21].

3.1 Fold Assignment

The first step in comparative modeling is to identify one or more templates (sequences with known 3D structure) for the modeling procedure. One way to do this is to search a database of experimentally determined structures extracted from the Protein Data Bank (PDB) [22] to find sequences that have detectable similarity to the target (*see Note 2*). To prepare this database (*see Note 3*), run the following command from the command line (*see Note 4*):

```
$ python make_pdb_95.py > make_pdb_95.log
```

This generates a file called `pdb_95.bin`, which is a binary representation of the search database (*see Note 5*) and a log file, `make_pdb_95.log`. Next, MODELLER's `profile.build()` command is used; this uses the local dynamic programming algorithm to identify sequences related to TvLDH [23]. In the simplest case, `profile.build()` takes as input the target sequence, in file `TvLDH.ali` (*see Note 6*), and the binary database and returns a set of statistically significant alignments (file `build_profile.prf`) and a MODELLER log file (`build_profile.log`). Run this step by typing

```
$ python build_profile.py > build_profile.log
```

The first few lines of the resulting `build_profile.prf` will look similar to (*see Note 7*) the following (note that the rightmost column, containing the primary sequence, has been omitted here for clarity):

```
# Number of sequences : 76
# Length of profile   : 335
# N_PROF_ITERATIONS  : 1
# GAP_PENALTIES_1D   : -500.0 -50.0
# MATRIX_OFFSET      : -450.0
# RR_FILE             : ${LIB}/blosum62.sim.mat
1  TvLDH  S  0  335  1  335  0  0  0  0  0. 0.0
2  1a5zA  X  1  312  75 242  63 229 164 28. 0.58E-07
3  2a92A  X  1  316  8  191  6 186 174 26. 0.11E-03
4  4aj2A  X  1  327  85 301  89 300 207 25. 0.24E-04
5  1b8pA  X  1  327  7  331  6 325 316 42. 0.0
```

The first six lines of this file contain the input parameters used to create the alignments. Subsequent lines contain several columns of data; for the purposes of this example, the most important columns are (1) the 2nd column, containing the PDB code of the related template sequences; (2) the 11th column, containing the percentage sequence identity between the TvLDH and template sequences; and (3) the 12th column, containing the *E*-values for the statistical significance of the alignments. These columns are shown in bold above.

The extent of similarity between the target-template pairs is usually quantified using sequence identity or a statistical measure such as *E*-value (*see Note 8*). Inspection of column 11 shows that a template with a high sequence identity with the target is the 1y7tA structure (45% sequence identity). Further inspection of column 12 shows that there are 15 PDB sequences, all but one corresponding to malate dehydrogenases (1b8pA, 1bdmA, 1civA, 3d5tA, 4h7pA, 4h7pB, 5mdhA, 7mdhA, 5nueA, 4tvoA, 4tvoB, 4uulA, 4uuoA, 4uupA, 1y7tA) that show significant similarities to TvLDH with *E*-values of zero.

3.2 Sequence-Structure Alignment

The next step is to align the target TvLDH sequence with the chosen template (*see Note 9*). Here, the 1y7tA template is used. This alignment is created using MODELLER's `align2d()` function (*see Note 10*). Although `align2d()` is based on a global dynamic programming algorithm [24], it is different from standard sequence-sequence alignment methods because it takes into account structural information from the template when constructing an alignment. This task is achieved through a variable gap penalty function that tends to place gaps in solvent exposed and curved regions, outside secondary structure segments, and not

between two positions that are close in space [16]. In the current example, the target-template similarity is so high that almost any method with reasonable parameters will result in the correct alignment (*see Note 11*).

This step is carried out by running:

```
$ python align2d.py > align2d.log
```

This script reads in the PDB structure of the template, and the sequence of the target (TvLDH) and calls the `align2d()` function to perform the alignment. The resulting alignment is written out in two formats. `TvLDH-1y7tA.ali` in the PIR format is subsequently used by MODELLER for modeling; `TvLDH-1y7tA.pap` in the PAP format is easier to read, for example to see which residues are aligned with each other.

3.3 Model Building

Models of TvLDH can now be built by running:

```
$ python model.py > model.log
```

The script uses MODELLER's `automodel` class, specifying the name of the alignment file to use and the identifiers of the target (TvLDH) and template (1y7tA) sequences. It then asks `automodel` to generate five models (*see Note 12*). Each is assessed with the normalized DOPE assessment method [20]. The five models are written out as PDB files with names `TvLDH.B9999[0001-0005].pdb`.

3.4 Model Evaluation

The log file produced by the model building procedure (`model.log`) contains a summary of each calculation at the bottom of the file. This summary includes, for each of the five models, the MODELLER objective function (*see Note 13*) [14] and the normalized DOPE score (*see Note 14*). These scores can be used to identify which of the five models produced is likely to be the most accurate model (*see Note 15*).

Since the DOPE potential is simply a sum of interactions between pairs of atoms, it can be decomposed into a score per residue, which is termed in MODELLER an “energy profile.” This energy profile can be generated for the model with the best DOPE score by running the `make_energy_profile.py` script. The script outputs the profile, `TvLDH.profile`, in a simple format that is easily displayed in any graphing package. Such a profile is useful to detect local regions of high pseudo-energy that usually correspond to errors in the model (*see Notes 16 and 17*).

3.5 Use of Multiple Templates

One way to potentially improve the accuracy of generated models is to use multiple template structures. When there are multiple templates, different template structures may be of higher local sequence identity to the target (or higher quality) than others in different regions, allowing MODELLER to build a model based on the most useful structural information for each region in the protein. The procedure is demonstrated here using five templates that have high sequence identity to the target (1b8pA, 4h7pA, 4h7pB, 5mdhA, 1y7tA). Input files can be found in the ‘multiple’ subdirectory of the zipfile. The first step is to align all of the templates with each other, which can be done by running:

```
$ python salign.py > salign.log
```

This script uses MODELLER’s `salign()` function [17] to read in all of the template structures and then generate their best structural alignment (*see Note 18*), written out as `templates.ali`.

Next, just as for single-template modeling, the target is aligned with the templates using the `align2d()` function. The function’s `align_block` parameter is set to 5 to align the target sequence with the pre-aligned block of five templates, and not to change the existing alignment between individual templates:

```
$ python align2d.py > align2d.log
```

Finally, model generation proceeds just as for the single-template case (the only difference is that `automodel` is now given a list of all five templates):

```
$ python model.py > model.log
```

Comparison of the normalized DOPE scores from the end of this logfile with those from the single-template case shows an improvement in the DOPE score of the best model from -0.92 to -1.19 . Figure 2 shows the energy profiles of the best scoring models from each procedure (generated using the `plot_profiles.py` script). It can be seen that some of the predicted errors in the single-template model (peaks in the graph) have been resolved in the model calculated using multiple templates.

3.6 External Assessment

Models generated by MODELLER are stored in PDB files, and so can be evaluated for accuracy with other methods if desired. One such method is the ModEval web server at <https://salilab.org/evaluation/>. This server takes as input the PDB file and the MODELLER PIR alignment used to generate it. It returns not only the normalized DOPE score and the energy profile, but also the GA341 assessment score [25, 26] and an estimate of the C α

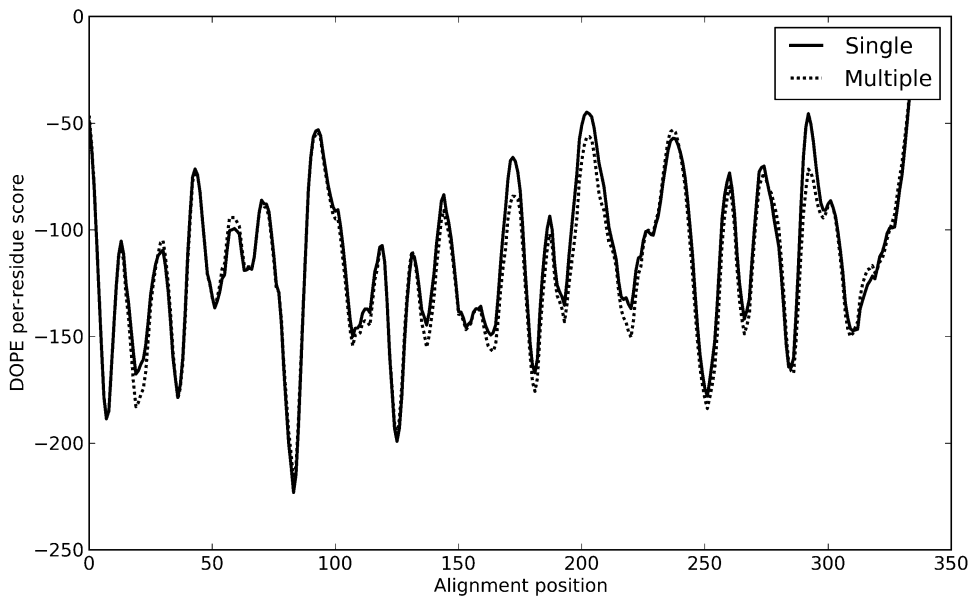


Fig. 2 The DOPE [20] energy profiles for the best-assessed model generated by modeling with a single template (solid line) and multiple templates (dotted line). Peaks (local regions of high, unfavorable score) tend to correspond to errors in the models

RMSD and native overlap between the model and its hypothetical native structure, using the TSVMOD method [27]; native overlap is defined as the fraction of C α atoms in the model that are within 3.5 Å of the same C α atom in the native structure after least squares superposition.

3.7 Structures of Complexes

The example shown here generates a model of a single protein. However, MODELLER can also generate models of complexes of multiple proteins if templates for the entire complex are available; examples can be found in the MODELLER manual. In the case where only templates for the individual subunits in the complex can be found, comparative models can be docked in a pairwise fashion by molecular docking [28, 29] or assembled based on various experimental data to generate approximate models of the complex using a wide variety of integrative modeling methods [30–33]. For example, if a cryoelectron microscopy density map of the complex is available, a model of the whole complex can be constructed by simultaneously fitting comparative models of the subunits into the density map using the MultiFit method [34] or its associated web server at <https://salilab.org/multifit/> [35]. Alternatively, if a small angle X-ray (SAXS) profile of a dimer is available, models of the dimer can be generated by docking the two subunits, constrained by the SAXS data, using the FoXSdock web server at <https://salilab.org/foxsdock/> [36, 37]. Both of these methods are part of the open source *Integrative Modeling Platform* (IMP) package [31].

4 Notes

1. The MODELLER website also contains a full manual, a mailing list, and more example MODELLER scripts. A license key is required to use MODELLER, but this can also be obtained from the website.
2. The sequence identity is a useful predictor of the accuracy of the final model when its value is $>30\%$. It has been shown that models based on such alignments usually have, on average, more than $\sim 60\%$ of the backbone atoms correctly modeled with a root-mean-squared-deviation (RMSD) for $C\alpha$ atoms of less than 3.5 \AA (Fig. 3). Sequence-structure relationships in the “twilight zone” [38] (corresponding to relationships with statistically significant sequence similarity with identities generally

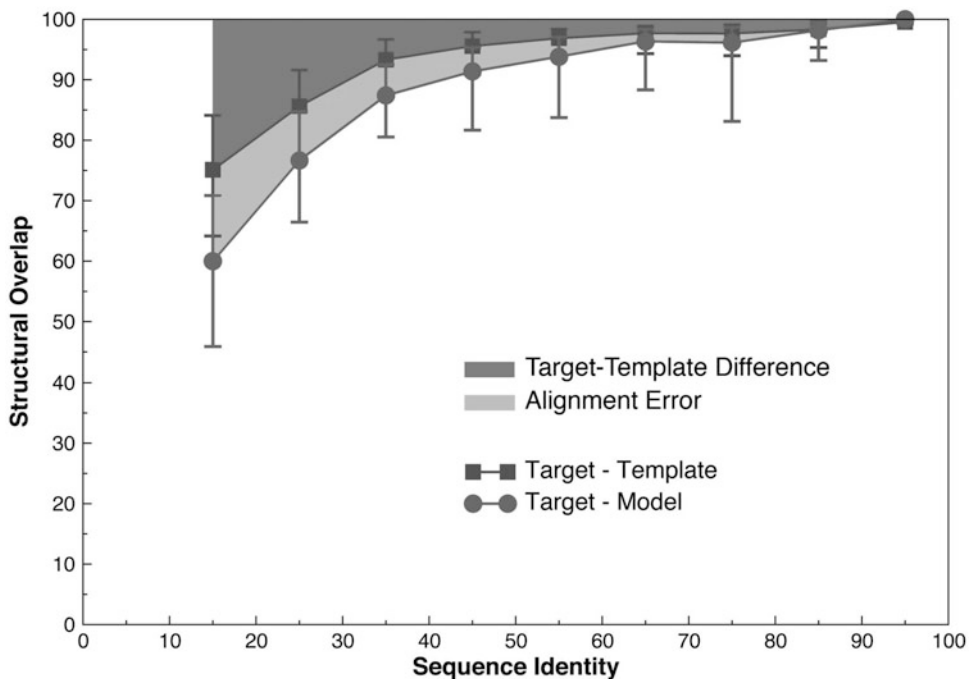


Fig. 3 Average model accuracy as a function of sequence identity [62]. As the sequence identity between the target sequence and the template structure decreases, the average structural similarity between the template and the target also decreases (dark gray area, squares) [63]. Structural overlap is defined as the fraction of equivalent $C\alpha$ atoms. For the comparison of the model with the actual structure (circles), two $C\alpha$ atoms were considered equivalent if they belonged to the same residue and were within 3.5 \AA of each other after least squares superposition. For comparisons between the template structure and the actual target structure (squares), two $C\alpha$ atoms were considered equivalent if they were within 3.5 \AA of each other after alignment and rigid-body superposition. The difference between the model and the actual target structure is a combination of the target-template differences (dark gray area) and the alignment errors (light gray area). The figure was constructed by calculating ~ 1 million comparative models based on single template of varying similarity to the targets. All targets had known (experimentally determined) structures

in the 10–30% range), or the “midnight zone” [38] (corresponding to statistically insignificant sequence similarity), typically result in less accurate models.

3. The database contains sequences of the structures from PDB. To increase the search speed, redundancy is removed from the database; the PDB sequences are clustered with other sequences that are at least 95% identical, and only the representative of each cluster is stored in the database. This database is termed `pdb_95`. A copy of this database is included in the downloaded zipfile as `pdb_95.pir`. Newer versions of this database, updated as new structures are deposited in PDB, can be downloaded from the MODELLER website at <https://salilab.org/modeller/supplemental.html>.
4. MODELLER is a command line tool, so all commands must be run by typing at the command line. All of the necessary input files for this demonstration are in the downloaded zipfile; simply download and extract the zipfile and change into the newly created directory (using the ‘`cd`’ command at the command line). After this, MODELLER scripts can be run as shown in the text. All MODELLER scripts are Python scripts, compatible with both Python 2 and Python3, and so should be run with the ‘`python`’ or ‘`python3`’ commands. (On some systems the full path to the Python interpreter may be necessary, such as `/usr/bin/python` on a Linux or Mac machine or `C:\python27\python.exe` on a Windows system.) MODELLER scripts can also be run from other Python frontends, such as IDLE, if desired. On a Windows system, it is generally **not** a good idea to simply ‘double click’ on a MODELLER Python script, since any output from the script will disappear as soon as it finishes. Finally, if Python is not installed, MODELLER includes a basic Python 2.3 interpreter as ‘`mod-<version>`’. For example, to run the first script using MODELLER version 9.21’s own interpreter, run ‘`mod9.21 make_pdb_95.py`’. Note that `mod9.21` automatically creates a ‘`make_pdb_95.log`’ logfile.
5. The binary database is much faster to use than the original text format database, `pdb_95.pir`. Note, however, that it is not necessarily smaller. This script does not need to be run again unless `pdb_95.pir` is updated.
6. `TvLDH.ali` simply contains the primary sequence of the target, in MODELLER’s variant of the PIR format (which is documented in more detail in the MODELLER manual). This file is included in the zipfile.
7. Although MODELLER’s algorithms are deterministic, exactly the same job run on different machines (e.g., a Linux box versus a Windows or Mac machine) may give different results.

This difference may arise because different machines handle rounding of floating point numbers and ordering of floating point operations differently, and the minor differences introduced can be compounded and end up giving very different outputs. This variation is normal and to be expected, and so the results shown in this text may differ from those obtained by running MODELLER elsewhere.

8. The sequence identity is not a statistically reliable measure of alignment significance and corresponding model accuracy for values lower than 30% [38, 39]. During a scan of a large database, for instance, it is possible that low values occur purely by chance. In such cases, it is useful to quantify the sequence-structure relationship using more robust measures of statistical significance, such as *E*-values [40], that compare the score obtained for an alignment with an established background distribution of such scores.

One other problem of using sequence identity as a measure to select templates is that, in practice, there is no single generally used way to normalize it [39]. For instance, local alignment methods usually normalize the number of identically aligned residues by the length of the alignment, while global alignment methods normalize it by either the length of the target sequence or the length of the shorter of the two sequences. Therefore, it is possible that alignments of short fragments produce a high sequence identity but do not result in an accurate model. Measures of statistical significance do not suffer from this normalization problem because the alignment scores are corrected for the length of the aligned segment before the significance is computed [40, 41].

9. After a list of all related protein structures and their alignments with the target sequence has been obtained, template structures are usually prioritized depending on the purpose of the comparative model. Template structures may be chosen based purely on the target-template sequence identity or a combination of several other criteria, such as the experimental accuracy of the structures (resolution of X-ray structures, number of restraints per residue for NMR structures), conservation of active-site residues, holo-structures that have bound ligands of interest, and prior biological information that pertains to the solvent, pH, and quaternary contacts. In this case an MDH template with a moderately high sequence identity was chosen. (In practice, since the modeling is generally inexpensive, it can be simply repeated with a different template or set of templates and the resulting models compared for utility.) One of the detected templates, 4uulA, is TvLDH itself, the structure of which was recently determined in a study of convergent evolution of LDH and MDH [42]; this template was excluded from

selection in order to demonstrate the comparative modeling method.

10. Although fold assignment and sequence-structure alignment are logically two distinct steps in the process of comparative modeling, in practice almost all fold assignment methods also provide sequence-structure alignments. In the past, fold assignment methods were optimized for better sensitivity in detecting remotely related homologs, often at the cost of alignment accuracy. However, recent methods simultaneously optimize both the sensitivity and alignment accuracy. For the sake of clarity, however, they are still considered as separate steps in the current chapter.
11. Most alignment methods use either the local or global dynamic programming algorithms to derive the optimal alignment between two or more sequences and/or structures. The methods, however, vary in terms of the scoring function that is being optimized. The differences are usually in the form of the gap penalty function (linear, affine, or variable) [16], the substitution matrix used to score the aligned residues (20×20 matrices derived from alignments with a given sequence identity, those derived from structural alignments, and those incorporating the structural environment of the residues) [43], or combinations of both [44–47]. There doesn't yet exist a single universal scoring function that guarantees the most accurate alignment for all situations. Above 30–40% sequence identity, alignments produced by almost all methods are similar. However, in the twilight and midnight zones of sequence identity, models based on the alignments of different methods tend to have significant variations in accuracy. Improving the performance and accuracy of methods in this regime remains one of the main tasks of comparative modeling [48, 49].
12. To generate each model, MODELLER takes a starting structure, which is simply the target sequence threaded onto the template backbone, adds some randomization to the coordinates, and then optimizes it by searching for the minimum of its scoring function. Since finding the global minimum of the scoring function is not guaranteed, it is usually recommended to repeat the procedure multiple times to generate an ensemble of models; the randomization is necessary otherwise the same model would be generated each time. Computing multiple models is particularly important when the sequence-structure alignment contains different templates with many insertions and/or deletions. Calculating multiple models allows for better sampling of the different template segments and the conformations of the unaligned regions. The best scoring model among these multiple models is generally more accurate than the first model produced.

13. The MODELLER objective function is a measure of how well the model satisfies the input spatial restraints. Lower values of the objective function indicate a better fit with the input data and, thus, models that are likely to be more accurate [14].
14. The Discrete Optimized Protein Energy (DOPE) [20] is an atomic distance-dependent statistical potential based on a physical reference state that accounts for the finite size and spherical shape of proteins. The reference state assumes that a protein chain consists of noninteracting atoms in a homogeneous sphere of equivalent radius to that of the corresponding protein. The DOPE potential was derived by comparing the distance statistics from a nonredundant PDB subset of 1472 high-resolution protein structures with the distance distribution function of the reference state. By default, the DOPE score is not included in the model building routine, and thus can be used as an independent assessment of the accuracy of the output models. The DOPE method assigns a score for a model by considering the positions of all non-hydrogen atoms, with lower scores predicting more accurate models. Since DOPE is a pseudo-energy dependent on the composition and size of the system, DOPE scores are only directly comparable for models with the same set of atoms (so can, for example, be used to rank multiple models of the same protein, but cannot be used without additional approximations to compare models of a protein and its mutant). The normalized DOPE (or *z*-DOPE) score, however, is a *z* score that relates the DOPE score of the model to the average observed DOPE score for “reference” protein structures of similar size [27]. Negative normalized DOPE scores of -1 or below are likely to correspond to models with the correct fold.
15. Different measures to predict errors in a protein structure perform best at different levels of resolution. For instance, physics-based force-fields may be helpful at identifying the best model when all models are very close to the native state (<1.5 Å RMSD, corresponding to $\sim 85\%$ target-template sequence identity). In contrast, coarse-grained scores such as atomic distance statistical potentials have been shown to have the greatest ability to differentiate models in the ~ 3 Å C α RMSD range. Tests show that such scores are often able to identify a model within 0.5 Å C α RMSD of the most accurate model produced [50]. When multiple models are built, the DOPE score generally selects a more accurate model than the MODELLER objective function.
16. Segments of the target sequence that have no equivalent region in the template structure (i.e., insertions or loops) are among the most difficult regions to model [13, 51–53]. This difficulty is compounded when the target and template are distantly

related, with errors in the alignment leading to incorrect positions of the insertions and distortions in the loop environment. Using alignment methods that incorporate structural information can often correct such errors [16]. Once a reliable alignment is obtained, various modeling protocols can predict the loop conformation, for insertions of up to approximately 15 residues long [13, 51, 54–57].

17. As a consequence of sequence divergence, the mainchain conformation of a protein can change, even if the overall fold remains the same. Therefore, it is possible that in some correctly aligned segments of a model, the template is locally different ($<3 \text{ \AA}$) from the target, resulting in errors in that region. The structural differences are sometimes not due to differences in sequence, but are a consequence of artifacts in structure determination or structure determination in different environments (e.g., packing of subunits in a crystal and ligands). The simultaneous use of several templates can minimize this kind of error [58, 59].
18. It is particularly important to generate the best alignment of the structures to minimize conflicting information (e.g., one template suggesting that two $C\alpha$ atoms in the target are close, and another suggesting they are widely separated). SALIGN [17] uses both sequence- and structure-dependent features to align multiple structures. It employs an iterative procedure to determine the input parameters that maximize the structural overlap of the generated alignment.

Acknowledgments

We are grateful to all members of our research group. This review is partially based on our previous reviews [60, 61]. We also acknowledge support from National Institutes of Health (U54 GM094625) as well as computing hardware support from Ron Conway, Mike Homer, Hewlett-Packard, NetApp, IBM, and Intel.

References

1. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17(6):333–351. <https://doi.org/10.1038/nrg.2016.49>
2. Holcomb J, Spellmon N, Zhang Y, Doughan M, Li C, Yang Z (2017) Protein crystallization: eluding the bottleneck of X-ray crystallography. *AIMS Biophys* 4(4):557–575. <https://doi.org/10.3934/biophys.2017.4.557>
3. Baker D, Sali A (2001) Protein structure prediction and structural genomics. *Science* 294(5540):93–96
4. Schwede T, Sali A, Honig B, Levitt M, Berman H, Jones D, Brenner S, Burley S, Das R, Dokholyan N, Dunbrack RJ, Fidelis K, Fiser A, Godzik A, Huang Y, Humblet C, Jacobson M, Joachimiak A, Krystek SJ, Kortemme T, Kryshchukovych A, Montelione G, Moulton J, Murray D,

- Sanchez R, Sosnick T, Standley D, Stouch T, Vajda S, Vasquez M, Westbrook J, Wilson I (2009) Outcome of a workshop on applications of protein models in biomedical research. *Structure* 17(2):151–159
5. Zhang Y (2008) Progress and challenges in protein structure prediction. *Curr Opin Struct Biol* 18(3):342–348. <https://doi.org/10.1016/j.sbi.2008.02.004>
 6. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29:291–325
 7. Eswar N, Sali A (2009) Protein structure modeling. In: Sussman JL, Spadon P (eds) *From molecules to medicine, structure of biological macromolecules and its relevance in combating new diseases and bioterrorism, NATO science for peace and security series - A: Chemistry and biology*. Springer, Dordrecht, the Netherlands, pp 139–151
 8. Ginalski K (2006) Comparative modeling for protein structure prediction. *Curr Opin Struct Biol* 16(2):172–177. <https://doi.org/10.1016/j.sbi.2006.02.003>
 9. Das R, Baker D (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem* 77:363–382. <https://doi.org/10.1146/annurev.biochem.77.062906.171838>
 10. Zhang Y, Skolnick J (2004) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci U S A* 101(20):7594–7599. <https://doi.org/10.1073/pnas.0305695101>
 11. Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins Suppl* 3:171–176
 12. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H, Yang Z, Meng EC, Pettersen EF, Huang CC, Datta RS, Sampathkumar P, Madhusudhan MS, Sjolander K, Ferrin TE, Burley SK, Sali A (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res* 39:465–474
 13. Fiser A, Do RKG, Sali A (2000) Modeling of loops in protein structures. *Protein Sci* 9(9):1753–1773
 14. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779–815
 15. Marti-Renom MA, Madhusudhan MS, Sali A (2004) Alignment of protein sequences by their profiles. *Protein Sci* 13(4):1071–1087
 16. Madhusudhan MS, Marti-Renom MA, Sanchez R, Sali A (2006) Variable gap penalty for protein sequence-structure alignment. *Protein Eng Des Sel* 19(3):129–133
 17. Madhusudhan MS, Webb BM, Marti-Renom MA, Eswar N, Sali A (2009) Alignment of multiple protein structures based on sequence and structure features. *Protein Eng Des Sel* 22:569–574
 18. Brooks BR, Brooks CL III, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. *J Comput Chem* 30(10):1545–1614. <https://doi.org/10.1002/jcc.21287>
 19. Sali A, Overington JP (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci* 3(9):1582–1596
 20. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* 15(11):2507–2524
 21. Wu G, Fiser A, ter Kuile B, Sali A, Muller M (1999) Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc Natl Acad Sci U S A* 96(11):6285–6290
 22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242
 23. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147(1):195–197
 24. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48(3):443–453
 25. John B, Sali A (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res* 31(14):3982–3992. <https://doi.org/10.1093/nar/gkg460>
 26. Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. *Protein Sci* 11(2):430–448. <https://doi.org/10.1110/ps.22802>
 27. Eramian D, Eswar N, Shen M, Sali A (2008) How well can the accuracy of comparative protein structure models be predicted? *Protein Sci* 17(11):1881–1893

28. Vajda S, Kozakov D (2009) Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol* 19 (2):164–170. <https://doi.org/10.1016/j.sbi.2009.02.008>
29. Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. *Proteins* 78(15):3073–3084. <https://doi.org/10.1002/prot.22818>
30. Alber F, Forster F, Korkin D, Topf M, Sali A (2008) Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem* 77:443–477
31. Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A (2012) Putting the pieces together: integrative structure determination of macromolecular assemblies. *PLoS Biol* 10 (1):e1001244
32. Robinson C, Sali A, Baumeister W (2007) The molecular sociology of the cell. *Nature* 450 (7172):973–982
33. Ward A, Sali A, Wilson I (2013) Integrative structural biology. *Science* 339(6122):913–5. <https://doi.org/10.1126/science.1228565>
34. Lasker K, Sali A, Wolfson HJ (2010) Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins* 78:3205–3211
35. Tjioe E, Lasker K, Webb B, Wolfson H, Sali A (2011) MultiFit: a web server for fitting multiple protein structures into their electron microscopy density map. *Nucleic Acids Res* 39:167–170
36. Schneidman-Duhovny D, Hammel M, Sali A (2011) Macromolecular docking restrained by a small angle X-ray scattering profile. *J Struct Biol* 3:461–471
37. Schneidman D, Hammel M, Tainer J, Sali A (2016) FoXS, FoXSDock, and MultiFoXS: single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res* 44(W1):W424–W429. <https://doi.org/10.1093/nar/gkw389>
38. Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng* 12 (2):85–94
39. May AC (2004) Percent sequence identity; the need to be explicit. *Structure* 12(5):737–738. <https://doi.org/10.1016/j.str.2004.04.001>
40. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402
41. Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. *J Mol Biol* 276(1):71–84. <https://doi.org/10.1006/jmbi.1997.1525>
42. Steindel PA, Chen EH, Wirth JD, Theobald DL (2016) Gradual neofunctionalization in the convergent evolution of trichomonad lactate and malate dehydrogenases. *Protein Sci* 25 (7):1319–1331. <https://doi.org/10.1002/pro.2904>
43. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89(22):10915–10919
44. Zhou H, Zhou Y (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* 58 (2):321–328. <https://doi.org/10.1002/prot.20308>
45. McGuffin LJ, Jones DT (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* 19 (7):874–881
46. Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* 51(4):504–514. <https://doi.org/10.1002/prot.10369>
47. Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 310(1):243–257. <https://doi.org/10.1006/jmbi.2001.4762>
48. Dunbrack RL Jr (2006) Sequence comparison and protein structure prediction. *Curr Opin Struct Biol* 16(3):374–384. <https://doi.org/10.1016/j.sbi.2006.05.006>
49. Xiang Z (2006) Advances in homology protein structure modeling. *Curr Protein Pept Sci* 7 (3):217–227
50. Eramian D, Shen M, Devos D, Melo F, Sali A, Marti-Renom M (2006) A composite score for predicting errors in protein structure models. *Protein Sci* 15(7):1653–1666
51. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins* 55(2):351–367. <https://doi.org/10.1002/prot.10613>
52. Zhao S, Zhu K, Li J, Friesner RA (2011) Progress in super long loop prediction. *Proteins* 79 (10):2920–2935. <https://doi.org/10.1002/prot.23129>
53. Fernandez-Fuentes N, Oliva B, Fiser A (2006) A supersecondary structure library and search

- algorithm for modeling loops in protein structures. *Nucleic Acids Res* 34(7):2085–2097. <https://doi.org/10.1093/nar/gkl156>
54. van Vlijmen HW, Karplus M (1997) PDB-based protein loop prediction: parameters for selection and methods for optimization. *J Mol Biol* 267(4):975–1001. <https://doi.org/10.1006/jmbi.1996.0857>
 55. Coutsias EA, Seok C, Jacobson MP, Dill KA (2004) A kinematic view of loop closure. *J Comput Chem* 25(4):510–528. <https://doi.org/10.1002/jcc.10416>
 56. Karami Y, Guyon F, De Vries S, Tuffery P (2018) DaReUS-Loop: accurate loop modeling using fragments from remote or unrelated proteins. *Sci Rep* 8(1):13673. <https://doi.org/10.1038/s41598-018-32079-w>
 57. Nguyen SP, Li Z, Xu D, Shang Y (2017) New deep learning methods for protein loop modeling. *IEEE/ACM Trans Comput Biol Bioinform* 16:596. <https://doi.org/10.1109/TCBB.2017.2784434>
 58. Sanchez R, Sali A (1997) Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins Suppl* 1:50–58
 59. Srinivasan N, Blundell TL (1993) An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng* 6(5):501–512
 60. Webb B, Sali A (2014) Protein structure modeling with MODELLER. In: Kihara D (ed) *Methods in molecular biology*, vol 1137. Springer, New York, pp 1–15
 61. Webb B, Sali A (2017) Protein structure modeling with MODELLER. *Methods Mol Biol* 1654:39–54
 62. Sanchez R, Sali A (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc Natl Acad Sci U S A* 95(23):13597–13602
 63. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* 5(4):823–826



Chapter 15

Parameterization of a Dioxygen Binding Metal Site Using the MCPB.py Program

Pengfei Li and Kenneth M. Merz Jr.

Abstract

The MCPB.py program greatly facilitates force field parameterization for metal sites in metalloproteins and organometallic compounds. Herein we present an example of MCPB.py to the parameterization of the dioxygen binding metal site of peptidylglycine-alpha-hydroxylating monooxygenase (PHM), which contains a copper ion. In this example, we also extend the functionality of MCPB.py to support molecular dynamics (MD) simulations in GROMACS through a python script. Illustrative MD simulations were performed using GROMACS and the results were analyzed. Notes about the program were also provided in this chapter, to assist MCPB.py users for metal site parameterizations.

Key words Force field, Metal ion, Software, Molecular dynamics, AMBER, GROMACS

1 Introduction

Metal ions play vital roles in the structure and function of a myriad of proteins, stabilization of DNA, etc. [1–4]. Along with the increase in computational power, computer simulations play a more and more active role in scientific research. Among various methods, the classical force field approach has been extensively employed in the modeling of biomolecules such as proteins, nucleic acids, and lipids [5–10]. However, force field parameterization for metal ion containing systems remains a challenge [11]. Different approaches have been proposed to model metal ions, such as the bonded model [12, 13], the nonbonded model [14, 15], and the cationic dummy atom model [16, 17]. Among these models, the bonded model for metal ions requires a number of parameters and its parameterization can be tedious.

MCPB.py [18], a python-based metal center parameter builder, has been developed in recent years and can significantly

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-1-0716-0892-0_15) contains supplementary material, which is available to authorized users.

Yu Wai Chen and Chin-Pang Benu Yiu (eds.), *Structural Genomics: General Applications*, Methods in Molecular Biology, vol. 2199, https://doi.org/10.1007/978-1-0716-0892-0_15, © Springer Science+Business Media, LLC, part of Springer Nature 2021

decrease the amount of work needed to build bonded models for metalloproteins and organometallic compounds. Herein, we provide an example parameterization of a dioxygen binding metal site. This example is complementary to MCPB.py tutorials available online. Moreover, the original development of MCPB.py targeted the AMBER community. Considering the significant number of users of the GROMACS software package [19, 20], herein we also extend the MCPB.py protocol to support molecular dynamics (MD) simulations in GROMACS. To avoid confusion, we note that the extended protocol still depends on the AmberTools software package [21]. Finally, this chapter also intends to help users to become more familiar with the MCPB.py program and provide tips for its normal use. Several modeling files for the example case are provided in the Electronic Supplementary Materials, included on the chapter's webpage, in order to facilitate the reproduction of this tutorial.

2 Materials

The metal site for parameterization is shown in Fig. 1, where a copper ion coordinates with two histidine (HIS) residues, one methionine (MET), and a dioxygen group. The initial preparation of the necessary files is essential for MCPB.py. It is best to be careful in these initial steps to avoid having to start over again when errors are encountered. Note that the commands in this tutorial are performed under the linux or unix platform.

2.1 Preparation of the PDB File

1. Download the PDB file of peptidylglycine-alpha-hydroxylating monooxygenase (PHM) from the Protein Data Bank (PDB ID: 1SDW) [22]. Note that the structure has the first 44 residues missing, and the residues will be re-sequenced in the current protocol for preparation of the PDB file. Attention should be paid to the format of the PDB file (*see Note 1*), and tips are provided for “cleaning” the PDB file and adding hydrogen atoms to the PDB file (*see Note 2*).
2. Use *awk* to extract the parts of PDB file we want to keep:

```
$ cat 1sdw.pdb | awk '$1=="ATOM"' > 1sdw_protein.pdb
$ cat 1sdw.pdb | awk '$1=="HETATM"' | awk '$6==358' > 1sdw_cu.pdb
$ cat 1sdw.pdb | awk '$1=="HETATM"' | awk '$4=="OXY"' > 1sdw_oxy.pdb
```

In the “HETATM” section, only the copper ion with residue ID of 358 and the dioxygen molecule were kept because they are involved in the metal site that we want to parameterize (Fig. 1). There are another two metal ions in the PDB file that we have not considered in this parameterization

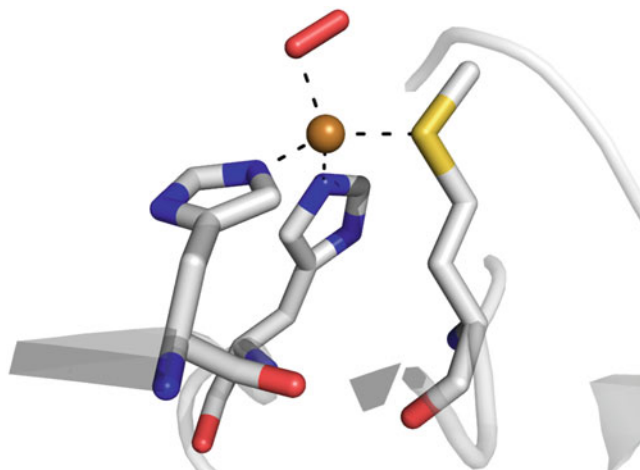


Fig. 1 The metal site selected for parameterization in this work. In which the copper ion is coordinated to HIS242 (left), HIS244 (middle), MET314 (right), and OXY360 (upper). Structure is from the PDB entry 1SDW

(i.e., treating these sites as apo), while in an actual research project on this system these metal sites may also need to be parameterized, for which the current protocol can be adapted.

3. Add hydrogen atoms to the `1sdw_protein.pdb` file using the *H++* webserver [23], which is at <http://biophysics.cs.vt.edu/index.php>. Then download the generated AMBER topology and coordinate files, and use the *ambpdb* program in AmberTools to create a protein PDB file based on these two files:

```
$ ambpdb -p 0.15_80_10_pH6.5_1sdw_protein.top -c
0.15_80_10_pH6.5_1sdw_protein.crd > 0.15_80_10_pH6.5_1sdw_protein.pdb
```

4. Combine the generated protein PDB file, the copper PDB file, and the dioxygen group PDB file into one single PDB file:

```
$ cat 0.15_80_10_pH6.5_1sdw_protein.pdb 1sdw_cu.pdb 1sdw_oxy.pdb | awk
'$1!="END"' > 1sdw_H.pdb
```

5. Check the PDB structure using the VMD program, [33] which can be obtained from <https://www.ks.uiuc.edu/Research/vmd/>, to see whether the protonation states of the metal-ligand residues are correct or not. Here the residue 200 (this residue number has been re-sequenced, which corresponds to residue 244 in the PDB entry 1SDW) should be HID instead of HIP (in the AMBER naming scheme, HID, HIE, HIP are HIS residues which have proton on the N_{δ} , N_{ϵ} , and both atoms in the imidazole ring, respectively), so we edit the `1sdw_H.pdb`

file by changing the residue name of residue 200 from HIP to HID and deleting its HE2 atom. Afterwards we use *pdb4amber* in AmberTools to clean up the PDB file:

```
$ pdb4amber -i 1sdw_H.pdb -o 1sdw_H_clean.pdb
```

2.2 Preparation of the mol2 Files

Relevant mol2 files (and maybe frcmod files) are required for the “unnatural” residues (*see Note 3*), and they can be created based on the PDB file (*see Note 4*). However, attention should be paid to the format of the mol2 files (*see Note 5*). In the current study, we treat the copper ion as Cu^{2+} and the dioxygen group as superoxide, while in the proposed mechanism the complex exists as a set of resonance structures: $\text{Cu}^{2+}\text{-OO}^-$ and $\text{Cu}^+\text{-OO}$ [24]. If one wants to model the $\text{Cu}^+\text{-OO}$ state, the following procedure can be adapted by changing the charge of copper to +1 and changing the charge of the dioxygen group to zero.

1. Create a PDB file which contains only the copper ion and convert it to a mol2 file (named CU.mol2) using *antechamber* [25]:

```
$ cat 1sdw_H_clean.pdb | awk '$1=="HETATM"' | awk '$4=="CU"' > CU.pdb
$ antechamber -fi pdb -fo mol2 -i CU.pdb -o CU.mol2 -rn CU -pf y
```

Then enter the mol2 file and change the atom type and charge of the copper ion to “CU” and 2.0, respectively. Note that certain antechamber versions cannot perform this conversion, users can use the *metaldpb2mol2.py* script instead (please check the webpage <http://ambermd.org/tutorials/advanced/tutorial20/mcpbpy.htm> for details).

2. Create a PDB file that contains only the dioxygen group, convert it to a mol2 file (named OXY.mol2):

```
$ cat 1sdw_H_clean.pdb | awk '$1=="HETATM"' | awk '$4=="OXY"' > OXY.pdb
$ antechamber -fi pdb -fo mol2 -i OXY.pdb -o OXY.mol2 -rn OXY -pf y
```

Then enter the mol2 file and change both of the two charges to -0.5 . These charges do not need to be accurate but their sum should be correct (*see Note 5*). Afterwards we can use *parmchk2* to generate the missing parameters for the dioxygen group:

```
$ parmchk2 -f mol2 -I OXY.mol2 -o OXY.frcmod
```

The generated OXY.frcmod does not contain any parameters. This is because we already have parameters for the bond type o-o in the general amber force field (GAFF), which we are going to use to build the system, hence here we

do not need a frcmol file for the OXY.mol2 file (*see Note 6*). However, frcmol files are necessary for some cases, and tips of how to generate them are provided in **Note 7**.

2.3 Create a MCPB.py Input File

An input file is required for the MCPB.py modeling (*see Note 8*). Herein we create a MCPB.py input file, in which we set the variables for the original PDB file (which means the PDB file prepared for the MCPB.py protocol), group name, atomic ID for the copper ion, distance cutoff for determining the connectivity between the metal ion and ligating O/N/S atoms, the mol2 file for the copper ion, the mol2 file for the dioxygen group, and the optimization option of the large model. Here we use a 2.3 Å cutoff because the default value 2.8 Å will cause MCPB.py to assign a nonphysical coordination bond between the copper ion and the other oxygen atom in the dioxygen group. It is better to measure the distances between the metal ion and surrounding atoms before choosing an appropriate cutoff value. Since GAFF is used by default, we do not need to set the gaff variable in the input file.

To assist users to reproduce the example, the following files are attached in the Electronic Supplementary Materials: (1) the original PDB file; (2) the mol2 files for the copper ion and the dioxygen group; (3) the MCPB.py input file.

3 Methods

Note the following modeling is based on the release of MCPB.py in AmberTools19 [26].

3.1 Perform the First Step of the MCPB.py Protocol

Simply run the command:

```
$ MCPB.py -i 1sdw_MCPBpy.in -s 1
```

This step will generate the PDB files for the small, standard, and large models, as well as the fingerprint files for the standard and large models, along with the quantum input files for the small and large models. Inside the fingerprint file for the standard model, the last section (with lines beginning with “LINK”) is for the coordinate bonds between metal ion and ligating atoms. Herein, each atom is represented by its atom IDs in the PDB file followed by a dash sign followed by its atom name. We can see that all the four coordination bonds are correctly assigned and no other coordination bonds are assigned so we progress to the next step. Otherwise users can edit these “LINK” lines to meet their own needs (*see Note 9*).

3.2 Quantum Calculations by Gaussian16 [27]

Both Gaussian and GAMESS-US outputs are supported by MCPB.py (*see Note 10*) but their formats are different (*see Note 11*). In the present example, we use Gaussian16 to perform the quantum

calculations. There are three quantum calculations that need to be performed: (1) geometry optimization of the small model; (2) frequency calculation based on the optimized geometry in order to get the Hessian matrix of the small model; (3) Merz–Kollman population analysis [28] of the large model. The Hessian matrix will be used for force constant calculation in the second step of the MCPB.py protocol, and the population analysis results will be used for the restrained electrostatic potential (RESP) charge fitting [29] in the third step of the MCPB.py protocol.

One can (or may have to) edit the input files of the calculations. Relevant tips are provided in **Note 12**. In addition, all the quantum calculations need to finish normally (i.e., without any errors). Herein we change the multiplicity to 3 in these calculations by considering the electronic structure of the $\text{Cu}^{2+}\text{-OO}^-$ state. We also modify these input files to use 16 cpus and 16 GB memory to facilitate the calculations.

1. We run the Gaussian16 using the following commands:

```
$ g16 < PHM_small_opt.com > PHM_small_opt.log
$ g16 < PHM_small_fc.com > PHM_small_fc.log
```

Tips for this step are provided in **Note 13**. Herein we can see that the geometry optimization finished normally and the convergence criteria were satisfied in this step. However, the same convergence criteria were not satisfied in the frequency calculation. This is because during the geometry optimization, an estimated force constant matrix is used while in the frequency calculation the accurate force constant matrix (which is calculated at the same level of theory) is used. So we need to further optimize the structure. We can copy the input file for frequency calculation to a new file named PHM_small_fc2.com. And then manually change the keyword “Freq” to “Opt=CalcAll”. This tells Gaussian to do a geometry optimization with the accurate force constant matrix that is updated at every step. This keyword also tells Gaussian to perform a frequency calculation automatically following the geometry optimization. The same chk file (PHM_small_opt.chk) will be used for this calculation, and it will be updated during the calculation. After modifying the input file, we run the command:

```
$ g16 < PHM_small_fc2.com > PHM_small_fc2.log
```

This geometry optimization fully converged and there were no imaginary frequencies generated. Next we convert the chk file to the fchk file by the following command:


```
$ formchk PHM_small_opt.chk
```

2. For the Merz–Kollman population analysis of the large model, we run the command:

```
$ g16 < PHM_large_mk.com > PHM_large_mk.log
```

Because we set the `large_opt` variable to 1 in the `MCPB.py` input file, the generated Gaussian input file for the Merz–Kollman population analysis (`PHM_large_mk.com`) includes a geometry optimization for the hydrogen atoms before the final population analysis. This calculation finished normally and we will use the output file for the RESP charge fitting procedure. Tips for this step are provided in **Note 14**.

3.3 Perform the Remaining Steps of the MCPB.py Protocol

1. Herein we use the Seminario method to parameterize the metal site based on the Cartesian Hessian matrix that is saved in the `fchk` file. The `fchk` file should be available in the working directory where `MCPB.py` is being used. Tips for this step are provided in **Note 15**. Herein, simply run the command:

```
$ MCPB.py -i lsdw_MCPBpy.in -s 2
```

It will generate a parameter file named “`PHM_mcpbpy.frcmod`”. This file will be used in the final LEaP modeling step. During the current step, each ligating atom will be assigned a new atom type, in order to differentiate them from other atom types and also from each other. No unusual force constants are observed in the `frcmod` file, so we go ahead to the next step.

2. Next we perform RESP charge fitting for the metal site residues. The output file of Merz–Kollman population analysis should be in the working directory for the `MCPB.py` program. Tips for this step are provided in **Note 16**. Herein, simply run the command:

```
$ MCPB.py -i lsdw_MCPBpy.in -s 3
```

It will perform RESP charge fitting and generate the updated `mol2` files for the metal site residues. These files are `HD1.mol2`, `HD2.mol2`, `MT1.mol2`, `CU1.mol2`, and `OY1.mol2`. These `mol2` files contain the fitted RESP charges. As indicated in the file names, the metal site residues were renamed, in order to differentiate them from the original residues, because they have updated atom types for the ligating atoms and the updated partial charges. There are no unusual

(e.g., too large) partial charges in the generated mol2 files, hence we proceed to the next step.

3. Simply run the command:

```
$ MCPB.py -i lsdw_MCPBpy.in -s 4
```

It will generate a *LEaP* input file (“PHM_tleap.in”) along with an updated PDB file (“PHM_mcpbpy.pdb”) for the LEaP step. In this PDB file, the residue names of the metal site residues are updated. However, the generated LEaP input file may need edits before progressing to the next step (*see Note 17*).

3.4 Perform the LEaP Step and Check the Generated Files

1. In this step, LEaP is used to create the AMBER topology and coordinate files. Herein we use the ff14SB force field [30] for the protein, TIP3P water model [31] for the solvent, van der Waals (VDW) parameters of the Cl⁻ ions are from the hydration free energy (HFE) parameter set [32], and VDW parameters of the Cu²⁺ ion are from the ion oxygen distance (IOD) parameter set [14]. After checking that the LEaP input file is fine, we run the command to generate the topology and coordinate files:

```
$ tleap -s -f PHM_tleap.in > PHM_tleap.out
```

2. After this step it is best to check the generated topology and coordinate files using VMD and ParmEd (*see Notes 18 and 19*). One can use the VMD program [33] through the command:

```
$ vmd -parm7 PHM_solv.prmtop -rst7 PHM_solv.inpcrd
```

In this way we can validate that there are bonds between the metal ion and its ligating atoms, and peptide bonds between the metal site residues and their neighboring amino acids, as well as the disulfide bonds. Moreover, we use *ParmEd* in AmberTools to check whether the force field parameters of the metal site residues are assigned correctly or not. Simply run the command:

```
$ parmed -i parmed.in -p PHM_solv.prmtop -c PHM_solv.inpcrd
```

Where parmed.in is the ParmEd input file to check the parameters.

3.5 Convert the AMBER Topology and Coordinate Files to the GROMACS Format

This step needs a python script called *amb2gro_top_gro.py* in AmberTools19 or a more recent version of AmberTools. In which the ParmEd module will be utilized to convert the AMBER topology and coordinate files to the GROMACS topology and coordinate files (with .top and .gro suffices, respectively). Simply run the command for this step:

```
$ amb2gro_top_gro.py -p PHM_solv.prmtop -c PHM_solv.inpcrd -t PHM_solv.top
-g PHM_solv.gro -b PHM_solv.gromacs.pdb
```

Herein the “PHM_solv.top” is the GROMACS topology file, which contains all the force field parameters, while “PHM_solv.gro” is the GROMACS coordinate file. Only these two files are needed to perform minimization or MD simulations in GROMACS, while the PDB file “PHM_solv.gromacs.pdb” is just a reference. Note that the current tutorial only serves as an illustrative example so herein we have skipped benchmark calculations. However, benchmark calculations on the energies and forces between the two software packages are highly recommended in an actual research project to make sure the conversion is correct (*see Note 20*).

3.6 Perform MD simulations in GROMACS and Analyze the Results

Afterwards, we can perform the minimization and MD simulations. Note that the current example can be useful to other simulations that involve steps switch from AMBER to GROMACS (*see Note 21*). Herein the minimization and MD input files are adapted from the GROMACS tutorial for simulating lysozyme in water (<http://www.mdtutorials.com/gmx/lysozyme/index.html>). All these simulations were performed based on the GPU implementation of GROMACS (through the “-nb gpu” option) in the version of 2018.3 [19, 20].

1. Minimization:

```
$ gmx grompp -f min.mdp -c PHM_solv.gro -p PHM_solv.top -o em.tpr > em.out
$ gmx mdrun -v -deffnm em -nb gpu
```

2. 5 ns NVT equilibration:

```
$ gmx grompp -f nvt.mdp -c em.gro -r em.gro -p PHM_solv.top -o nvt.tpr > nvt.out
$ gmx mdrun -deffnm nvt -nb gpu
```

3. 5 ns NPT equilibration:

```
$ gmx grompp -f npt.mdp -c nvt.gro -r nvt.gro -t nvt.cpt -p PHM_solv.top -o npt.tpr > npt.out
$ gmx mdrun -deffnm npt -nb gpu
```

4. 20 ns NVT production:

```
$ gmx grompp -f md.mdp -c npt.gro -t npt.cpt -p PHM_solv.top -o md_0_1.tpr > md_0_1.out
$ gmx mdrun -deffnm md_0_1 -nb gpu
```

In the production simulation, we have 20 ns sampling using 2 fs time-step, with snapshots being saved every 10 ps.

5. Post-processing of the trajectory:

```
$ gmx trjconv -s md_0_1.tpr -f md_0_1.xtc -o md_0_1_noPBC.xtc -pbc mol -center
```

This command accounts for the periodicity and works interactively. We select the protein as the group for centering (option 1 in the first input), and then select the whole system as the group for output (option 0 in the second input). The `md_0_1_noPBC.xtc` file generated in this step will be used for the analysis.

6. Here we use the *cpptraj* program [34] in AmberTools to analyze the results. Simply run the command:

```
$ cpptraj -p PHM_solv.top -i cpptraj.in > cpptraj.out
```

It calculates the root-mean-square deviation (RMSD) values of the backbone C_{α} atoms relative to the crystal structure, the RMSD values of the heavy atoms in the metal site residues relative to the crystal structure, and the bond distances between the metal ion and its ligating atoms. These results are shown in Fig. 2. Even though these values are reasonable, the present tutorial is only for illustrative purposes and adjustments

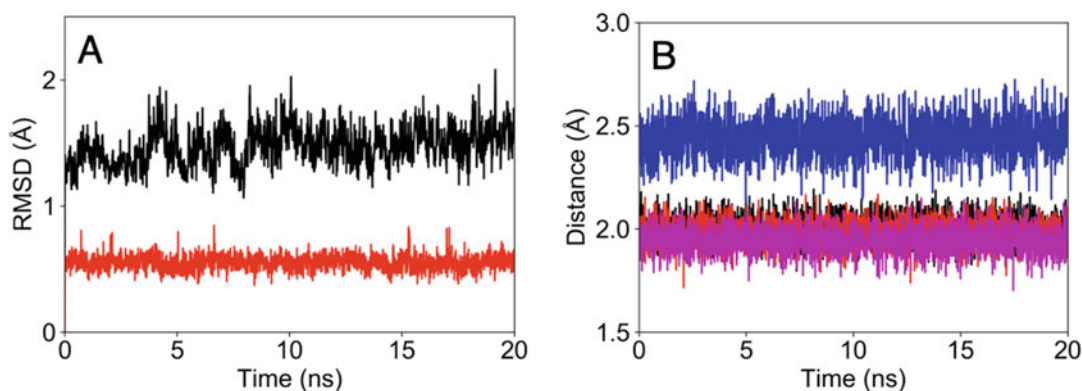


Fig. 2 (a) The RMSD values of the backbone C_{α} atoms (black) and heavy atoms in the metal site residues (red) during the production run; **(b)** The bond distances of metal-ligand bonds for the copper ion, the ligating atoms are HID242@NE2 (black), HID244@NE2 (red), MET314@SD (blue), and OXY360@O1 (magenta), according to the residue numbers in the PDB entry 1SDW

of the force field parameters may be necessary to reach the desired level of accuracy. In addition, even though one can successfully reproduce the present example, more hands-on practices may be needed to master the use of MCPB.py (*see Note 22*).

The following files are also attached in the Electronic Supplementary Materials: (1) the Gaussian fchk file after the last frequency calculation of the small model; (2) the Gaussian log file of the Merz–Kollman population analysis for the large model; (3) the LEaP input file; (4) the AMBER topology and coordinate files of the solvated system; (5) the ParmEd input file for the parameter check; (6) the cpptraj input file for data analysis.

4 Notes

1. The PDB file should follow the format of PDB version 3.0, which is consistent with the force field settings in the AMBER software package, while other formats can cause recognition errors of the atoms. If users want to use GROMACS for the simulations, there should not be discontinuous residues in this PDB file. Otherwise errors will be issued when using the GROMACS topology and coordinate files that are obtained based on the conversion of the AMBER topology and coordinate files. The PDB file should have the metal site in the correct protonation state, and cannot have two atoms sharing the identical atom name inside a certain residue. In the PDB file, the metal ions should be separated into independent residues, and have both the residue name and atom name as its element symbol with all letters capitalized. This rule applies for both isolated ions such as a “ZN” residue, and for embedded ions such as an iron inside a HEME group (where the iron should be separated from the HEME group into an independent residue). If there is more than one metal ion in the metal site, each ion should be separated into an independent residue. If there are ligands in the PDB file, each atom inside the ligands should have its atom name with the first letter corresponding to its element (e.g., “HB11” instead of “1HB1”), in order for MCPB.py to correctly identify the element.
2. PDB files can be readily downloaded from the RCSB website: <https://www.rcsb.org/>. The *pdb4amber* program in AmberTools can be used to “clean” the PDB file. Note that if alternative conformations exist in the PDB file, the *pdb4amber* program can be used to “clean” the PDB file by keeping the “A” conformation or the most populous alternative conformation. Alternative conformations may be available for the metal

site residues, so special attention should be paid when choosing the appropriate one. The *H++* webserver can be used to add the hydrogen atoms for the amino acids. Since the *H++* webserver will delete all the crystal waters, metal ions, or ligands in the PDB file, the hydrogen atoms of these fragments can be added by the *reduce* and *LEaP* programs in AmberTools. Because different pieces were handled separately in the process of adding hydrogen atoms, users need to correct the protonation states of the metal site residues afterwards if they are wrong. Then users need to combine these fragments into a single PDB file. Afterwards *pdb4amber* can be used to re-sequence the PDB file. The final PDB file should meet the criteria mentioned above.

3. Users need to prepare relevant mol2 files (and maybe frcmod files, *see Note 6*) for the “unnatural” residues. This is because MCPB.py only has the force field information such as atom types, partial charges, and force constants for the standard amino acids, which are available in the AMBER force fields. Therefore users need to provide the force field information for the “unnatural” residues by themselves. Based on this information (which are templates) and the quantum calculations, MCPB.py will generate the missing parameters for the metal site.
4. Based on the PDB file obtained, users can create independent PDB files for the “unnatural” residues (e.g., metal ions, ligands, and water molecules). Through these PDB files, users can generate the mol2 and frcmod files. The *antechamber* program in AmberTools can be used for this purpose. For residues that are identical except the coordinates (e.g., the water molecules in the PDB file), only one mol2 file is enough. One can choose to use the AMBER atom type or general AMBER force field (GAFF) [35] atom type when using *antechamber*. Which one to choose will depend on the specific case at hand. For example, for an organic ligand, the GAFF atom type is recommended, while for a water molecule (or a hydroxyl group) that is coordinated to the metal ion, the AMBER atom type is recommended. For consistency consideration, users may need to adjust the automatically assigned atom types according to the AMBER parameter files of the force field to be used.
5. The mol2 file of an “unnatural” residue should have the identical residue name and atom names that are used in the PDB file. This is because during the parameter assignment procedure, MCPB.py cannot match the atoms in the PDB file and the atoms in the mol2 file unless both their residue and atom names match. Based on these matches, MCPB.py will assign atom types and charges to the atoms inside the “unnatural”

residues according to the mol2 file. For an “unnatural” residue that is in the metal site, the total charge in its mol2 file should be physical—the initial partial charges do not need to be accurate (because during the RESP charge fitting of the metal site residues, these partial charges will be updated), but the total charge of each residue should be physical for the reason stated below (e.g., for a Zn^{2+} ion, its mol2 file should have a charge of +2; while for a Fe^{3+} ion, its mol2 file should have a charge of +3; for a ligating organic ligand, its mol2 file should have the total charge of the group (typically, 0, +1, or -1, etc.)). This is because MCPB.py determines the total charge of the metal site by adding up the charges of the metal site residues. Therefore charges of these mol2 files are crucial for MCPB.py to correctly assign the charges when generating the input files for quantum calculations, and when performing the subsequent RESP charge fitting. Incorrect assignments will cause errors such as exaggerated partial charges for the metal site residues after RESP charge fitting. For an “unnatural” residue that is outside the metal site, the partial charges will not be fitted again, so these charges should be accurate and ready for simulations.

6. The *frmod* files contain the force field parameters for the “unnatural” residues, and these parameters are missing in the AMBER force field (including the general AMBER force field, which can be specified through the *gaff* variable) that the users are going to use. If these parameters are already available in the AMBER force field, users do not need to provide any particular *frmod* files.
7. After obtaining the mol2 files, the *parmchk2/parmchk* program in AmberTools can be used to generate the *frmod* files. The *parmchk2* program has an improved algorithm for generating the missing parameters and is recommended. However, it is not guaranteed that *parmchk2* can handle all the cases that *parmchk* can handle. So when it fails, users can try *parmchk*. If users choose to use the AMBER atom type for a certain residue, they should indicate the path of the *parmfile* when generating the corresponding *frmod* file. This *parmfile* is the *dat* file in the $\$AMBERHOME/\text{dat}/\text{leap}/\text{parm}$ directory which corresponds to the force field to be used.
8. Users need to manually create an input file for MCPB.py, in which the PDB, mol2, and (if any) *fromod* files are indicated. This input file are utilized for all the four steps of the MCPB.py protocol. A number of variables are available for the input file, and users can consult the AMBER manual for details. Of these variables, the following three are indispensable:
 - (a) *original_pdb*, which is the PDB file MCPB.py uses;
 - (b) *ion_ids*, which is the atom ID of the metal ion in the

metal site. If there are more than one metal ion available, all of their atom IDs need to be provided in one line and separated by spaces; (c) `ion_mol2files`, which is the mol2 file for the metal ion(s) in the metal site. If the metal site has more than one metal ion, and these ions are the same element and have the same formal charge, only one single mol2 file is needed for all of them.

9. Users can manually add/delete any coordinating bonds by adding or deleting the “LINK” lines. MCPB.py will treat these bonds as metal-ligand bonds and will generate relevant parameters based on them.
10. Gaussian03, Gaussian09, and Gaussian16 are supported by the MCPB.py program released in AmberTools19. In terms of the GAMESS-US program, at least the version 2011.08.11(R1) is supported. Overall, any version of GAMESS-US [36] that has the same output formats of the Cartesian Hessian matrix and Merz–Kollman population analysis as 2011.08.11(R1) is supported.
11. If Gaussian was used to do the quantum calculations, the Z-matrix Hessian matrix will be saved in the output file of the frequency calculation, which will be used when generating the force field parameters based on the Z-matrix method. However, the Cartesian Hessian matrix is saved in the binary checkpoint file after the frequency calculation. One needs to convert it to a fchk file, which will be used when generating the force field parameters based on the Seminario method. In comparison, if GAMESS-US was used, the Cartesian Hessian matrix will be saved in the output file of the frequency calculation. Currently MCPB.py only supports force constant generation based on the Seminario method when using GAMESS-US. Users need to put the output or fchk file under the working directory of MCPB.py. By default, MCPB.py will try to find the needed file having the default name of `(group_name)_small_opt.fchk` or `(group_name)_small_fc.log`. If the file has a name different from the default name, users can tell MCPB.py the file name through the `--fchk` option (for the fchk file) or the `--logf` option (for the output file).
12. One can modify the quantum calculation input files according to the resources available and specific needs. For example, the number of processors, the memory requirement, the DFT functional, the basis set, whether to use a solvation model, etc. Note that MCPB.py assigns the multiplicity of 1 or 2 to the input files of quantum calculations, based on the number of electrons available. However, this assignment may not be correct, especially for transition metal containing systems, whose multiplicity can be complicated. Hence users may need to make

corrections to the multiplicity before carrying out the quantum calculations. An incorrect multiplicity can cause convergence failure and generate incorrect force field parameters. We emphasize that during the quantum calculations, one cannot rearrange the atom sequence in the quantum input files. This is because MCPB.py will generate the force field parameters and partial charges based on the atom sequence in the generated PDB files and fingerprint files created in the first step of the MCPB.py protocol. Hence any sequence mismatch between these files and the output files of the quantum calculations will cause errors. If users can only successfully perform the optimization by reordering some of the atoms, they need to order them back in the final force constant calculation to match the atom sequence in the PDB and fingerprint files. The same rule applies to the quantum calculation of the Merz–Kollman population analysis. Similarly, any addition or deletion of the atoms during the quantum calculations will cause errors in later steps. So it is important to make sure that all the files have atom sequences matching the original PDB file in order to prevent this kind of problem.

13. For the small model, note that full geometry optimization will only find a stationary point, which may be a saddle point but not a local minimum. However, it is required to have a local minimum for the frequency calculation. Hence users need to check the convergence and frequencies in the output file of the frequency calculation. The convergence criteria need to be satisfied and there should be no imaginary frequencies. Otherwise, one needs to further optimize the structure and redo a frequency calculation until these requirements are met. Afterwards, the generated fchk or output file can be used for modeling in the next step. If GAMESS-US is used for these calculations, one should manually copy the optimized coordinates into the input file for frequency calculation. This step is not necessary for Gaussian because the frequency calculation will read the charge, multiplicity, and coordinates from the checkpoint file generated by the geometry optimization.
14. For the Merz–Kollman population analysis of the large model, a full optimization is not recommended because of the computational cost. A single point calculation can be performed if the positions of the hydrogen atoms are reasonable. Otherwise a geometry optimization for the hydrogen atoms is recommended. These different options can be set through the “large_opt” variable.
15. By default, the equilibrium values of bonds and angles involving the metal ion will be calculated based on the coordinates used in the frequency calculation. If users want to get these numbers according to the PDB file of the small model, the

variable “xstru” should be set to 1 in the MCPB.py input file. Besides, step “2b” should be used when users want to generate a frcmol file that have all the parameters except the ones of bonds and angles involving the metal ion (e.g., if they want to fill out these parameters later by hand). In this case, quantum calculations of the small model are not necessary. If the force constants in the generated frcmol file are all reasonable, one can progress to the next step.

16. There are different options for the RESP charge fitting step that are about charge restraints on different groups, users can consult the AMBER manual for details. Users need to check whether the partial charges are reasonable in the generated mol2 files. Generally the partial charge of a metal ion should be between 0 and +2, and smaller than its oxidation state, even for a metal ion which has an oxidation state of +3 or +4. Too large of a partial charge in the generated mol2 files may indicate incorrect charge assignments in the original mol2 files. Users can consult **Note 5** in this book chapter on this issue. In order to prevent this problem, one needs to make sure the original mol2 files have the correct charge settings, along with the correct charge and multiplicity assignments for the quantum calculations.
17. The LEaP input file may need to be adapted for different situations. Users need to check the LEaP input file carefully and make corrections if necessary. Special attention should be paid to the “bond” commands. This input file is supposed to have the disulfide bond settings. The metal site residues should have correct connections with each other through the “bond” commands. Moreover, since the metal site amino acid residues were renamed (which will be treated as “unnatural” residues by LEaP) and there is no head or tail information in the generated mol2 files, their connections with the neighboring amino acids need to be set through the “bond” commands as well. These settings should be added into the LEaP input file if they are missing.
18. It is highly recommended to use a visualization program such as VMD to check the generated topology and coordinate files. VMD will indicate the bond connections based on the molecular topology; hence, one can see whether or not the metal site residues have the correct connections with each other and with their neighboring residues. If any connections are incorrect or missing, one needs to modify the LEaP input file and regenerate the topology and coordinate files. Note that using VMD to simply check the PDB file is not enough because VMD will assign the bond connections automatically based on the distances between atoms.

19. *ParmEd* in AmberTools can be used to numerically check the parameters. Specifically, the “printBonds,” “printAngles,” “printDihedrals,” and “printDetails” commands can be used for this purpose.
20. Before performing minimization and MD simulations in GROMACS, it is better to perform benchmark calculations for energies and forces in AMBER and GROMACS for comparison. Even though an exact match is not expected due to the different simulation settings, the deviations should not be significant.
21. Moreover, the current example can be helpful to switch any simulation from AMBER to GROMACS. For example, a NPT equilibration by AMBER followed by a production run by GROMACS. This tutorial will help users to take advantage of the functionalities in both software packages.
22. Hands-on practice is essential to grasp any simulation tool. There are a large number of parameters in a force field; hence, care is essential, especially for complicated systems containing transition metal ions. Even though MCPB.py can significantly decrease the human effort involved in the parameterization of a metal site, it is dangerous to treat it as a “black box.” In order to better grasp the MCPB.py program, practice is highly recommended. We recommend the users to explore the example in this book chapter as well as the tutorials online, for better understanding and grasp of the MCPB.py program.

Acknowledgments

We acknowledge Prof. Justin Lemkul (Virginia Tech) for the GROMACS tutorial for simulating lysozyme in water. We acknowledge the computational support from the High Performance Computing Center (HPCC) at the Institute for Cyber-enabled Research (iCER) at Michigan State University (MSU). Pengfei Li gratefully acknowledges financial support through Prof. Sharon Hammes-Schiffer by the National Institutes of Health (Grant Number GM056207).

References

1. Woodson SA (2005) Metal ions and RNA folding: a highly charged topic with a dynamic future. *Curr Opin Chem Biol* 9(2):104–109. <https://doi.org/10.1016/j.cbpa.2005.02.004>
2. Dupureur CM (2008) Roles of metal ions in nucleases. *Curr Opin Chem Biol* 12(2):250–255. <https://doi.org/10.1016/j.cbpa.2008.01.012>
3. Andreini C, Bertini I, Cavallaro G, Holliday G, Thornton J (2008) Metal ions in biological catalysis: from enzyme databases to general principles. *J Biol Inorg Chem* 13(8):1205–1218. <https://doi.org/10.1007/s00775-008-0404-5>
4. Waldron KJ, Robinson NJ (2009) How do bacterial cells ensure that metalloproteins get the correct metal? *Nat Rev Microbiol* 7

- (1):25–35. <https://doi.org/10.1038/nrmicro2057>
- Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 117(19):5179–5197
 - MacKerell AD, Bashford D, Bellott M, Dunbrack R, Evanseck J, Field MJ, Fischer S, Gao J, Guo H, Ha S (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102(18):3586–3616. <https://doi.org/10.1021/jp973084f>
 - MacKerell AD, Banavali N, Foloppe N (2000) Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* 56(4):257–265. [https://doi.org/10.1002/1097-0282\(2000\)56:4<257::AID-BIP10029>3.0.CO;2-W](https://doi.org/10.1002/1097-0282(2000)56:4<257::AID-BIP10029>3.0.CO;2-W)
 - Klauda JB, Venable RM, Freites JA, O'Connor JW, Tobias DJ, Mondragon-Ramirez C, Vorobyov I, MacKerell AD Jr, Pastor RW (2010) Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J Phys Chem B* 114(23):7830–7843. <https://doi.org/10.1021/jp101759q>
 - Cheatham TE, Case DA (2013) Twenty-five years of nucleic acid simulations. *Biopolymers* 99(12):969–977. <https://doi.org/10.1002/bip.22331>
 - Dickson CJ, Madej BD, Skjevik ÅA, Betz RM, Teigen K, Gould IR, Walker RC (2014) Lipid14: the amber lipid force field. *J Chem Theory Comput* 10(2):865–879
 - Li P, Merz KM (2017) Metal ion modeling using classical mechanics. *Chem Rev* 117(3):1564–1686. <https://doi.org/10.1021/acs.chemrev.6b00440>
 - Lin F, Wang R (2010) Systematic derivation of AMBER force field parameters applicable to zinc-containing systems. *J Chem Theory Comput* 6(6):1852–1870
 - Peters MB, Yang Y, Wang B, Füsti-Molnár L, Weaver MN, Merz KM Jr (2010) Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). *J Chem Theory Comput* 6(9):2935–2947. <https://doi.org/10.1021/ct1002626>
 - Li P, Roberts BP, Chakravorty DK, Merz KM Jr (2013) Rational design of particle mesh Ewald compatible Lennard-Jones parameters for +2 metal cations in explicit solvent. *J Chem Theory Comput* 9(6):2733–2748. <https://doi.org/10.1021/ct400146w>
 - Li P, Merz KM Jr (2014) Taking into account the ion-induced dipole interaction in the non-bonded model of ions. *J Chem Theory Comput* 10(1):289–297
 - Åqvist J, Warshel A (1990) Free energy relationships in metalloenzyme-catalyzed reactions. Calculations of the effects of metal ion substitutions in staphylococcal nuclease. *J Am Chem Soc* 112(8):2860–2868
 - Pang Y-P, Xu K, Yazal JE, Prendergast FG (2000) Successful molecular dynamics simulation of the zinc-bound farnesyltransferase using the cationic dummy atom approach. *Protein Sci* 9(10):1857–1865. <https://doi.org/10.1110/ps.9.10.1857>
 - Li P, Merz KM Jr (2016) MCPB.py: a Python based metal center parameter builder. *J Chem Inf Model* 56(4):599–604. <https://doi.org/10.1021/acs.jcim.5b00674>
 - van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. *J Comput Chem* 26(16):1701–1718. <https://doi.org/10.1002/jcc.20291>
 - Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4(3):435–447. <https://doi.org/10.1021/ct700301q>
 - Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. *J Comput Chem* 26(16):1668–1688. <https://doi.org/10.1002/jcc.20290>
 - Prigge ST, Eipper BA, Mains RE, Amzel LM (2004) Dioxygen binds end-on to mononuclear copper in a precatalytic enzyme complex. *Science* 304(5672):864–867
 - Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A (2005) H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* 33(Suppl 2):W368–W371. <https://doi.org/10.1093/nar/gki464>
 - Klinman JP (2006) The copper-enzyme family of dopamine β-monooxygenase and peptidyl-glycine α-hydroxylating monooxygenase: resolving the chemical pathway for substrate hydroxylation. *J Biol Chem* 281(6):3013–3016
 - Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical

- calculations. *J Mol Graph Model* 25 (2):247–260. <https://doi.org/10.1016/j.jmgm.2005.12.005>
26. Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham III TE, Cruzeiro VWD, Darden TA, Duke RE, Ghoreishi D, Giambasu G, Giese T, Gilson MK, Gohlke H, Goetz AW, Greene D, Harris R, Homeyer N, Huang Y, Izadi S, Kovalenko A, Krasny R, Kurtzman T, Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Man V, Mermelstein DJ, Merz KM, Miao Y, Monard G, Nguyen C, Nguyen H, Onufriev A, Pan F, Qi R, Roe DR, Roitberg A, Sagui C, Schott-Verdugo S, Shen J, Simmerling CL, Smith J, Swails J, Walker RC, Wang J, Wei H, Wilson L, Wolf RM, Wu X, Xiao L, Xiong Y, York DM, Kollman PA (2019), AMBER 2019, University of California, San Francisco
 27. Frisch M, Trucks G, Schlegel H, Scuseria G, Robb M, Cheeseman J, Scalmani G, Barone V, Petersson G, Nakatsuji H (2016) Gaussian 16, revision A. 03. Gaussian Inc, Wallingford, CT
 28. Besler BH, Merz KM Jr, Kollman PA (1990) Atomic charges derived from semiempirical methods. *J Comput Chem* 11(4):431–439
 29. Bayly CI, Cieplak P, Cornell W, Kollman PA (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J Phys Chem* 97(40):10269–10280
 30. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput* 11(8):3696–3713. <https://doi.org/10.1021/acs.jctc.5b00255>
 31. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926–935. <https://doi.org/10.1063/1.445869>
 32. Li P, Song LF, Merz KM Jr (2015) Systematic parameterization of monovalent ions employing the nonbonded model. *J Chem Theory Comput* 11(4):1645–1657
 33. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14(1):33–38. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5)
 34. Roe DR, Cheatham TE III (2013) PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* 9(7):3084–3095. <https://doi.org/10.1021/ct400341p>
 35. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. *J Comput Chem* 25(9):1157–1174
 36. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su S (1993) General atomic and molecular electronic structure system. *J Comput Chem* 14(11):1347–1363. <https://doi.org/10.1002/jcc.540141112>



Parameterization of Large Ligands for Gromacs Molecular Dynamics Simulation with LigParGen

Yu Wai Chen, Yong Wang, Yun-Chung Leung, and Kwok-Yin Wong

Abstract

Molecular dynamics (MD) simulation is a powerful method of investigating the interaction between molecular species. Defining the mechanical properties and topologies for all components involved is critical. While parameters for proteins are well established, those for the wide range of ligands and substrates are not. Here we introduce a very useful service which is designed for small organic molecules. We describe a protocol to extend this tool to beyond its current size (200 atoms) and formal charge (2+ to 2-) limits.

Key words Molecular dynamics (MD) simulation, Ligand parameters, Protein-ligand complex, OPLS-AA/L forcefield, Gromacs, Peptidoglycan glycosyltransferase, Moenomycin A, Fluorescein

1 Introduction

In recent years, molecular dynamics simulation of macromolecular systems has emerged as a powerful tool in structural analyses. In 2013, the Nobel Prize in Chemistry was awarded to Arieh Warshe, Michael Levitt, and Martin Karplus for their contributions in originating and advancing the art of computational simulation with molecular mechanics. Over the years, the vast improvement in both computer hardware and software helped to popularize these techniques. Nowadays, researchers can perform simulation of simple proteins on an average home computer.

Gromacs is one of the most popular simulation software [1], which is free for academic research. The project is under active development to take advantage of newer algorithms for use with a broad spectrum of hardware from personal computers to high-performance clusters.

Gromacs (see Note 1) has been applied to studying a wide range of biological phenomenon. Researchers are continually pushing the boundaries of what they can achieve. The software was developed initially for proteins, for which the chemical and structural properties (simulation parameters) are optimized. However, the same

cannot be said for other macromolecules (e.g., carbohydrates, nucleic acids, and lipids), small organic molecules, or modified amino acids. In the past, researchers have very limited means of progressing when they came across a species of which the simulation parameters are not known. Sometimes, the only way ahead was to use parameters based on similar structures guided by chemical intuitions.

In the past few years, several utilities have been developed to aid the parameterization of molecular species. These include the *LigParGen* [2], and the *ATB* [3], both are available as web servers. In this article, the use of the former on a 219-atom species will be described, as an example. The overall method is based on a *Gromacs* tutorial available on the *LigParGen* web server [4].

The biological system under investigation is the bifunctional *Staphylococcus aureus* protein, penicillin-binding protein 2 (PBP2), which is essential for cell wall synthesis. We are only interested in its peptidoglycan glycosyltransferase (PGT) domain. In our laboratory, we modified the PGT inhibitor, moenomycin A (MoeA; Fig. 1a), by replacing its A-ring with a label moiety (fluorescein with a C₄ linker) to become “F-4-MoeA.” The fluorescent properties of free and protein-bound F-4-MoeA are different, thus allowing the protein-inhibitor binding event to be monitored spectroscopically.

The purpose of this chapter is to demonstrate how to obtain reasonable parameters for the large ligand so that a simulation exercise can be set up to study the dynamics of the non-covalent protein:ligand (PGT:F-4-MoeA) complex. Our starting material is a crystal structure in the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) which consists of the full periplasmic body of PBP2 complexed with moenomycin A (PDB ID: 2OLV) [5]. We demonstrate how the ligand parameters are obtained with a 2-part divide-and-conquer strategy combining server-generated parameters with manual editing.

2 Software

All the software needed in this chapter, except the *ChemOffice Professional* suite, are free for academic or nonprofit use.

2.1 Web Server

1. *LigParGen* server (<http://zarbi.chem.yale.edu/ligpargen>).
2. To accompany its use, the web page for *Gromacs* tutorial and for downloading *Python* scripts and *Gromacs* input files is: http://zarbi.chem.yale.edu/ligpargen/gmx_tutorial.html.

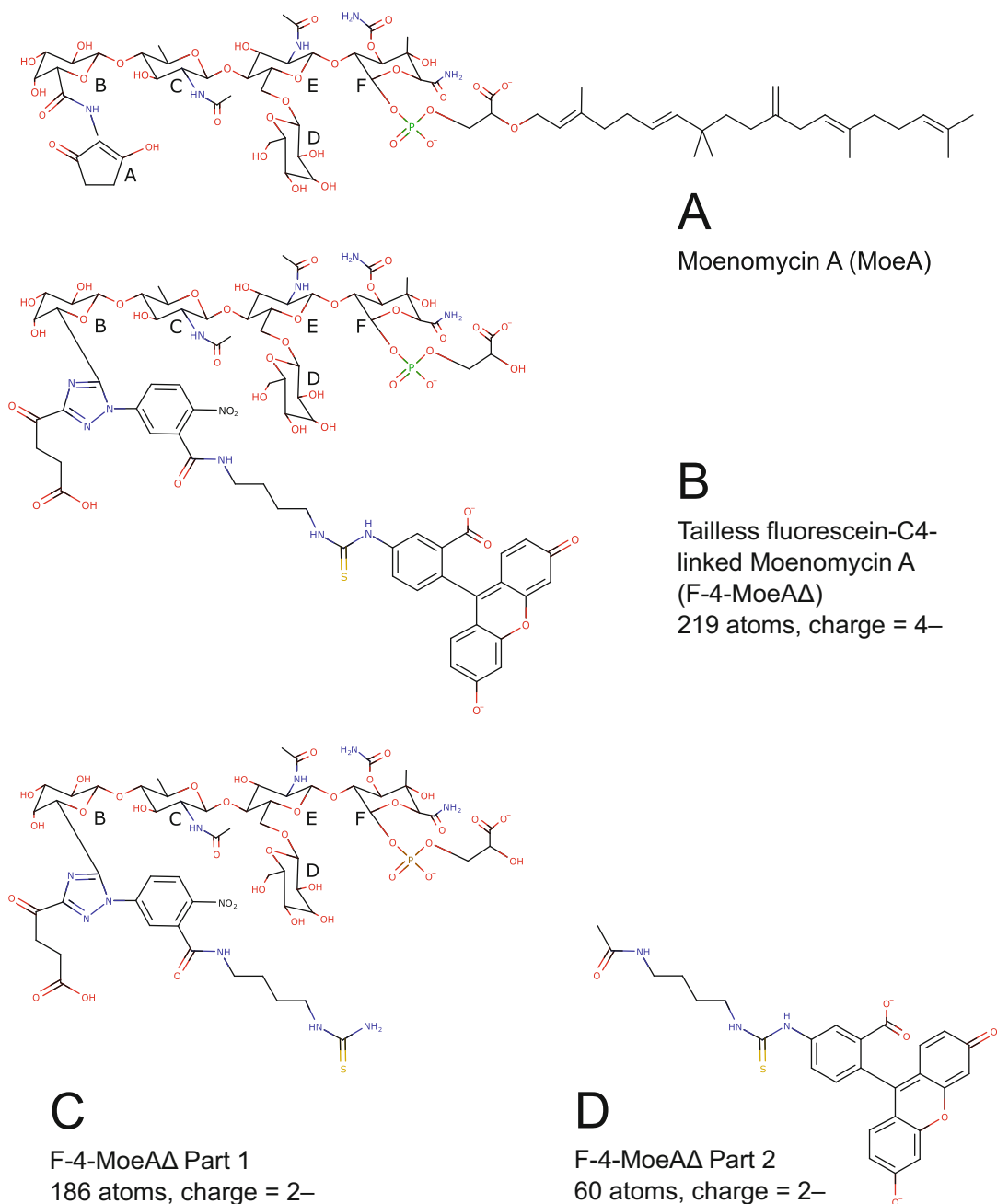


Fig. 1 Chemical structures of the ligand and its derivatives. This figure and Fig. 2 were prepared with *MarvinSketch* and *Inkscape*

2.2 Installed on Local Linux Computer

1. *Modeller* version 9.X (<https://salilab.org/modeller/>)—for comparative modeling of proteins from structure templates (*.pdb* files).
2. *Coot* (<https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/>)—for viewing structures (*.pdb* files) graphically and manipulation of atomic coordinates.

3. *Gromacs* version 4.6.7 (<http://www.gromacs.org>) (higher versions will need modifications of input described in this chapter).
4. *Open Babel* (<http://openbabel.org/>)—for interconversion of chemical structure in different formats.
5. *Avogadro* (<https://avogadro.cc/>)—for graphical visualization and energy minimization of structures (*.mol2* files).
6. *PyMOL* (<https://pymol.org/>)—for viewing structures (*.pdb* files) graphically and addition of missing hydrogen atoms.
7. *VMD* (<https://www.ks.uiuc.edu/Research/vmd>)—for viewing simulation structures (*.gro* files) graphically.

2.3 Commercial Software (Windows)

ChemDraw and *Chem3D* (*ChemOffice Professional* suite, PerkinElmer).

3 Protein Model

The protein structure needs to contain the full sequence and has all the hydrogen atoms. Residues 137–144 are missing in the crystal structure. The starting model of PGT is built using *Modeller* employing chain B of PDB ID: 2OLV and including other homologous PGT domain structures as reference templates (PDB IDs: 3VMA, 3FWL, 3NB6, 2OQO, 6FTB). The details of *Modeller* operations are beyond the scope of this chapter, and the readers are referred to its webpage for tutorials and instructions. Briefly, the following steps are performed:

1. Download all the crystal structures (*.pdb* files) from the RCSB PDB (<https://www.rcsb.org>).
2. Superimpose all PGT domain structures on a molecular graphics program (e.g. *Coot*) which allows the display of electron density maps generated from deposited data. By careful inspection, a multiple-sequence alignment was constructed which serves as the input for *Modeller* (see **Note 2**).
3. The side-chain rotamer conformations of the *Modeller*-best model are checked against the template structures with their respective electron density maps and adjusted if necessary.
4. The protonation states of essential residues are checked and revised if necessary (see **Note 3**). Name the protein-only coordinate file, e.g., *protein.pdb*.

4 Ligand Model

The ligand, moenomycin A, that is bound to *S. aureus* PBP2 in the crystal structure serves as the starting point. The C₂₅ lipid tail is not defined and the ligand will be kept truncated in the simulation because the tail is presumably associating with the membrane which will not be simulated. The goal of this step is to prepare a structure model of the “tailless” ligand, F-4-MoeAΔ (Fig. 1b). First, we shall produce a purely geometric 3D model from drawing out its chemical structure (Subheading 4.1). This will then be used to generate atomic coordinates of the parts of the ligand which was chemically extended (fluorescein plus linker) on top of the MoeA framework (Subheading 4.2).

4.1 Preparation of a Complete Geometric Model

1. The 2D structure of F-4-MoeAΔ is drawn in *ChemDraw* (*ChemOffice Professional*), and its 3D coordinates are generated by *Chem3D* (*ChemOffice Professional*), saved into the *mol2* format (e.g. *F_4_MoeA.mol2*) and passed to *Avogadro* for energy minimization.
2. The *mol2*-format structure was used to generate a coordinates file in *pdb* format using *Open Babel*. In the *Linux* environment, the command is:

```
$ babel -imol2 F_4_MoeA.mol2 -opdb F_4_MoeA.pdb
```

The atoms in this file do not have unique atom names (they only have element types: N, O, C, etc.) nor a residue type (default residue is “***”).

4.2 Manual Rebuilding of Structural Model

The MoeA framework from the crystal structure needs to be manually rebuilt with the fluorescein moiety added to form the complete ligand such that, when complexed with the protein, the ligand structure makes chemical sense and does not lead to major steric clashes. Rebuilding consists of rounds of manipulation of the model coordinates in a molecular graphics program (*Coot*), exporting the moved coordinates, and grafting of the exported (moved) coordinates into the starting model of that round.

1. Starting with the known MoeA coordinates (residue “M0E B 901” of PDB ID: 2OLV), several atoms corresponding to the A-ring of MoeA (atom names: CCM, OCQ, NCS, CCT, CCU, OCV, CCW, CCX, CCY, OCZ) are deleted. These atoms will be replaced by the linker atoms (Fig. 1a, b).
2. The linker and fluorescein parts will be added. The *Coot* function, “Rotate Translate Zone/Chain/Molecule”, is used for adjusting rotatable bonds so that the whole ligand fits into the binding site (*see Note 4*). Each round of these manipulations

results in rotation along one bond. It may take several rounds to achieve the final model. In the case of F-4-MoeA Δ , it took five rounds.

3. The new atomic coordinates are revised so that each has a unique atom name. The residue name of all atomic coordinates is to be the same as that of the MoeA (i.e., “MOE B 901”).
4. The model is examined in the graphics program *PyMOL*, and hydrogen atoms are added and carefully examined and revised according to chemical knowledge (*see* **Note 5**).
5. The ligand, F-4-MoeA Δ , consists of 219 atoms and exceeds the limit of 200 atoms imposed by the *LigParGen* server. In addition, the ligand has a formal charge of 4[−] (2[−] on the fluorescein dianion, and 2[−] on the MoeA framework; Fig. 1b). The *LigParGen* server imposes a maximum charge of from 2⁺ to 2[−]. The solution is to upload the ligand in two parts and combine them afterward.

5 Ligand Parameterization

5.1 Generating Parameters in Two Parts

1. The ligand coordinate *pdb* file is divided into two parts, with extensive overlap in the C₄ linker (Fig. 1c, d). Each part has fewer than 200 atoms and a charge of 2[−] (*see* **Note 6**).
2. The complete ligand F-4-MoeA Δ *pdb* file is edited into the two part-*pdb* files by deleting unneeded atom lines. Unfulfilled coordination at the new termini is filled with hydrogen atoms using *PyMOL*.
3. The two files (*part-1.pdb* and *part-2.pdb*) are submitted separately to the *LigParGen* server. Since both parts carry charges, the 1.14*CM1A charge model is employed. On completion, the two parameterization runs will produce the sets of *.itp* and *.gro* files (parameters) for *Gromacs*.

5.2 Merging Geometry Parameters (.gro)

1. First, we need to identify a group of atoms which have similar partial charges in both parts to form the merge junction. The two *.itp* files (topology) are examined. Under the [atoms] section, all atoms are listed with their partial charges. In this case, the −NH−CS− atoms (H2C, N28, C2B, S2E of Part 1 and their equivalents, H1C, N0O, C0M, S0N of Part 2) have similar charges (Fig. 2a).
2. The coordinates of the two parts cannot be simply merged as they are from individual *LigParGen* runs. With both loaded into *Coot*, the base structure (*part-1.pdb*) is kept stationary, and the fluorescein part (*part-2.pdb*) is moved so that the junction atoms overlap in space; then the new coordinates of part-2 atoms are exported.

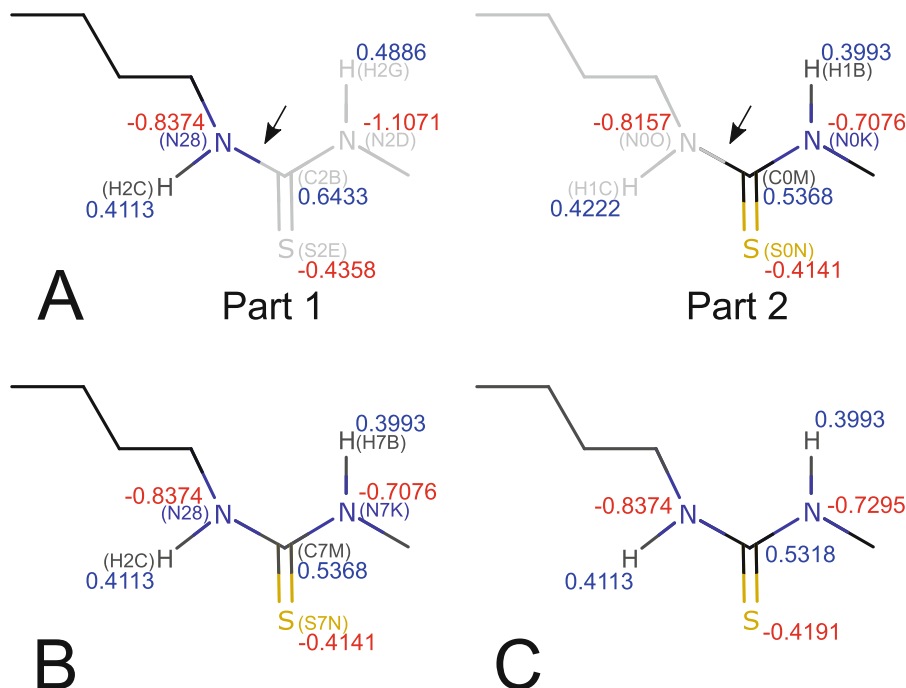


Fig. 2 Details of atomic charges at the merge junction. Positive charges are in blue and negative charges are in red. Bracketed three-character atom names are shown next to element symbols. (a) The two parts that form the complete ligand with the junction indicated by an arrow: Part 1 (all atoms to the left of the arrow) and Part 2 (all atoms to right of the arrow). The parts that are not used are in grey: these atoms and their parameters are deleted. (b) The merged model and its atomic charges. (c) The final model with rebalanced charges

- Before merging, make sure all (three-character) atom names are unique by checking and renaming if needed, in *part-2.gro* (see **Note 7**; compare Fig. 2a, b).
- Concatenate the *part-1.gro* and *part-2.gro* into a new file called *FOM.gro*. The new ligand “residue” is referred to as “FM0” (for “fluorescein-labeled moenomycin”) throughout the simulation.
- Delete unwanted redundant atom lines from each *.gro* file (Fig. 2a).
- In *FOM.gro*, all the part-2 XYZ coordinates need to be updated. These values are extracted from the *Coot*-moved *part-2.pdb* and divided by 10 so that they are in nm (for *Gromacs*).
- Revise the total number of atoms in *FOM.gro* (for F-4-MoeAΔ, it is 219).
- Edit all residues (first column) to be “1FM0” (residue number 1, residue name “FM0”).

5.3 Combining Topology Parameters (.itp)

1. The two *.itp* files (topology) are edited separately before merging.
2. Edit *part-1.itp*: [*atomtypes*] and [*atoms*] sections: comment out redundant atom entries (Fig. 2a). *LigParGen* assigns each atom with an individual atomtype in the format *opls_nnnn* where *nnnn* is a unique number (e.g. *opls_883*). Note that each atom is assigned an “atom number” (first field in [*atoms*]; see **Note 8**). All the subsequent geometric parameters are defined using this “atom number.”
3. Edit *part-1.itp*: In the geometry definition sections: [*bonds*], [*angles*], [*dihedrals*], [*pairs*], identify all the lines which contain any of the redundant atoms (identified by their “atom number”) and comment all out.
4. Edit *part-1.itp*: In the geometry definition sections: [*bonds*], [*angles*], [*dihedrals*], [*pairs*], identify those entries which contain the junction atoms and mark down for further editing.
5. Edit *part-2.itp*: [*atomtypes*] and [*atoms*] sections: comment out redundant atom entries (Fig. 2a). All “*opls_nnnn*” atomtypes will be clashing with those in *part-1.itp*. Give them unique new numbers (e.g. change *opls_800* – *opls_859* to *opls_9800* – *opls_9859*). All “atom numbers” are also clashing so rename them from 1, 2,... to 201, 202,... (but also see **Note 8**).
6. Edit *part-2.itp*: In the geometry definition sections: [*bonds*], [*angles*], [*dihedrals*], [*pairs*], identify all the lines which contain any of the redundant atoms (identified by their “atom number”) and comment all out. Rename all atom names to those assigned in Subheading 5.2, **step 3**.
7. Edit *part-2.itp*: In the geometry definition sections: [*bonds*], [*angles*], [*dihedrals*], [*pairs*], identify those entries which contain the junction atoms and edit them. The part-2 atom numbers should be associated with the respective part-1 atom numbers to recreate the proper geometric definitions.
8. Merge the two edited files, *part-1-edited.itp* and *part-2-edited.itp*, into a new file called *FOM.itp* (Fig. 2b). Make the new residue, “FM0”, in the [*atoms*] section of *FOM.itp*. Check all atom numbers (see **Note 8**).
9. The partial charges of atoms at and near the junction are checked and manually revised so that the overall charge of the F-5-MoeΔ ligand is an integer (−4) (Fig. 2c).

5.4 Rebuilding Coordinates (.pdb)

1. Merge the two files, *part-1.pdb* and *part-2.pdb*, into a new file called *FOM.pdb*.
2. Delete redundant atom lines (Fig. 2a).

3. Rename all part-2 atom names to those assigned in Subheading 5.2, step 3.
4. Name the new residue as “A chain”, residue number 1, with residue name “FM0.”
5. Change all atom lines to begin with “HETATM” (instead of “ATOM”) to conform to the PDB convention for ligands.

6 Running the MD Simulation (Using the Parameters)

6.1 Preparation

Set up the simulation box and the solvation system (*see Note 9*). *Python* scripts are downloaded from the *LigParGen* tutorial for *Gromacs* webpage.

1. Prepare the coordinates file for the complex with the help of a *python* script. The OPLS-AA/L forcefield must be chosen for compatibility.

```
$ pdb2gmx -f protein.pdb -o protein.gro -water tip3p -ignh
$ python combineGro_prot_lig.py protein.gro FM0.gro > cpx.gro
```

Edit *topol.top* (output of *pdb2gmx*) and do (a) add the line “#include FM0.itp” below the forcefield definition “#include oplsaaff/forcefield.itp”; (b) at the end of the file, under the [molecules] section, add a molecule type “FM0 1”.

2. Prepare the simulation environment: a cubix box with 1-nm-thick walls (*see Notes 10 and 11*).

```
$ editconf -f cpx.gro -o cpx_box.gro -c -d 1.0 -bt cubic
$ genbox -cp cpx_box.gro -cs spc216.gro -o cpx_box_W.gro -p topol.top
```

3. Set up the system with 0.15-M ions and to neutralize the charges on the protein:ligand complex.

```
$ grompp -f ions.mdp -c cpx_box_W.gro -p topol.top -o ions.tpr -maxwarn 2
$ genion -s ions.tpr -o cpx_box_Wi.gro -p topol.top -neutral -conc 0.15
```

(Add the ions to the SOL molecule group in response to this command.)

6.2 Energy Minimization (See Notes 10–12)

```
$ grompp -f em.mdp -c cpx_box_Wi.gro -p topol.top -o em.tpr -maxwarn 2
$ mdrun -v -defnm em
```

6.3 Equilibration, Production (See Notes 10–12)

Once energy minimization has been run successfully, the MD system is set up properly. The remaining of the MD steps: equilibration in NVT and NPT ensembles with restraints and the unrestrained production MD will run. The details of these steps are beyond the scope of this chapter. A snapshot of the PGT:F-4-MoeAΔ complex after a 10-ns MD run is shown in Fig. 3.

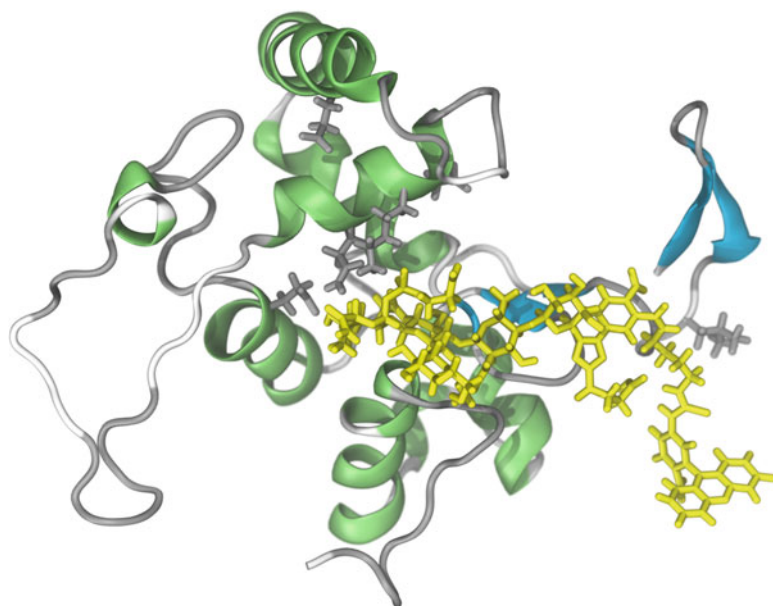


Fig. 3 The model of the PGT:F-4-MoeA Δ complex. This is the structure of the last frame (1001) of the production MD simulation. The protein is shown in secondary-structure cartoon with helices, strands and loops colored in green, cyan, and gray, respectively. The residues which are important in interacting with MoeA are shown with their side-chain atoms and colored in gray. The structure of F-4-MoeA Δ is colored in yellow. This figure is produced with *VMD* and *ImageMagick*

7 Notes

1. Text format conventions: computer software, file type, and filenames are in italics; commands entered in Linux prompt are boxed and in monospaced fonts.
2. The most time-consuming step for *Modeller* rebuilding is the construction of the template sequence alignment. The quality of the model produced will be dependent on the quality of the manual structure-based alignment. Automated sequence-based alignment is generally inferior.
3. Protonation states of protein side chains are often not known in crystal structures except those of very high resolution. Sometimes, protonation states can be inferred from known chemical properties.
4. Manual rotation (dihedral angle adjustment) in *Coot*. The non-standard geometry of the unknown ligand is not understood by *Coot*. Therefore, the “Edit χ Angles” or “Torsion General”

(edit dihedral angles) tools cannot be used. We did not go down the route of defining all the parameters for the rebuilding of the ligand. Instead, we employed a “brute-force” manual rebuilding by eye judgment and chemical intuition. In a typical round, a ligand model (starting coordinates set) was loaded twice into *Coot* so that we have two molecules: one for visual reference (coordinates will be fixed); the other for rebuilding (coordinates will be revised). Rotation along a bond is achieved by carrying out manual “Rotate Translate Zone/Chain/Molecule” actions while keeping the bond being rotated aligned in both molecular objects. The “rotated” (moved) coordinate set is exported. The new “rotated” atomic coordinates on the “moved” side of the bond are grafted into and replace the equivalent coordinates of the “starting” set.

5. Checking the protonation states of chemical groups is important. In this case, F-4-MoeAΔ has four ionizable groups, two on fluorescein, one on the phosphate, and one on the terminal carboxylate distal to the fluorescein. Hydrogen atoms on these groups are deleted.
6. Due to the maximum charges allowed on the *LigParGen* server, it may be necessary to keep some ionizable groups protonated. In the case of F-4-MoeAΔ, a terminal carboxylic acid has been kept protonated (Fig. 1c) to allow the other two groups, which play essential roles in protein binding, to be ionized.
7. Use a convention to rename part-2 atoms so that they are unique as well as can be easily identified. For example, change all middle characters from “0” to “7” and from “1” to “8” (i.e. atom “C0X” becomes “C7X”; “N0K” becomes “N7K”; “O11” becomes “O81”). Of course, one already checked that none of the part-1 atoms has “7” or “8” as its middle character.
8. It is crucial to note that the atom numbering in the final *FM0.itp* file has to start from 1 and is continuous without gaps. Otherwise, *Gromacs* will not run.
9. Use TIP3P water for the OPLS-AA/L forcefield.
10. *ions.mdp*, *em.mdp*, *nvp.mdp*, *npt.mdp*, *md.mdp* can be downloaded from “http://zarbi.chem.yale.edu/ligpargen/gmx_tutorial.html”
11. Use “vdw-type = Cut-off” and “cutoff scheme = Verlet” in all *.mdp* files to avoid getting fatal error due to large charge groups.
12. Include “FM0” as one of the energy groups (e.g., in *em.mdp*, “energygrps = Protein FM0”).

Acknowledgments

We acknowledge support from the Innovation and Technology Commission of Hong Kong, the Hong Kong Polytechnic University and the Life Science Area of Strategic Fund 1-ZVH9.

References

1. Berendsen HJC, Vandespoel D, Vandrunen R (1995) Gromacs - a message-passing parallel molecular-dynamics implementation. *Comput Phys Commun* 91(1–3):43–56
2. Dodda LS et al (2017) LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res* 45(W1):W331–W336
3. Malde AK et al (2011) An automated force field topology builder (ATB) and repository: Version 1.0. *J Chem Theory Comput* 7(12):4026–4037
4. Jorgensen WL (2019) LigParGen: OPLS/CM1A parameter generator for organic ligands. <http://zarbi.chem.yale.edu/ligpargen/>. Accessed 22 Mar 2019
5. Lovering AL et al (2007) Structural insight into the transglycosylation step of bacterial cell-wall biosynthesis. *Science* 315(5817):1402–1405



Chapter 17

Simulation of Proteins Modified with a Fluorescent Label

Zoe Chan and Yun-Chung Leung

Abstract

Fluorescent labeling of protein has been widely used in microbiology for detection and analysis. Molecular dynamics simulations provide vital supporting information for predictions and interpretations of experimental results. While force fields for proteins with regular amino acids are readily available, parameters for covalently attached fluorophores have to be incorporated into these force fields before they can be used for simulations. In this chapter, we shall discuss the methods to parameterize a fluorescent probe (fluorescein) attached to a cysteine, as a modified residue, for performing simulations with *GROMACS*.

Key words Molecular dynamics simulation, Parameterization, Mutated amino acid residue, Fluorescent probe, CHARMM force field, GROMACS, Fluorescein

1 Introduction

Fluorescent probe is commonly used in microbiology for analytical and detection purposes [1–5]. A fluorophore can be attached in the vicinity of the active site of an enzyme to detect enzyme–substrate complex formation. Binding of a substrate or an inhibitor alters the microenvironment of the fluorophore and induces a change in the fluorescence signal, allowing fast detection and screening of potential target substrate or inhibitors. For fast and accurate detection, the fluorescent probe should be attached to a residue close to the binding pocket with minimal hindrance to the enzymatic function. MD simulations can be used to study different conditions and combinations of fluorophores at various locations for potential designs of the sensor. The experimental results can also be studied against MD simulations for more thorough understandings and interpretations. However, the common fluorophores, such as fluorescein and rhodamine, are not normally incorporated in currently available force fields for MD simulations.

Force fields include all the parameters necessary in simulations, such as atomic charges, bond lengths, and bond angles. All the calculations are done based on these parameters. Therefore, the

parameters of the fluorescent probe have to be included in the force field used.

In this chapter, a step-by-step method for parameterizing a fluorescent probe, fluorescein-5-maleimide, attached to a cysteine as a mutated residue and incorporating it into CHARMM27 force field will be discussed. This is one of the most popular ways of fluorescent labeling a protein target to make a biosensor. Some solutions to common errors that may occur during simulations with this mutated residue will be discussed in the hope of helping fellow users to tackle the known issues.

2 Software and Protein Structure

2.1 Software

All the software used in this chapter are free for academic purposes and are installed on a Linux computer with Ubuntu version 16.04 (*see Note 1*).

1. *ACPYPE* (or AnteChamber PYthon Parser interfacE) [6]—for parameterization.
2. *Antechamber* version 17.3 [7] from *AmberTools17* package.
3. *Coot* (Crystallographic Object-Oriented Toolkit) version 0.8.8 [8]—for protein structure visualization.
4. *GROMACS* version 4.6.7 [9–11].
5. *JLigand* version 1.0.40 [12]—for fluorophore-amino acid complex drawing.
6. *Open Babel* version 2.3.2 [13]—for interconversion of complex structure in different formats.
7. *VMD* version 1.9.3 [14]—for MD simulation structures visualization.

2.2 Protein Structure

The fluorophore-protein complex structure was modeled based on the well-defined crystallographic structures of BlaC β -lactamase obtained from the Protein Data Bank [15] (PDB ID: 3CG5 [16]). Make sure there are no missing residues in the crystal structure (*see Note 2*).

3 Fluorescent Probe Structure Building and Optimization

1. The structure of fluorescein-5-maleimide-attached cysteine is built using *JLigand*. The mutated amino acid, CMF (Cysteine-Maleimide-Fluorescein), is constructed by first drawing the covalent structure, then regularizing the structure to check the chemistry and minimize steric hindrance.

2. The atom names can be changed by double-clicking the atom and edit the “Atom Id”. Some atom names may cause errors in *GROMACS*, refer to Subheading 8.1 for more details.
3. R and S stereoisomers can be controlled at each chiral center by double-clicking the corresponding atom to open the “Edit Atom Details” tab.
4. The coordinates and bond information of the residue is then saved in PDB and CIF formats. The coordinate file can be edited manually by a plain-text editor to remove all the non-aromatic hydrogen atoms for a clean visualization.
5. Check the structure on *Coot* for connectivity and torsional, steric, and angular strains before proceeding to the next step.

4 Parameterization of the Fluorescent Label Using ACPYPE

1. The topology file can then be generated from the coordinate file through *Antechamber*, a program in the *AmberTools* package. It gives the parameters compatible with AMBER, CHARMM, and OPLS force fields.

```
$ acpype -i CMF.pdb
```

2. If the molecule is charged, for example carrying a total charge of -2 , they should be specified by “ $-n$ ”

```
$ acpype -i CMF.pdb -n -2
```

3. Check for abnormalities in the charges of the generated topology files using a plain-text editor and the coordinates on *Coot*. Abnormalities are likely caused by the misidentification of atom and bond types or missing atoms, especially hydrogen atoms. See Subheading 8.2 for more details.
4. Carefully check the atom and bond types in the MOL2 file as they greatly affect the charge calculation. SYBYL-format atom and bond types used in the MOL2 file are listed in Tables 1 and 2, respectively.
5. Remove any extra hydrogen (e.g., ionizable hydrogens) from the MOL2 file, and strategically change the charge to maintain the correct overall charge.
6. Check the total number of atoms and bonds.
7. Save the MOL2 file after modification and use it to generate the topology files.

```
$ acpype -i CMF_modified.mol2 -n -2
```

Table 1
SYBYL atom types

Hydrogen	H
Carbon sp ³	C.3
Carbon sp ²	C.2
Carbon sp	C.1
Carbon aromatic	C.ar
Carbocation (guanidinium)	C.cat
Nitrogen sp ³	N.3
Nitrogen sp ²	N.2
Nitrogen sp	N.1
Nitrogen aromatic (pyridine)	N.ar
Nitrogen amide	N.am
Nitrogen trigonal planar (nitro, pyrrole)	N.pl3
Nitrogen positively charged sp ³ (lysine)	N.4
Oxygen sp ³	O.3
Oxygen sp ²	O.2
Oxygen in carboxylates and phosphates	O.co2
Sulfur sp ³	S.3
Sulfur sp ²	S.2
Sulfoxide sulfur	S.o
Sulfone sulfur	S.o2
Phosphorus sp ³	P.3
Halogens and metals	Element symbol (F, Cl, Ca, Zn, etc.)

Table 2
SYBYL bond types

Single	1
Double	2
Triple	3
Aromatic	ar
Amide	am
Delocalized (carboxylate, guanidinium)	ar

8. A subsidiary subdirectory, `CMF.acpype/`, should be created by *ACPYPE*. It contains the coordinate files of the residue in MOL2 and PDB format, and the corresponding files for AMBER, CHARMM, GROMACS, and OPLS force fields.
9. The `CMF_GMX` files are generated for *GROMACS*, the GRO file (`CMF_GMX.gro`) contains the molecular structure in Gromos87 format; the TOP file (`CMF_GMX.top`) is the topology file. The ITP file (`CMF_GMX.itp`) includes all the parameters needed: atomic charges, bond types, angles, and improper and proper dihedrals.

5 Incorporation of the Fluorophore Parameters into the CHARMM27 Force Field

5.1 Setting Up the Force Field

1. All the force field subdirectories are located under the directory (*see Note 3*):
`gromacs/top/`, each with the extension “.ff”. All the parameters required are found in the force field subdirectory.
2. Duplicate the entire force field subdirectory for the target force field before customizing it (*see Note 4*). At the `gromacs/top/` directory:

```
$ cp --r charmm27.ff charmm27mod.ff
```
3. From here on, work in the `gromacs/top/charmm27mod.ff` subdirectory (*see Note 6*).
4. CMF is treated as a mutated residue, and therefore, must be specified in the force field file. The following files (Subheadings 5.2–5.6) must be modified to incorporate the parameters calculated using *ACPYPE* into the topology.

5.2 forcefield.doc

`gromacs/top/charmm27mod.ff/forcefield.doc`

1. To distinguish the modified force field from the original in simulation, rename the first line of the `forcefield.doc`. In this example, the force field was named “CHARMM27-CMF all-atom force field with CMF (with cmap) --version 2.0.”

5.3 residuetypes.dat

`gromacs/top/residuetypes.dat`

1. For the program to recognize CMF as an amino acid, it has to be specified in `residuetype.dat`.
2. The five recognized types are: **Protein**, **DNA**, **RNA**, **Water**, and **Ion**.
3. It is recommended to copy the `residuetypes.dat` file into the working directory and make any amendment on the copy instead of the original file.

5.4 aminoacids.rtp gromacs/top/charmm27mod.ff/aminoacids.rtp

```

[ CMF ]
[ atoms ]
      N      NH1      -0.529001      0
      HN      H       0.359800      1
      CA      CT1      0.018500      2
[ bonds ]
      CB      CA
      OG      CB
[ impropers ]
      N      -C      CA      HN
      C      CA      +N      O
[ cmap ]
      -C      N      CA      C      +N

```

1. Edit the aminoacids.rtp file (rtp = residue topology parameter file) according to the CMF_GMX.itp file generated by *ACPYPE*.
2. The first line [CMF] is the residue name.
3. [atoms] gives the atom names in the first column; atom types in the second column; their charge in the third column; and the sequential number on the last column starting from 0.
4. The bonds and improper angles are listed using the atom names under [bonds] and [impropers], respectively. Note that “-C” and “+N” are the backbone carbon and nitrogen of the connecting residues.
5. [cmap] is unique in the CHARMM force field. It is a grid-based energy correction map on the ϕ/ψ torsion angles in the Ramachandran space.
6. Make sure all the atom names match the atom names in the PDB file (*see Note 5*). The error may arise from atom names, *see Common Errors in Subheading 8.1*.

5.5 aminoacids.hdb gromacs/top/charmm27mod.ff/aminoacids.hdb

```

CMF      3
1        1      HN      N      -C      CA
1        5      HA      CA      N      C      CB
2        6      HB      CB      OG      CA

```

Table 3
Hydrogen types in HDB file [17]

Hydrogen types		
1	1 Planar H	Rings or peptide bond
2	1 Single H	Hydroxyl
3	2 Planar H	Ethylene sp^2 $R=CH_2$, or amide $-C(=O)NH_2$
4	2/3 Tetrahedral H	sp^3 $-CH_3$
5	1 Tetrahedral H	sp^3 R_3CH
6	2 Tetrahedral H	sp^3 R_2CH_2
7	2 Water H	
10	3 Water H	
11	4 Water H	

1. Hydrogen atoms absent in the protein structure will be added by the “`pdb2gmx`” program (part of *GROMACS*) in the simulation according to `aminoacids.hdb`.
2. If the hydrogens are omitted in the input structure of the modified residue, include the hydrogen information of the fluorescent-labeled residue in the HDB file.
3. The first row shows the name of the residue, *CMF*, and the number of lines describing this residue, three lines following have to be read in this case.
4. The first column is the number of hydrogen atoms to be added; the second line is the type of hydrogen atom (*see* Table 3). The third column is the name of hydrogen atoms added, and if there is more than one hydrogen atom to be connected, for example on row 4, the new hydrogen atoms will be named HB1 and HB2 [17]. The fourth column is the atom the new hydrogen atoms to be connected to, and two to three control atoms are included in the following columns to fix the orientations.

5.6 *ffbonded.itp*

`gromacs/top/charmm27mod.ff/ffbonded.itp`

The atomic interactions are simulated based on the parameters described in the force field. The parameters for bonded interactions, including bond length, angles, and proper and improper dihedrals, are included in `ffbonded.itp`. If the parameters are not already included in the force field, they have to be added by editing the `ffbonded.itp` according to the `CMF_GMX.itp` file.

5.6.1 Bonds

A. Bonds

```
[ bondtypes ]
;i j func b0 k
;###CMF
S CP2 1 0.18392 180670
```

1. [bondtypes] specifies the bonds through defying the bond length b_0 and the force constant k .
2. Both parameters can be found in CMF_GMX.itp file under section [bonds], where the column after the semicolon is the description of the corresponding atom.
3. The default function in the CHARMM27 force field is 1. Other function of bond types can be found in Table 4.

5.6.2 Angles

```
[ angletypes ]
;i j k func th0 cth ub0 cub
;###CMF
O P5 O 5 1.1580e+02 3.8351e+02 0.0 0.0
```

Table 4
Bond types in `ffbonded.itp` [17]

Name of interaction	Function type	Parameters (units)
Bond	1	b_0 (nm); k_b (kJ/mol/nm ²)
G96 bond	2	b_0 (nm); k_b (kJ/mol/nm ⁴)
Morse	3	b_0 (nm); D (kJ/mol); β (nm ⁻¹)
Cubic bond	4	b_0 (nm); $C_{i=2,3}$ (kJ/mol/nm ^{<i>i</i>})
Connection	5	
Harmonic potential	6	b_0 (nm); k_b (kJ/mol/nm ²)
FENE bond	7	b_m (nm); k_b (kJ/mol/nm ²)
Tabulated bond	8	Table number (≥ 0); k (kJ/mol)
Tabulated bond ^a	9	Table number (≥ 0); k (kJ/mol)
Restraint potential	10	Low, up_1 , up_2 (nm); k_{dr} (kJ/mol/nm ²)

^aNo connection, and so no exclusions, are generated for this interaction

Table 5
Angle types in `ffbonded.itp` [17]

Name of interaction	Function type	Parameters (units)
Angle	1	θ_0 (deg); k_θ (kJ/mol/rad ²)
G96 angle	2	θ_0 (deg); k_θ (kJ/mol)
Cross bond-bond	3	$r_{1e}r_{2e}$ (nm); $k_{rr'}$ (kJ/mol/nm ²)
Cross bond-angle	4	$r_{1e}r_{2e}r_{3e}$ (nm); $k_{r\theta}$ (kJ/mol/nm ²)
Urey–Bradley	5	θ_0 (deg); k_θ (kJ/mol/rad ²); r_{13} (nm); k_{UB} (kJ/mol/nm ²)
Quartic angle	6	θ_0 (deg); $C_{i=0,1,2,3,4}$ (kJ/mol/rad ⁱ)
Tabulated angle	8	Table number (≥ 0); k (kJ/mol)

- Angles are defined in `[angletypes]`. The default function Urey–Bradley vibration includes a harmonic potential (angle θ^0 (th0) and constant k^θ (cth)) and a harmonic correction term (bond r^0 (ub0) and constant k^{UB} (cub)).
- However, the angle parameters calculated by *ACPYPE* are function 1 and only include angle θ^0 (theta) and constant k^θ (cth). Refer to Table 5 for angle function types.
- Beware of the function type when incorporating the data into the force field or put both r^0 and k^{UB} as zero to ignore the harmonic correction term.
- Using different function types causes errors in *GROMACS* preprocessor.
- Manually changing the function type of the corresponding angles in the topology file or using the patch `topolpatch`. and `topolpatch.pl` in Appendix C to solve the problem.

```

$ ./topolpatch.pl

```
- Replace the topology file with the edited topology file output by the patch.

5.6.3 Proper and Improper Dihedral Angles

```
[ dihedraltypes ]
;i j k l func phi0 cp mult
;###MOD
X C CT2 X 9 180.00 0.675 2
[ dihedraltypes ]
;i j k l func q0 cq
;###MOD
CPB CPA NPH CPA 2 0.0000 174.0544
```

1. The proper and improper dihedral angles are both under section [dihedraltypes]. Refer to Figs. 1 and 2 in Appendix A for a more detailed explanation on proper and improper dihedrals.
2. “X” can be used to define the four atoms i , j , k , and l as “any atom type.”
3. The default function for proper dihedrals is function 9: proper dihedral (multiple), while the default function for improper dihedrals is function 2: improper dihedral. The different dihedral function types are listed in Table 6.
4. Proper dihedral (multiple) is defined by ϕ_0 , ϕ_s , the angle between plane ijk and plane jkl , constant cp , k_ϕ , and the multiplicity.
5. Improper dihedrals are defined by angle ξ_0 and constant k_ξ .
6. Note that the angles ξ_0 and ϕ_s are in degree while the constant k_ξ is in kJ/mol/rad^2 .
7. The improper dihedrals calculated using *ACPYPE* are given in function 9 and are treated as proper dihedrals in *GROMACS*.
8. Similar to angle function type, errors arise from the difference in the dihedral function type in topology and `ffbonded.itp`. Changing the dihedral function type in topology or using the patch will resolve the errors.

6 Structure Construction of Protein-Fluorophore Complex

1. The fluorophore is covalently bonded to the protein. To construct the protein-fluorophore complex, open the protein structure, `3CG5.pdb`, and the fluorophore built by *JLigand*, `CMF.pdb`, on *Coot*.

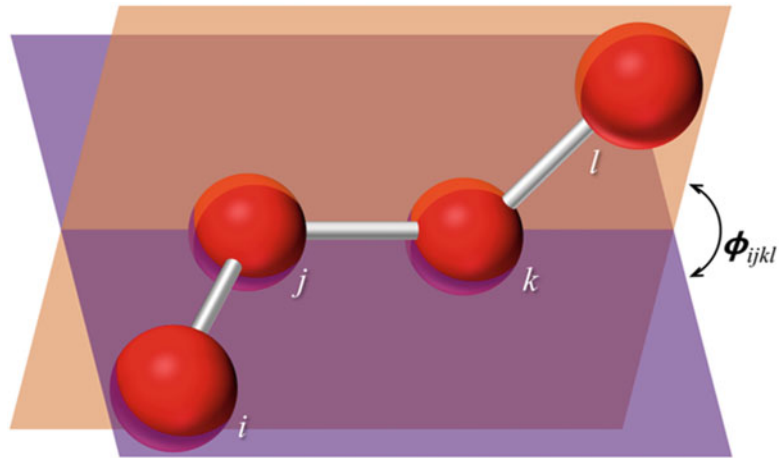


Fig. 1 Proper dihedral angle

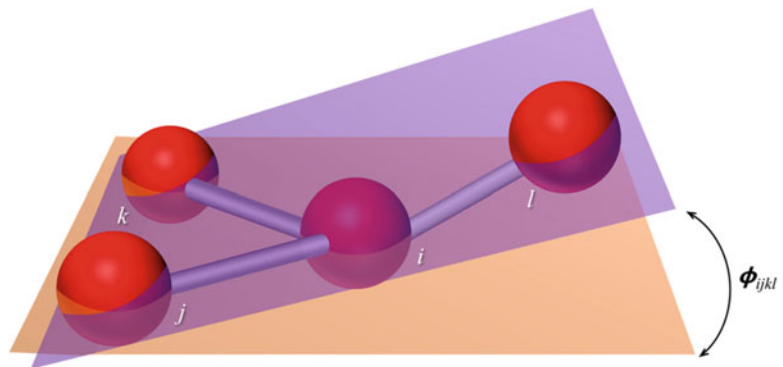


Fig. 2 Improper dihedral angle

2. Overlap the fluorophore-attached amino acid CMF with the target residue spatially and save a copy of the coordinates of the fluorophore-attached amino acid as `CMF_translated.pdb`.
3. Open both the PDB coordinate files of the fluorophore-amino acid complex `CMF_translated.pdb` and the protein `3CG5.pdb` using a plain-text editor.
4. Insert the coordinates of the fluorophore into the protein PDB file manually and save a copy of the new complex coordinates as `3CG5_CMF.pdb` (see **Note 7**).
5. Check the protein-fluorophore complex on *Coot* for any strain or abnormalities.
6. Make sure all the possible chiral structures are built for MD simulation.

Table 6
Dihedral types in `ffbonded.itp` [17]

Name of interaction	Function type	Parameters (units)
Proper dihedral	1	ϕ_s (deg); k_ϕ (kJ/mol); multiplicity
Improper dihedral	2	ξ_0 (deg); k_ξ (kJ/mol/rad ²)
Ryckaert-Belleman dihedral	3	$C_0, C_1, C_2, C_3, C_4, C_5$ (kJ/mol)
Periodic improper dihedral	4	ϕ_s (deg); k_ϕ (kJ/mol); multiplicity
Fourier dihedral	5	C_1, C_2, C_3, C_4 (kJ/mol)
Tabulated dihedral	8	Table number (≥ 0); k (kJ/mol)
Proper dihedral (multiple)	9	ϕ_s (deg); k_ϕ (kJ/mol); multiplicity

7 Molecular Dynamics Simulation of Fluorescent Probe-Labeled Protein

The MD simulation largely follows the *GROMACS* tutorial written by Justin Lemkul (http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/lysozyme_old/index.html). All the commands are written according to *GROMACS* 4.6.7. For later version, please refer to the corresponding *GROMACS* User Manual [17].

7.1 Generate Topology File

1. The MD simulation starts with converting the PDB file to a GRO file and generating a topology file and a positional restraint file.

```
$ pdb2gmx -f 3CG5_CMF.pdb -o 3CG5_CMF.gro -p -water tip3p
```

2. The input PDB file is converted to the output GRO file, and the topology file is generated with the option “-p”. The water model can be specified by “-water” as “none,” “spc,” “spce,” “tip3p,” “tip4p,” or “tip5p”.
3. The force field will then be chosen from `gromacs/top/`. Select the modified force field from the list.

7.2 Topology

1. Topology contains all the information that defines the molecule. After the comment lines, the force field used is specified.

```
; Include forcefield parameters
#include "charmm27mod.ff/forcefield.itp"
```

2. Check the name of the force field to ensure the correct force field is used.
3. Next [moleculetype] indicate that the chain A protein in PDB file was read.

```
[ moleculetype ]
; Name          nrexcl
Protein_chain_A 3
```

4. Each atom is then listed in [atoms]. The missing hydrogen atoms are added according to aminoacids.hdb.

```
[ atoms ]
; nr  type  resnr  residue  atom  cgnr  charge  mass  typeB  charge  massB
; residue 43 ASP  rtp  ASP  q  0.0
  1  NH3   43   ASP    N    1    -0.3   14.007 ; qtot -0.3
  2  HC    43   ASP    H1   2     0.33   1.008 ; qtot 0.03
  3  HC    43   ASP    H2   3     0.33   1.008 ; qtot 0.36
```

nr: atom number

type: atom type

resnr: amino acid residue number

residue: amino acid residue name

atom: atom name, note that as mentioned in Subheading 5.4, some atom names cause clashes in *GROMACS*. Refer to Subheading 8.1 for further information.

cgnr: charge group number

charge: atomic charge, qtot in the comment line keeps a running total charge

mass: atomic mass

5. The subsequent sections [bonds], [pairs], [angles], and [dihedrals] detail the connections and interactions in the protein molecule.

6. Positional restraints are included by the following lines. The restraint file is generated by the `pdb2gmx` command.

```
; Include Position restraint file
#ifdef POSRES
#include "posre.itp"
#endif
```

7. This is the end of the protein definition. The next section defines the solvent. The topology and restraints for the chosen water model are included, followed by the topology of ions.

```
; Include water topology
#include "charmm27mod.ff/tip3p.itp"

#ifdef POSRES_WATER
; Position restraint for each water oxygen
[ position_restraints ]
; i funct      fcx      fcy      fcz
  1  1      1000      1000      1000
#endif

; Include topology for ions
#include "charmm27zc.ff/ions.itp"
```

8. Finally, the name of the system is stated in `[system]` and the molecules, including protein and solvent molecules, are all listed in `[molecules]`.

```
[ system ]
; Name
PROTEIN in water

[ molecules ]
; Compound      #mols
Protein_chain_A  1
```

7.3 Define the Box

1. In MD simulation, the protein is contained in a periodic unit cell. The unit cell is defined by

```
$ editconf -f 3CG5_CMF.gro -o newbox.gro -c -d 1.5 -bt dodecahedron
```

2. The command “-c” center the protein, “-d” specify the size of the box by placing the protein molecule at least 1.5 nm from the box edge.
3. The box type can be defined as “triclinic,” “cubic,” “dodecahedron” (rhombic dodecahedron), or “octahedron” (truncated octahedron) by “-bt”.

7.4 Add Water Molecules

1. Having the box defined, water molecules are added to fill the box in order to mimic the protein molecule in aqueous.

```
$ genbox -cp newbox.gro -cs spc216.gro -o solv.gro --p
```

2. The configuration of the protein is contained in the GRO output file from the previous step and specified by “-cp”.
3. The configuration of water as solvent can be found in standard GRO file in the *GROMACS* program by “-cs”.
4. The topology has to be updated by “-p” to match the number of atoms in GRO file and topology file.
5. The volume of the box, the density, and the number of solvent molecules added to the system will be shown on the CLI. The number of solvent molecules may be checked at the end of the topology under the section [molecules].

```
[ molecules ]
; Compound      #mols
Protein_chain_A    1
SOL                14781
```

7.5 Add Ions

1. Proteins are often stabilized in a salt solution to balance its charge. To add the ions into the solution, “genion” is used to replace water molecules with ions (*see Note 8*).
2. Prior to running “genion,” the coordinate and topology files have to be converted into an atomic-level input TPR file, produced by GROMACS preprocessor, “grompp”.

```
$ grompp -f ions.mdp -c solv.gro -p -o ions.tpr
```

3. The MDP input file (molecular dynamics parameter file) contains parameters for various ions. A sample can be found here

(http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/lysozyme_old/Files/ions.mdp).

4. Errors may occur if there is any inconsistency between the topology file and the force field or between the topology file and the GRO coordinate file. *See* Subheading 8.2 for more details.
5. Once the TPR file is generated, ions can be added using “genion”.

```
$ genion -s ions.tpr -o ions.gro -p -pname K -nname CL -neutral -conc 0.3
```

6. The TPR input file produced by “grompp” entered by “-s”.
7. The positive and negative ions are indicated by “-pname” and “-nname”, respectively.
8. The topology is updated by “-p” to match the number of atoms in the GRO file and the topology file.
9. The ion concentration, in M, is determined by “-conc”. For charged protein, the charge can be neutralized using “-neutral”.
10. A line will be displayed on the prompt asking which group the ions should replace. Select “SOL” to replace the water molecule added in the previous step. Again, the added ions can be found under [molecules] in topology.

```
[ molecules ]
; Compound      #mols
Protein_chain_A    1
SOL                14781
K                  103
CL                 89
```

7.6 Energy Minimization

Prior to MD, to avoid system collapse, steric strain and inappropriate geometry have to be resolved.

1. The energy of the solvated system is minimized using the steepest descent method. A sample set of parameters for the energy minimization, `minim.mdp`, can be found here (http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/lysozyme_old/Files/minim.mdp).
2. *GROMACS* preprocessor, “grompp” is used to process the coordinate file before “mdrun”.

```
$ grompp -f minim.mdp -c ions.gro -p -o em.tpr
```

- The “-v” command makes `mddrun` verbose. The name of the input TPR file and output files are defined by “-deffnm”.

```
$ mddrun -v -deffnm em
```

- Four files will be produced:
 - `em.log`: ASCII-text log file of the EM process
 - `em.edr`: Binary energy file
 - `em.trr`: Binary full-precision trajectory
 - `em.gro`: Energy-minimized structure
- A successful energy minimization should give negative potential energy, E_{pot} , on the order of 10^5 – 10^6 , and the maximum force, F_{max} , should be smaller than 1000 kJ/mol, as set in the MDP file.

```
Steepest Descents converged to Fmax < 1000 in 332 steps
Potential Energy = -7.7104625e+05
Maximum force    = 8.9544427e+02 on atom 3101
Norm of force    = 3.6544601e+01
```

7.7 Canonical Equilibration (NVT)

The system is in a reasonable structure after energy minimization. However, while the solvent and ions have optimized among themselves, they are not necessarily equalized around the protein. To bring the system to the desired temperature, different ensembles are used for equilibration.

- The canonical ensemble is first used on the restrained system. The positional restraint force is applied to the heavy atoms of the protein through the `posre.itp` file (positional restraint file) created by “`pdb2gmx.`” It allows the protein to move in a controlled manner to equilibrium.

```
$ grompp -f nvt.mdp -c em.gro -p -o nvt.tpr
$ mddrun -v -deffnm nvt
```

- As the unit cell is periodic, the protein may partition between unit cells and appears to be cut in the middle. This causes errors in “`grompp.`” See Subheading 8.4 for the solution.

7.8 Gibbs Equilibration (NPT)

With the temperature of the system stabilized by the canonical ensemble, the pressure, hence the density, is then stabilized by the Gibbs ensemble in a similar way.

- A stepwise relaxation in positional restraint allows the protein more freedom to reach the most stable state.

2. A script to relax the restrain from 1000 kJ/nm to 0.1 kJ/nm, “npt.”, can be found in Appendix B.
3. Before performing this stepwise relaxation, a separate positional restrain file has to be created.

```
$ genrestr -f nvt.gro -o posre_mod.itp -fc 1000 1000 1000 -n index.ndx
```

4. Select the protein to be restrained by its number.
5. Check the npt. file to see if the topology file name matches. Manually edit the topology by replacing “POSRES” by “POSRES_MOD” and “posre.itp” by “posre_mod.itp” under

```
; Include Position restraint file
#ifdef POSRES_MOD
#include "posre_mod.itp"
#endif
```

6. Initiate the equilibration

```
$ ./npt. >&npt.log
```

7.9 MD Simulation

Finally, the system has reached a fairly stable state and is ready for MD simulation.

1. Check the md.mdp parameter file carefully before starting the simulation.
2. For protein, the simulation is usually 10 ns, but the timescale can be changed to match the time frame of the predicted molecular movement.

```
$ grompp -f md.mdp -c npt8.gro -t npt8.cpt -p -o md.tpr
```

3. The trajectory from the previous step can be input by “-t”.

```
$ mdrun -v -deffnm md
```

4. The results of the simulation can be visualized by *VMD*.

8 Common Errors

8.1 Atom Name: O1, O2, OC1, OC2

GROMACS considers the following atom as backbone atoms: N, CA, C, O, O1, O2, OC1, OC2, OT, OXT, H1, H2, H3, H, and HN. When naming the modified residue, using the above atom names causes errors. Avoid these names always.

8.2 Misidentification of Atom or Bond Types in ACPYPE

To generate topology files, there can be no missing atoms in the molecule. For example, although zwitterion forms between the backbone nitrogen and carbon, for the program to correctly identify the sp^3 hybridization of nitrogen, both amine and carboxyl

groups have to contain the correct number of hydrogens. If the residue coordinate file has any missing hydrogen atoms, use *Open Babel* to add the missing hydrogen to the structure. Check the generated PDB file on *Coot*.

```
$ babel -h CMF_no_hydrogen.pdb CMF.pdb
```

Atom and bond types can only be specified and modified in the MOL2 file format but not PDB file format. Errors may also arise when the program cannot calculate the charges due to misidentification of atom and bond types. In that case, add all the missing hydrogen atoms and generate a MOL2 file using *Open Babel*. Edit the MOL2 file carefully before using it for parameterization.

```
$ babel -ipdb CMF.pdb -omol2 CMF.mol2
```

8.3 No Default U-B/ Improper Dih. Types

For customized force field using *ACPYPE*, the incorporated angle and dihedrals may be in a different function type than the force field default function type. As the *GROMACS* system assumes all the parameters being in the default function, errors arise from the inconsistency. This is a fatal error and would cause the system to abort.

```
ERROR 1 [file topol.top, line 24020]:
  No default U-B types
Fatal error:
There were 50 errors in input file(s)
```

The function types of the corresponding lines in topology have to be changed according to the `ffbonded.itp`. This can either be done manually or using the patch in Appendix C.

8.4 The Sum of the Two Largest Charge Group Radii Is Larger than *rlist*

The unit cell is treated as periodic in *GROMACS*. During equilibration, the protein may partition between unit cells and appear to be cut in the middle. This causes problems in the *GROMACS* preprocessor, “`grompp`,” especially when the protein is charged, as the charge group are too far apart. To contain the protein in a single unit cell, center the protein using trajectory conversion “`trjconv`”.

```
$ trjconv -s em.tpr -f em.gro -o em_centre.gro -center -pbc mol -n index.ndx
```

The protein is centered by the flag “`-center`” and the type of periodic boundary condition treatment is set by “`-pbc`” as `mol` to put the center of mass of molecules in the box. Other treatments include `res` to put the center of mass of residues in the box and `atom` to put all atoms in the box. The prompt will ask the user to

select the group to be centered (protein) and the group to be output (system). Select by their number.

If ligands or other ions linked to the protein have to be moved together with the protein, create an NDX index file to specify them together.

```
$ make_ndx -f em.gro -o
```

Select the groups by number and combine them using a bar symbol “|” (logical or) between groups. End the section by “q”.

9 Notes

- Conventions of font styles used:
 - Computer software/package names: italicized Roman (e.g., *GROMACS*).
 - Linux Command line input: monospaced Courier (e.g., `$ acpype -i CMF.pdb`).
 - File or directory names mentioned in text: Courier (e.g., `CMF.pdb`, `/usr/bin/`).
 - Contents of plain-text type files: Courier (e.g., `[molecule]`)
- Before using the structure downloaded from the PDB in MD simulations, always check if there are any missing residues as they cause errors and clashes in *GROMACS*. If there are, download homologous protein structures and build the absent residues using *Coot*, referencing the Ramachandran plot to select the most probable torsion angles.
- The absolute path (i.e., location) of the *GROMACS* system topology directory depends on how the system was installed on the local Linux system. From the *GROMACS* main level directory (e.g., `/usr/share/gromacs`), it is one level down, i.e., the `top/` subdirectory.
- If there are problems in copying the subdirectory, check write permission of the files and directories. The user must have write permissions in all *GROMACS* directories to perform all procedures in this chapter.
- Consistency is crucial. Check the residue names, atom names, and atomic charges in all the related files to make sure they are all consistent. A checklist of the files involved may help when changes have to be made.
- In order to trace and reverse the changes made, especially on system files, duplicate the original files before making any amendments. Make a compressed file if needed.
- Be very careful with the invisible tabs and spaces as tabs are not recognized in some files, for example in PDB files. Open the

files with a plain-text editor to search for tabs if unexplainable errors are encountered.

8. Phosphate buffer is a common solvent used in biological experiments. However, it is not parameterized in some force fields. It is recommended to use other anions with the same ionic strength as phosphate ions. The ionic strength equation is

$$I = \frac{1}{2} \sum_{i=1}^n c_i z_i^2$$

where c is the concentration; z is the ionic charge. For example, 300 mM KCl is equivalent to 50 mM K_3PO_4 in ionic strength.

Appendix A: Proper and Improper Angle (*See Figs. 1 and 2*)

Appendix B: npt

```
#!/bin/csh -f
#
# Script to run NPT with reducing restraints from 1000 to 0.1
kJ/mol/nm
#
# First restraint files (1000 kJ/mol/nm) made by pdb2gmx
# edit npt1.mdp file next

# restraint file: posre_Protein_chain_B.itp
# restraint file: posre_Ion_chain_Z.itp
# restraint file: posre_OH-.itp

grompp -f npt1.mdp -c nvt.gro -t nvt.cpt -p -o npt1.tpr
mdrun -v -deffnm npt1

sed -i 's/1000 1000 1000/500.0 500.0 500.0/g' posre_mod.itp
grompp -f npt1.mdp -c npt1.gro -t npt1.cpt -p -o npt2.tpr
mdrun -v -deffnm npt2

sed -i 's/500.0 500.0 500.0/200.0 200.0 200.0/g' posre_mod.itp
grompp -f npt1.mdp -c npt2.gro -t npt2.cpt -p -o npt3.tpr
mdrun -v -deffnm npt3

sed -i 's/200.0 200.0 200.0/100.0 100.0 100.0/g' posre_mod.itp
grompp -f npt1.mdp -c npt3.gro -t npt3.cpt -p -o npt4.tpr
mdrun -v -deffnm npt4

sed -i 's/100.0 100.0 100.0/ 50.0 50.0 50.0/g' posre_mod.itp
grompp -f npt1.mdp -c npt4.gro -t npt4.cpt -p -o npt5.tpr
mdrun -v -deffnm npt5

sed -i 's/50.0 50.0 50.0/10.0 10.0 10.0/g' posre_mod.itp
grompp -f npt1.mdp -c npt5.gro -t npt5.cpt -p -o npt6.tpr
mdrun -v -deffnm npt6

sed -i 's/10.0 10.0 10.0/1.0 1.0 1.0/g' posre_mod.itp
grompp -f npt1.mdp -c npt6.gro -t npt6.cpt -p -o npt7.tpr
mdrun -v -deffnm npt7

sed -i 's/1.0 1.0 1.0/0.1 0.1 0.1/g' posre_mod.itp
grompp -f npt1.mdp -c npt7.gro -t npt7.cpt -p -o npt8.tpr
mdrun -v -deffnm npt8

#!/bin/rm -f posre_Protein_chain_B.itp posre_Ion_chain_Z.itp
posre_OH-.itp
exit
```

Appendix C: topolpatch

```
#!/bin/csh -f
#
# direct grompp output into a file:
# grompp -f ions.mdp -c solv.gro -p -o ions.tpr > & grompp.log
#

rm -f *.err

grep "ERROR"      --after-context=1 grompp.log > grompp.err
grep "U-B"        --before-context=1 grompp.err > grompp-ub.err
grep "Improper"  --before-context=1 grompp.err > grompp-im.err

exit

sed -i s// topol.top
```


topolpatch.pl

```

#!/usr/bin/perl

# perl script to correct for no default U-B and improper dihedral types
# in ZC's customerised Charm27 forcefield of fluorescein
# by Y.W. Chen, Jan 2018

# Required: grompp.log file containing error messages & topol.top file

open(GROMPPLOG, "<grompp.log") || die "Can't open grompp.log\n";
open(GMXTOPIN, "<topol.top") || die "Can't open topol.top";
open(GMXTOPOUT, ">topol.top.new");

while(<GROMPPLOG>) {
  if (/ERROR/) {

    my $nextline = <GROMPPLOG>;

    if ($nextline =~ /U\ -B/) {      # Fill U-B error array
#     print $nextline; # for debug
      ($junk1, $serrNo, $junk2, $junk3, $junk4, $serrLine) = split(/\ /,
$_) ;
      $serrLine =~ s/]:>//s;      # remove trailing "]:" characters
#     print $serrNo, " ", $serrLine, "\n"; # for debug
      $subError[$serrNo] = $serrLine;
    }

    elsif ($nextline =~ /Improper/) { # Fill improper dihedrals error
array
#     print $nextline; # for debug
      ($junk1, $serrNo, $junk2, $junk3, $junk4, $serrLine) = split(/\ /,
$_) ;
      $serrLine =~ s/]:>//s;      # remove trailing "]:" characters
#     print $serrNo, " ", $serrLine, "\n"; # for debug
      $idError[$serrNo] = $serrLine;
    }
  }
}

close (GROMPPLOG);

#print "\n", @ubError, "\n"; # for debugging
#print "\n", @idError, "\n"; # for debugging

@topolLine = <GMXTOPIN>;
close (GMXTOPIN);

foreach $subError (@ubError) {
  if ($subError > 0) { # skip handling $array[0]
    $stopolLine[$subError-1] =~ s/ 5/ 1/;
#     print $stopolLine[$subError-1]; # for debugging
  }
}

foreach $idError (@idError) {
  if ($idError > 0) { # skip handling $array[0]
    $stopolLine[$idError-1] =~ s/ 2/ 9/;
#     print $stopolLine[$idError-1]; # for debugging
  }
}

print GMXTOPOUT @topolLine;

close (GMXTOPOUT);

```

References

1. Toseland CP (2013) Fluorescent labeling and modification of proteins. *J Chem Biol* 6 (3):85–95
2. Ploetz E, Lerner E, Husada F, Roelfs M, Chung S, Hohlbein J, Weiss S, Cordes T (2016) Förster resonance energy transfer and protein-induced fluorescence enhancement as synergetic multi-scale molecular rulers. *Sci Rep* 6:33257
3. Rainey KH, Patterson GH (2019) Photo-switching FRET to monitor protein–protein interactions. *Proc Natl Acad Sci U S A* 116 (3):864–873
4. Söveges B, Imre T, Póti ÁL, Sok P, Kele Z, Alexa A, Kele P, Németh K (2018) Tracking down protein–protein interactions via a FRET-system using site-specific thiol-labeling. *Org Biomol Chem* 16(32):5756–5763
5. Margineanu A, Chan JJ, Kelly DJ, Warren SC, Flatters D, Kumar S, Katan M, Dunsby CW, French PMW (2016) Screening for protein–protein interactions using Förster resonance energy transfer (FRET) and fluorescence lifetime imaging microscopy (FLIM). *Sci Rep* 6:28186
6. Sousa da Silva AW, Vranken WF (2012) ACPYPE - AnteChamber PYthon Parser interface. *BMC Res Notes* 5(367):1–8
7. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 25:247–260
8. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66:486–501
9. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun* 91:43–56
10. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GRGMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4:435–447
11. Páll S, Abraham MJ, Kutzner C, Hess B, Lindahl E (2015) Tackling exascale software challenges in molecular dynamics simulations with GROMACS. In: *Solving software challenges for exascale*. Springer, Stockholm, pp 3–27
12. Lebedev AA, Young P, Isupov MN, Moroz OV, Vagin AA, Murshudov GN (2012) JLigand: a graphical tool for the CCP4 template-restraint library. *Acta Crystallogr D Struct Biol* 68 (4):431–440
13. O’Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. *Journal of Cheminformatics* 3(33):1–14
14. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14(1):33–38
15. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–242
16. Tremblay LW, Hugonnet JE, Blanchard JS (2008) Structure of the covalent adduct formed between Mycobacterium tuberculosis β -lactamase and clavulanate. *Biochemistry* 47 (19):5312–5316
17. van der Spoel D, Lindahl E, Hess B, GROMACS Development Team (2014) GROMACS User Manual version 4.6.7, Sweden



Protocol for Simulations of PEGylated Proteins with Martini 3

Fabian Grünewald, Peter C. Kroon, Paulo C. T. Souza, and Siewert J. Marrink

Abstract

Enhancement of proteins by PEGylation is an active area of research. However, the interactions between polymer and protein are far from fully understood. To gain a better insight into these interactions or even make predictions, molecular dynamics (MD) simulations can be applied to study specific protein-polymer systems at molecular level detail. Here we present instructions on how to simulate PEGylated proteins using the latest iteration of the Martini coarse-grained (CG) force-field. CG MD simulations offer near atomistic information and at the same time allow to study complex biological systems over longer time and length scales than fully atomistic-level simulations.

Key words Martini, Proteins, Polymers, PEGylation, Coarse-grained simulation, Modified proteins

1 Introduction

Since polyethylene glycol (PEG) was for the first time covalently attached to proteins in the late 1970s, this strategy, known as PEGylation, has become a viable tool for enhancing proteins [1]. For example, PEGylation is known to increase the stability of the secondary structure against temperature [2], increase circulation time of protein therapeutics [3], and decrease immune responses [3]. At the same time, it usually does not interfere with protein activity or secondary structure [1–6]. However, it has also been found that the PEG conformation is intimately connected to the efficiency of mentioned enhancements [7, 8]. Generally, it is assumed that the PEG chain adopts one of two possible conformational motifs: the so-called “shroud” conformation or the “dumbbell” conformation. In the shroud conformation, PEG wraps around the protein. In the dumbbell conformation, however, PEG exists as a coil next to the protein resembling one end of the dumbbell with the other end being the protein [4]. Overall, there

appears to be no consensus as to whether one conformation is favored over the other. Both direct and indirect evidence exists for either hypothesis and it seems to depend on molecular weight, the specific protein, as well as how many PEG chains are attached [4]. For example, in their hallmark study, Pai and coworkers used small-angle neutron scattering (SANS) to directly measure the polymer conformation of PEGylated lysozyme and human growth hormone. Their study shows that PEG assumes the dumbbell type conformation, where it exists as a coil next to the protein [4]. In a more recent study, LaCouer and coworkers showed that the activity of PEGylated hemoglobin peaks with a PEG chain length of 10 kg/mol. They hypothesized, also based on SANS experiments, that this change in activity is caused by PEG transitioning from the dumbbell to the shroud conformation, where the polymer wraps around the protein. However, it was not possible to find a sharp crossover point as function of increasing chain length where this conformational change would occur [8].

Molecular dynamics (MD) simulations have also been used to investigate the conformations of PEG in PEGylated proteins [9]. MD simulations are a powerful tool because they offer insight into biomolecular assemblies at the molecular level. Recently, Munasinghe and coworkers studied PEGylated Albumin Bovin Serum (BSA) using atomistic MD simulations and observed a conformational change with increasing chain length. They state that this is likely to be driven by strong interactions with specific amino acids. These interactions only become relevant as the PEG length increases, thus driving the conformational change [7]. If this proves true, it would mean that the conformations of PEGs and thus their enhancement capabilities are protein specific. However, as pointed out by Lin, Ramezanghorbani, Colina and coworkers, atomic level detailed simulations are limited in timescale, length, and complexity [7, 9]. To overcome these limitations and potentially have high throughput screenings, coarse-grained (CG) molecular dynamics simulations can be used.

CG simulations with the Martini force-field [10] are among the most popular for biomolecular applications. They have been widely applied to study complex biological systems such as the plasma membrane [11] and the thylakoid membrane including the light-harvesting complex II [12]. For a practical view on the Martini force-field, see the relevant chapter in the same book series [13]. Here we only recount the basic details of the model and its latest iteration (i.e., Martini 3).

The Martini force-field utilizes a building-block approach. Chemical moieties or small molecules of 2–5 non-hydrogen atoms are represented as one particle, called bead. Beads are the minimal building blocks and can be combined together to represent larger molecules. Each bead has a type, which defines how it interacts with the other beads in the force-field and its size. The

type is chosen from a predefined set of types. The best type is selected by closely matching experimental free energies of transfer from water to organic solvents of the underlying chemical fragment [10]. However, with Martini 3 other experimental properties such as miscibility are increasingly used for the selection and validation as well [14]. Bonded interactions between the beads, such as bond distances, angles and dihedral angles, are optimized to best represent the underlying molecular geometry, volume, and flexibility. They are derived by reproducing reference probability distributions obtained from atomistic simulations [15].

Using this approach, CG models for many biological [16–18] and synthetic molecules [19–22] have been created. One of the strengths of the Martini force-field is the compatibility between these different models. For example, it is no problem to combine PEG with proteins to represent PEGylated proteins as has been done before with Martini [23–25]. However, as detailed recently, Martini 2 has some pitfalls and drawbacks [26]. For instance, no standard bead was able to represent PEG with sufficient accuracy. This has led various authors to create special beads for PEG, which in turn limited their compatibility when used under different circumstances than the authors had designed it for [21]. A different drawback of Martini 2 are the overestimated interactions of proteins with each other [26] and potentially incorrect PEG protein interactions [23]. To overcome these limitations, we recommend the new Martini 3 force-field, which was specifically designed to increase the compatibility and to represent a wider variety of chemical fragments accurately. It has already been shown to overcome some of the drawbacks of the original Martini model [14, 15, 21, 27, 28].

In the next section, we will work through and explain how to generate parameters and input structures for PEGylated proteins using Martini 3. However, this guide can also be taken as an example for generating parameters and structures for simple proteins (Subheading 2.1) and polymers (Subheading 2.2). In Subheading 3, a detailed protocol for setting up a simulation and equilibrating it will be presented. Finally, Subheading 4 comprises useful practical tips and information.

2 Martini Parameters for PEGylated Proteins

Usually PEGylated proteins consist of (1) a protein, which is unmodified with respect to its native state [1] (i.e., preserved secondary structure and amino acid sequence), (2) a linker fragment attached to an amino acid, and (3) the PEG polymer chain. Following the Martini building-block approach, parameters for each of the parts are generated separately and then combined. In the last step an input structure for the PEGylated protein is

generated. Subheading 3 subsequently shows how to setup a simulation and equilibrate it. The flow chart in Fig. 1 shows this process in more detail. Whenever appropriate, we will discuss choices of parameters, input file formats and program options in a more general manner. In this way, the reader can adopt the options to their target problem.

2.1 Software Requirements

All the necessary input files to follow this tutorial can be downloaded from our website (<http://www.cgmartini.nl/>) or our Git-Hub page (<https://github.com/marrink-lab>). The tutorial requires a working installation of GROMACS (version 2016 or higher) [29], python 3, martinize 2 (<https://github.com/marrink-lab/vermouth-martinize>), and polyply (https://github.com/marrink-lab/polyply_1.0). For visualization, any program which can visualize gro files and PDB files, such as VMD or pymol, can be used. All programs required are open source. As detailed in **Note 1**, the tutorial—with minimal modification—can also be run on Windows or Mac OS. Note that commands, which need to be executed in the terminal are preceded by “\$.” All basic commands can also be found next to the flow chart (Fig. 1) as quick reference.

2.2 Martini Parameters for Simple Proteins

As example protein, we have chosen a mono-PEGylated lysozyme as used in the study of Pai and coworkers [4]. To begin, an appropriate structure (i.e., PDB code LZ3T) of lysozyme needs to be downloaded from the protein data bank. As discussed in **Note 2**, it is important to make sure that it is complete and contains all non-hydrogen atoms. Once the structure is obtained, the program martinize 2 [30] will be used to generate both CG itp files and starting structures.

2.2.1 Martinizing Lysozyme

Download the PDB file using the following command:

```
$ wget http://www.rcsb.org/pdb/files/3lzt.pdb
```

Martinize 2 requires definitions of the force-field and mappings of the amino acids. Files with these definitions are shipped with martinize 2, so only the name of the force-field (i.e., “-ff martini30b32”) needs to be provided. Furthermore, the atomistic PDB file of the protein, downloaded in the previous step, is required. The following command generates the basic parameters and a coarse-grained structure file.

```
$ martinize2 -f 3lzt.pdb -ff martini30b32 -x lysozyme_cg.pdb -o topol.top
```

In general, the basic options above should be supplemented by few more to generate appropriate parameters for Martini 3 proteins:

Example Commands

```

$ wget www.rcsb.org/pdb/files/3lzt.pdb

$ martinize2
-f 3lzt.pdb
-ff martini30b32
-x lysozyme_cg.pdb
-o topol.top
-dssp -scfix -cys auto
-elastic -p backbone

$ gmx editconf
-f lysozyme_cg.pdb
-o lysozyme_cg.gro
-box 8.5 8.5 8.5

$ polyply gen_itp
-f PEO.martini.3b.itp
  OH_end.itp OH_link.ff
-seq PEO:50 OHend:1
-o PEG_50_OH.itp
-name PEGOH

$ polyply gen_itp
-f molecule_0.itp MEE.itp
  PEG_50_OH.itp methoxy_link.ff
-seq molecule_0:1 MEE:1 PEGOH:1
-o lysoPEG.itp
-name lysoPEG

$ polyply gen_coords
-p system.top
-o lysoPEG.gro
-name lysoPEG
-c lysozyme_cg.gro

$ gmx solvate
-cp lysoPEG.gro
-o solvated.gro
-cs water.gro
-radius 0.21

$ gmx grompp
-f min.mdp
-c solvated.gro
-p system.top
-o dummy.tpr

$ gmx genion
-f solvated.gro
-o start.gro
-s dummy.tpr
-conc 0.15
-neutral

$ EM_EQ_run.sh
    
```

Flowchart

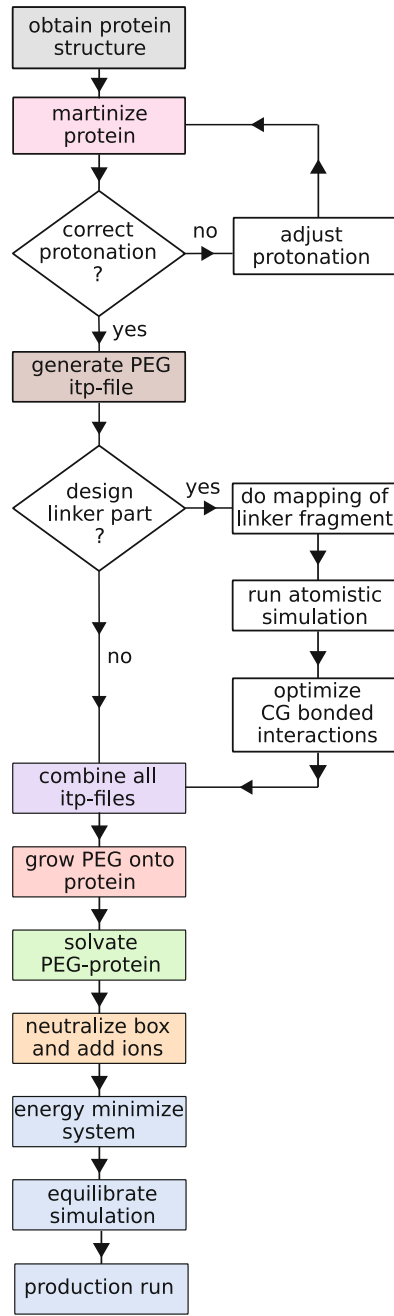


Fig. 1 Process of generating parameters and input structures for PEGylated proteins with Martini. The left column shows all commands required for the example

“*-dssp*” invokes the program DSSP [31, 32] to analyze the secondary structure of the reference structure in PDB file. This gives the necessary input to generate the bonded parameters of the protein model, which are secondary structure dependent [16]. It is not always appropriate to define the secondary structure based on a reference structure. In this case, you can define it manually as explained in **Note 3**.

“*-elastic*” is the option used to generate an elastic network to maintain the tertiary structure. In general, there are two different options for maintaining the tertiary structure: the elastic network approach [33], and the Go approach [34]. For this tutorial, the elastic network approach will be used. As outlined in **Note 4**, the Go approach can be used for more accurate protein dynamics.

“*-scfix*” tells martinize to apply the side chain fix. Herzog and coworkers have shown that including additional dihedral angle potentials for the amino acid side chains improves the protein dynamics [35]. The dihedrals are defined based on the reference atomistic structure and applied to the whole protein. So, in case you are simulating an unfolded peptide, you should not use the *-scfix* option. This so-called side chain fix (ScFix) is used by default for Martini3 folded proteins and applied for this tutorial.

“*-cys*” is used to let martinize 2 determine whether cysteine bridges exist. If the program finds a cysteine bridge, it will include the interactions that link the two beads participating in the cysteine bridge. Whereas linking PEG via the thiol group of cysteine is a very popular method for PEGylating proteins, the ligation usually targets cysteines that are not involved in cysteine bridges [1]. Therefore, this option can safely be used for all PEGylated proteins unless PEGylation is specifically known to disrupt a cysteine bridge. In that case, the relevant interactions need to be removed from the itp file manually after running martinize 2.

“*-p backbone*” can be used to generate position restraints for the protein backbone beads. These restraints are useful for the initial equilibration step. For example, it allows the water to penetrate cavities, which otherwise would collapse quickly. Because later, when this equilibration procedure is applied, these initial parameters are required.

Combining all the options above will generate a standard Martini 3 lysozyme protein with elastic network as well as a coarse-grained structure file. The final command is the following:

```
$ martinize2 -f 3lzt.pdb -ff martini30b32 -x lysozyme_cg.pdb
  -dssp -elastic -scfix -cys auto -p backbone -o topol.top
```


2.2.2 Checking Protonation States

Regular Martini 3 uses fixed protonation states, which means titratable amino acids are either neutral or charged. The pH of the simulation is usually assumed to be at physiological pH (i.e., pH 7.4). However, it should always be verified that the protonation states of the titratable amino acids are correct. The easiest way is to inspect the CG itp file. Aspartic acid, Glutamic acid, Lysine, Histidine, Tyrosine, Cysteine, the C-terminus, and the N-terminus can in principle change their protonation state. Their protonation state together with the number of Arginines (which are always charged at pH 7.4) determines the total charge of the protein, which can also serve as an indication for the protonation states. The following two commands will print all titratable amino acids and compute the total charge.

```
$ egrep 'ASP | LYS | GLU | HIS | TYR | CYS' molecule_0.itp
$ grep Q molecule_0.itp | awk '{sum += $7} END {print sum}'
```

Inspecting the output will show that lysozyme has a total charge of +8 and that all titratable amino acids except Tyrosine are charged. This is consistent with titration experiments of lysozyme [36, 37]. It is well known that pK_a values of amino acids can change as a result of their local environment. Therefore, in the absence of experimental data, the pK_a values of the amino acids in the protein should at least be estimated (e.g., using the H++ server <http://biophysics.cs.vt.edu/>) or more advanced tools and models should be used (*see Note 5*). To modify the protonation state with martinize 2, first a PDB file with the accurate protonation states needs to be obtained. For example, such a file can be downloaded from one of the servers performing the estimates. Make sure the format adheres to the specifications outlined in **Note 2**. Subsequently, the residue names of the amino acids, with changed protonation state, have to be changed according to the following scheme: Add a zero in front of the name and delete the last letter. For example, the residue name of aspartic acid changes from “ASP” to “0AS.” Using the modified PDB file, martinize can be run again to get the parameters for the protein with accurate protonation states.

2.3 Martini Parameters for PEG

In this section, we will show how to generate input parameters for Martini PEG using the program polyply. Polyply can be used to generate structures and input parameters for linear polymers or stitch together any existing itp files.

2.3.1 Homopolymer

To generate itp files for any polymer, polyply needs a monomer itp file (GROMACS format). This monomer itp file needs to contain all atoms part of the monomer repeat unit within the “[atomtypes]” directive. Furthermore, the file needs to include all the

bonded interactions that all the atoms listed before have with all following atoms. The idea is outlined in Fig. 2. Consider situation A, where we want a polymer of four monomers. Our input file should contain atom 1 in the “[atomtypes]” directive and all the interactions of 1 with the next monomers (indicated by arrows). The idea of this format is that the parameters for the rest of the monomers can be obtained by shifting the initial monomer (blue) one to the right and generating all bonded interactions accordingly. So the interaction between 1 and 2 becomes an interaction between 2 and 3 (Fig. 2b). This would generate an interaction between 2 and 5. However, since our chain only has four monomers this interaction would be dropped. As an example, the monomer input file for the Martini3 beta-version of PEG [21] is shown below. The PEG repeat unit is $-\text{[CH}_2\text{-O-CH}_2\text{]}-$ and modeled as one bead.

```
[ moleculetype ]
; name nexcl.
PEO 1
[ atoms ]
1 SN1a 1 PEO EO 1 0.000 45
[ bonds ]
1 2 1 0.37 7000
[ angles ]
1 2 3 2 135.00 50
1 2 3 10 135.00 75
[ dihedrals ]
1 2 3 4 1 180.00 1.96 1
1 2 3 4 1 0 0.18 2
1 2 3 4 1 0 0.33 3
1 2 3 4 1 0 0.12 4
```

Because Martini PEG is one bead per repeat unit, the “[atomtypes]” directive only has one atom. Furthermore, it has one bond with the next monomer, two angles, and four dihedral angle terms involving the next three monomers. Hence the highest atom index, which needs to be included in this itp file, is 4. Note that all monomer itp files always need to start with atom index 1. Polyly is also provided with a number of default monomer itp files (*see Note 6*). To generate a PEG of 3 kDa length (e.g., ~50 repeat units) only the monomer itp file, the sequence of monomers and the name of the molecule have to be provided as well as a name for the newly generated itp file. The sequence is provided using the “-seq” flag and consists of one or more blocks of the format “residue name: number of monomers.” The residue name must match at most one molecule name in the itp files provided with the “-f” flag.

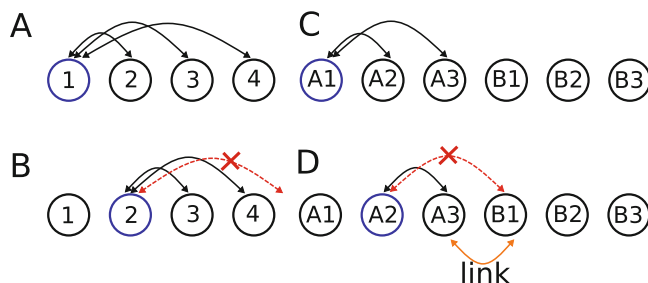


Fig. 2 Schematic of itp file generation using polyply; arrows indicate interactions (e.g., bonds, angles, etc.) between monomer fragments indicated by circles. Itp files of the polymer are generated from a single monomer itp file. This itp file (blue circle panel **a**) needs to define all interactions of that fragment with the next monomers (1–4 panel **a**). Then these interactions are used to generate the new interactions (panel **b**), but all interactions exceeding the maximum number of atoms (red arrow) are not taken into account. Similarly, in the case of block-copolymers the A block is built from a single monomer of each block (A1 and B1 panel **c**). When the interactions are generated, all interactions exceeding the single block (red arrow panel **d**) are not taken into account. This makes it necessary to supply an extra file defining all links between blocks (orange arrow, panel **d**)

```
$ polyply gen_itp -f PEO.martini.3b.itp -seq PEO:50 -o PEG_50.itp -name PEG
```

2.3.2 End Group and Block-Copolymers

Often, the end of a homopolymer chain is different from the monomeric repeat unit. For PEG, for example, one can either have a $\text{CH}_2\text{-CH}_3$ or $\text{CH}_2\text{-OH}$ group at the end. Especially for small PEG molecules the choice is important because the influence of the end group is proportionally larger than for longer chains. To illustrate the itp file generation for the case when an end group needs to be attached, the PEG chain will be terminated with a $\text{CH}_2\text{-OH}$ end group.

Attaching an end group follows the same procedure as attaching another block to form a block copolymer. All that is needed is another monomer itp file for the block and a link file, which uses the vermouth force-field format (i.e., “.ff”) [30]. This concept is illustrated in Fig. 2c, d. Polyply will generate the bonded interaction from the monomer itp files for each block separately, removing any overlapping interactions between the blocks. For example, the interaction between A1 and A3 would generate a new interaction between A2 and B1. However, since B1 belongs to a different block, the interaction is removed. Therefore, a second itp file specifying all the interactions linking the two blocks has to be supplied. This link file adheres to the vermouth force-field format. For use with polyply, each link file needs to contain the same first three lines. Those lines are shown in the example file below. These lines tell the program that it is a link and that the interactions listed

directly apply to the final polymer. After these three lines are defined all interactions specifying the link with the correct atom numbers as in the final polymer should follow. The link file cannot contain the “[moleculetypes]” and “[atomtypes]” directive. Note that the indices in the bond and angle directive do not start at 1 but are exactly the same as in the final polymer. For instance, for the end group, the link file should look as follows:

```
[ link ]
[ molmeta ]
by_atom_id true
[ bonds ]
50 51 1 0.280 7000.0
[ angles ]
49 50 51 2 140.00 25.0
```

And the CH₂-OH capping group monomer itp file only defines one bead:

```
[ moleculetype ]
; name nexcl.
OHend 1
[ atoms ]
1 TP1 1 OHend EO 1 0.000 36
```

This procedure of stitching together two itp files will also be used to combine the protein itp file and the PEG itp file together in the end. For now, the following command can be used to combine the PEG_50.itp with the end group. Note that the file extension of the link file is “.ff” and not “.itp.”

```
$ polyply gen_itp -f PEG_50.itp OH_end.itp OH_link.ff -seq
PEG:1 OHend:1 -o PEG_50_OH.itp -name PEGOH
```

The same result could also be obtained using one command:

```
$ polyply gen_itp -f PEO.martini.3b.itp OH_end.itp OH_link.ff -seq
PEO:50 OHend:1 -o PEO_50_OH.itp -name PEGOH
```

2.4 Linking PEG to the Protein

So far we have shown how to generate parameter files for lysozyme and PEG. However, before these are combined, it has to be defined how the PEG chain is attached to the protein. Attachment of PEGs to proteins is often done via the amine group of a Lysine or the N-terminus, or the thiol group of a Cystein [1].

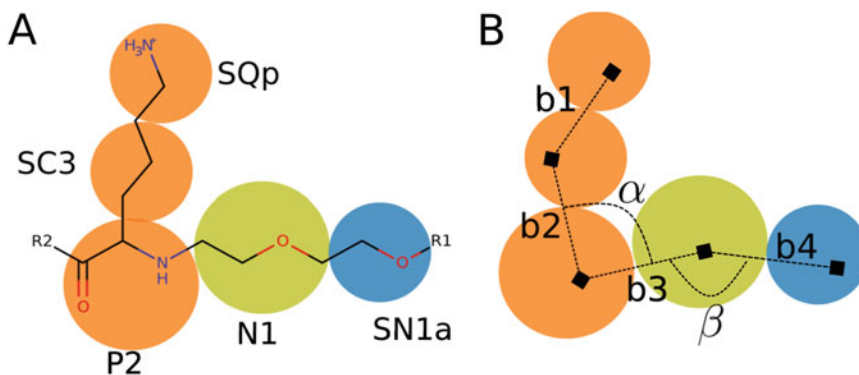


Fig. 3 Methoxyethane (yellow) linking PEG (R1, one repeat unit shown in blue) to the N-terminus of the protein (R2, terminal residue shown in orange). Panel **a** shows the mapping (i.e., each circle is one bead) and bead types. Panel **b** shows the definition of the bonded interactions b1–b4, α , and β

2.4.1 Designing Martini Parameters for the Linker

The linker motif used in this example is shown in Fig. 3a. It is the result of reductive alkylation, which is used to selectively attach PEG to the N-terminus. To design parameters for this linker, a mapping has to be designed which splits the linker into beads. The mapping is shown as circles in Fig. 3a. There is only one bead between the N-terminus and the PEG chain. Note that the mapping of the N-terminal amino acid (i.e., lysine) becomes the same as for the backbone in the rest of the protein. Thus the default backbone bead type (i.e., P2) should be used for this bead. The bead linking PEG and protein corresponds to methoxyethane. Therefore, the bead type that best resembles methoxyethane has to be found. The free energy of transfer from octanol to water of methoxyethane was estimated to be around 3.4 kJ/mol [10]. An N1 bead has a free energy of transfer of 4.1 kJ/mol and is thus the best match. The remaining beads are the regular PEG beads. Bonded interactions (i.e., bonds and angles), were obtained by simulating a fragment (R1=CH₃, R2=CH₂-O-CH₃) in water at the atomistic level using GROMOS [38] parameters obtained from the ATB [39] and reproducing the probability distributions at the CG level. We will not go into more detail for designing the linker, as it follows the normal design rules for Martini (*see Note 7*).

2.4.2 Combining Protein, Linker and PEG Parameters

Next, all itp files are combined together: first, the bead type of the N-terminal Lysine has to be changed to a neutral P2 bead. Open the lysozyme.itp file and change the type of the first bead to P2 and the charge to zero. The beginning of your edited file should look as follows:

```
[ moleculetype ]
molecule_0 1
```

```
[ atoms ]
1 P2 1 LYS BB 1 0.0
2 SC3 1 LYS SC1 2 0.0
3 SQp 1 LYS SC2 3 1.0
```

Next the itp file for the methoxyethane linker is defined. As it is only one bead, it looks the same as for the CH₂-OH end group.

```
[ moleculetype ]
; name nexcl.
MEE 1 ; methoxyethane-link
[ atoms ]
1 N1 1 MEE MEE 1 0.000 72
```

As done for the CH₂-OH end group before, a link file has to be created. The link file has to contain all bonded interactions that span the N-terminus, methoxyethane, and the first PEG bead. Figure 3b shows these interactions. Note that b1 and b2 are already defined in the protein itp file. In contrast to the file above defining the monomeric repeat unit, the link file uses the indices of the final itp file. In this example, the itp files are combined in the order: protein-linker-PEG. Thus the N-terminus will be the first bead with index 1. Because lysozyme has 292 beads, the methoxyethane linker bead will have the atom index 293 and the first PEG bead will have the index 294. Therefore, the link itp file needs to look as follows:

```
[ link ]
[ molmeta ]
by_atom_id true
[ bonds ]
1 293 1 0.41 2000 ; b3
293 294 1 0.39 5000 ; b4
[ angles ]
293 1 2 2 150 15 ; alpha
294 293 1 2 170 50 ; beta
```

Finally, all the files can be combined to obtain an itp file for the PEGylated lysozyme:

```
$ polyply gen_itp -f molecule_0.itp MEE.itp PEG_50_OH.itp methoxy_link.ff -seq
molecule_0:1 MEE:1 PEGOH:1 -o lysoPEG.itp -name lysoPEG
```

If all the files had already been generated, the links for the OH end group and methoxyethane could have been combined into one

file. Then it would have been possible to generate the itp file using a single command:

```
$ polyply gen_itp -f molecule_0.itp MEE.itp PEO.martini.3b.itp OH_end.itp
  combined_links.ff -seq molecule_0:1 MEE:1 PEG:50 OHend:1 -o lysoPEG.itp
  -name lysoPEG
```

However, it is common to use multiple invocations rather than doing everything at once.

2.5 Generation of Input Structures

Having obtained an itp file for PEGylated lysozyme, a starting structure can be generated. The protein structure generated by martinize 2 is supplied to polyply, which will add a PEG chain to it. Polyply only reads gro files. Thus the martinize 2 PDB file needs to be converted to gro format. In addition, the box size for the system can already be specified. As further detailed in **Note 8**, a sufficiently large box size should be used in order to stay below the overlap concentration. A cube with sides of 8.5 nm is sufficient for this purpose.

```
$ gmx editconf -f lysozyme_cg.pdb -o lysozyme_cg.gro -box 8.5 8.5 8.5
```

Polyply also requires an accurate topology file including the same information as used to run the simulation in vacuum. The topology file for PEGylated lysozyme will look as follows:

```
#include martini3/martini_v3.0.4.itp
#include lysoPEG.itp
[ system ]
lysoPEG in water
[ molecules ]
lysoPEG 1
```

This topology file has to be provided to polyply together with the name of the polymer and some other options as shown below:

```
$ polyply gen_coords -p system.top -o lysoPEG.gro -name lysoPEG
  -c lysozyme_cg.gro
```

Using the approach shown above polyply can in principle generate a starting conformation for any CG polymer with a few limitations as outlined in **Note 9**. Because polyply cannot generate the protein structure, the one obtained from martinize 2 is reused. Using this command an input structure as shown in Fig. 4 is generated. To generate just a PEG chain in vacuum, it would have been possible to omit the “-c” option and not define any initial structure to reuse.

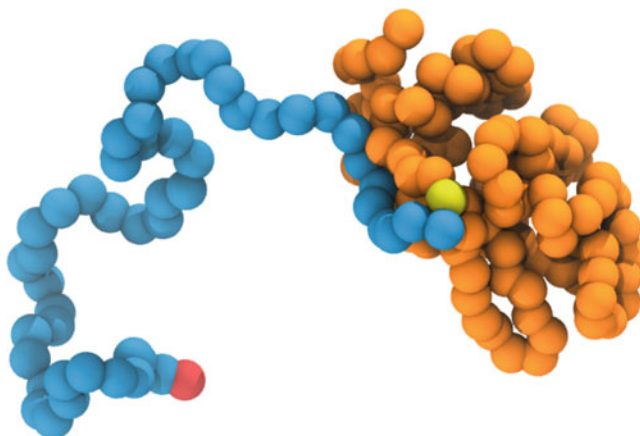


Fig. 4 PEGylated lysozyme initial structure; PEG is shown in blue, the linker bead in yellow and the protein backbone in orange

3 Setting Up and Running a Simulation

Once the starting structure for the PEGylated protein is generated, the system can be set up. Before solvating, the system should be energy minimized applying both positional restraints to the protein backbone and the polymer. This allows the initial structure to relax, but prevents it from coiling up. The positional restraints and the flexible options can be switched on by including the following line in the mdp file. All mdp files are provided with the tutorial files. General mdp files appropriate for Martini can be downloaded from our website (<http://www.cgmartini.nl/>).

```
define = -DFLEXIBLE -DPOSRES
```

Now run:

```
$ gmx grompp -f min.mdp -p system.top -c lysoPEG.gro -o min.tpr -r lysoPEG.gro
$ gmx mdrun -v -deffnm min
```

Next, the energy-minimized structure is solvated with water using the GROMACS tool `gmx solvate`:

```
$ gmx solvate -cs water.gro -cp min.gro -o protein_water.gro
  -radius 0.21 2>&1 | tee solv.out
```

This will add water into the simulation box. It is important to set the “-radius 0.21” option to account for the fact that Martini water is representing four water molecules at a time. If it is not set, GROMACS will pack too many water beads, which might lead to

instabilities during the equilibration. Because solvate adds water molecules to our system, the topology file needs to be revised. The following two commands will do it automatically.

```
$ water=$(grep "W (" solv.out | awk '{print $5}')
$ echo "WN ${water}" >> system.top
```

Besides water, salt needs to be added to the simulation box for two reasons: (1) As the protonation states are fixed, the solution needs to be neutralized; (2) to better mimic the biological environment of proteins, simulations of soluble proteins are generally run at 150 mM salt concentration. Of course, this can be adjusted to reproduce other experimental concentrations. The program `gmx genion` needs a `tpr` file to generate the ions and modify the topology file. For this purpose, a dummy `tpr` file using the energy minimization settings as before can be composed. The following commands generate the final box.

```
$ gmx grompp -f min.mdp -p system.top -c protein_water.gro
-o dummy.tpr -r protein_water.gro -maxwarn 1
$ echo WN | gmx genion -s dummy.tpr -neutral -conc 0.15
-p system.top -o start.gro
```

Before starting the production simulation, a series of energy minimization and equilibration should be run. This is especially important for large proteins and polymers. First, the final box is energy minimized using flexible bonds and position restraints.

```
$ gmx grompp -f min.mdp -p system.top -c start.gro -o min.tpr -r start.gro
$ gmx mdrun -v -deffnm min
```

Next, a short equilibration of 50 ns applying positional restraints is run. This equilibration allows the water to solvate the polymer and protein. Furthermore, through the use of the Berendsen barostat [40], the simulation will quickly relax to the final volume.

```
$ gmx grompp -f eq.mdp -p system.top -c min.gro -o eq.tpr
-r min.gro -maxwarn 1
$ gmx mdrun -v -deffnm eq
```

Finally, the positional restraints need to be released and another equilibration simulation using the Berendsen barostat should be run. Here the Berendsen barostat is used, as the simulation is more stable than with Parrinello–Rahman [41] pressure coupling.

```
$ gmx grompp -f eq2.mdp -p system.top -c eq.gro -o eq2.tpr -maxwarn 1
$ gmx mdrun -v -deffnm eq2
```

Now, everything is in place to perform the final production run. At this stage, the Parrinello–Rahman pressure coupling [41] should be used. However, before starting the simulation, a few properties should be checked: First open the output file in VMD [42] or any other visualization software and check whether the protein and polymer conformations look reasonable.

```
$ vmd eq2.gro
```

Next, it is useful to compute the box pressure, and temperature average to confirm that all have reached the intended target values.

```
$ gmx energy -f eq2.edr -o energy_eq.xvg
$ gmx analyze -f energy_eq.xvg
```

The values you obtain should be close to 1 bar for the pressure, 310 K for temperature, and a box volume which is constant. If they have converged, the production run can be started. Long simulation times are required to sufficiently sample the polymer conformational space. In the past, sampling times of around 10–30 μ s were used for polymer systems with Martini [19–21, 43]. For this example, the simulation time is set to 2 μ s.

```
$ gmx grompp -f NpT.mdp -p system.top -c eq2.gro -o run.tpr -maxwarn 1
$ gmx mdrun -v -deffnm run
```

After about 2 μ s of simulation, the PEGylated lysozyme has a conformation as shown in Fig. 5. It is clearly not in the shroud conformation. It exists as an extended chain fitting to the dumbbell conformation. This is the same conformation as found by Pai and coworkers for 30 kDa PEGylated lysozyme [4].

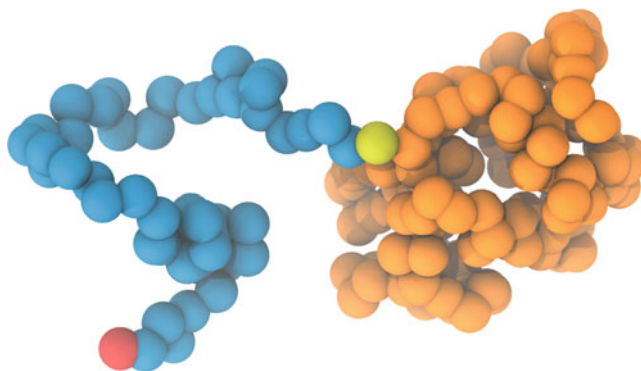


Fig. 5 Conformation of PEGylated lysozyme after 2 μ s of simulation

In this tutorial we have generated Martini 3 parameters for PEGylated lysozyme in three steps: (1) martinize 2 was used to obtain parameters for lysozyme; (2) polyply was used to generate parameters for OH terminated PEG-50; (3) polyply was used again to stitch together the protein and polymer parameters as well as the molecular fragment linking the two. It was also shown how to generate a starting structure for this molecule, and set up as well as equilibrate a simulation. Although a specific example was used to explain this procedure, the protocol is general and can be used for any PEGylated protein. More information about Martini and more tutorials are available on our web page (cgmartini.nl). Any questions regarding the realization of your project with martini, can be posed via our forum. We would also like to encourage reporting of bugs and problems regarding martinize 2 and polyply via the GitHub pages of the two programs.

4 Notes

1. To set up and run simulations on Windows or Mac OS, GRO-MACS, and Python 3 have to be installed. Both have appropriate versions available. Because martinize 2 and polyply are python programs, they can be run in operating system independent from within any python 3 environment. Note that in the tutorial, sometimes bash specific commands (e.g., “egrep” or “awk”) are used. These will only work on Linux OS running a bash shell. However, different solutions can easily be found suiting the OS of interest.
2. Martinize 2 poses some requirements on the input PDB file: There can be no missing residues; C and N atoms, which define the peptidic bond, must be present; and atom names should strictly adhere to the PDB format. In addition, residues in the PDB file are identified by their residue names, and the corresponding information must be present in the library of martinize 2. To add nonstandard residues to the library, please refer to a more specialized tutorial. In case of problems, the flags “-write-graph,” “-write-repair,” and “-write-canon” can be used to write out PDB files of the structure as interpreted by the program at various stages.
3. The bonded parameters of Martini proteins are assigned based on the secondary structure. [16] Usually, the secondary structure of a reference crystal structure is used for this purpose. However, this might not always be appropriate. In such cases, the secondary structure can be provided manually to martinize 2 using the “-ss” flag in a text-based format using a single-letter code. Each letter represents the basic secondary structure elements: H = helix, E = sheet/extended, and C = coil/turn.

For example, a tripeptide in a coil-like structure can have its secondary structure defined with the option “-ss CCC.”

4. All Martini protein models need special interactions to maintain the tertiary structure of the proteins. However, different approaches exist to accomplish this goal. The regular elastic network approach simply applies a bond between all backbone beads within a cutoff. This is sufficient to keep the structure stable and has reasonable properties [33]. However, it has been shown that improved flexibility of the protein structure is achieved with the Go [34] or ELNEDIN [33] approaches. Recently with Martini 3, in combination with the Go approach, it has even been shown that an allosteric pathway can be captured [27]. Therefore, if it is suspected that flexibility of certain domains is important for the PEG protein interactions, using a Go approach would improve the simulations.
5. It is common to treat titratable amino acids with fixed protonation state in MD simulations, even if the protonation states can change. However, it is also known that different environments can affect the pK_a of amino acids and thus their protonation state. It could be possible that PEG, especially in the shroud conformation, modifies the protonation states of titratable groups. If this is suspected, you can gain better insight by using a method with dynamic protonation states, which allows amino acids to change in the course of the simulation. We recently have put forward a proof-of-principle for such a method [44] and the GROMACS lambda dynamics approach [45] is also a suitable option.
6. Polyly has a library of default monomer itp files of different coarse-grained and atomistic polymers that come with the program. To see which files are available run “polyly gen_itp -list_lib.” To use these files, you can simply select the name from the list obtained with the previous command and then use the “-lib” flag instead or together with “-f.” For example, to generate the itp file for the PEG-50 polymer, you could also run:

```
polyly -lib martini3_beta -n_mon 50 -o PEG_50.itp -name PEG
```

7. A tutorial for linkers: <http://cgmartini.nl/index.php/tutorials-general-introduction-gmx5/parametrizing-new-molecule-gmx5>
8. When going from the dilute solution regime into the semi-dilute regime, polymer-polymer interactions become important or even dominating. The crossover point is indicated by the overlap concentration. As the concentration increases beyond the overlap concentration also the properties of

polymers change [46]. It is therefore important to choose an appropriate concentration when comparing to experiments. On one hand, if they are conducted in or extrapolated to the dilute solution regime, the simulation box needs to be sufficiently large. On the other hand, as pointed out by Pai et al., the osmotic pressure in cells is often higher than in dilute solution [4]. Therefore, PEGylated proteins should perhaps also be studied under crowded conditions. In this case, it would be appropriate to add more proteins and/or PEG chains to the simulation box to achieve higher concentrations. *See* ref. 21 on how to calculate the overlap fraction for PEG in water.

9. Polyly can generate structures for disordered, arbitrarily complex, polymers. However, this also means that coordinates for any polymer which has a certain long-range order should not be generated this way. Examples include proteins with a secondary structure, DNA, or polymers with large extended aromatic ring fragments. Another practical limitation applies: Polyly cannot generate polymers which contain virtual sides. Examples include Martini 2 P3HT [47].

References

1. Canalle LA, Löwik DWPM, Van Hest JCM (2010) Polypeptide-polymer bioconjugates. *Chem Soc Rev* 39:329–353
2. Pechar M, Kopečková P, Joss L et al (2002) Associative diblock copolymers of poly(ethylene glycol) and coiled-coil peptides. *Macromol Biosci* 2:199–206
3. Milton Harris J, Martin NE, Modi M (2001) Pegylation: a novel process for modifying pharmacokinetics. *Clin Pharmacokinet* 40:539–551
4. Pai SS, Hammouda B, Hong K et al (2011) The conformation of the poly(ethylene glycol) chain in mono-PEGylated lysozyme and mono-PEGylated human growth hormone. *Bioconjug Chem* 22:2317–2323
5. Daly SM, Przybycien TM, Tilton RD (2005) Adsorption of poly(ethylene glycol)-modified lysozyme to silica. *Langmuir* 21:1328–1337
6. Hamley IW (2014) PEG-peptide conjugates. *Biomacromolecules* 15:1543–1559
7. Munasinghe A, Mathavan A, Mathavan A et al (2019) Molecular insight into the protein-polymer interactions in N-terminal PEGylated bovine serum albumin. *J Phys Chem B* 123:5196–5205
8. Le Cœur C, Combet S, Carrot G et al (2015) Conformation of the poly(ethylene glycol) chains in DiPEGylated hemoglobin specifically probed by SANS: correlation with PEG length and in vivo efficiency. *Langmuir* 31:8402–8410
9. Lin P, Colina CM (2019) Molecular simulation of protein-polymer conjugates. *Curr Opin Chem Eng* 23:44–50
10. Marrink SJ, Risselada HJ, Yefimov S et al (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B* 111:7812–7824
11. Ingólfsson HI, Melo MN, Van Eerden FJ et al (2014) Lipid organization of the plasma membrane. *J Am Chem Soc* 136:14554–14559
12. Thallmair S, Vainikka PA, Marrink SJ (2019) Lipid fingerprints and cofactor dynamics of light-harvesting complex II in different membranes. *Biophys J* 116:1446–1455
13. Bruininks BMH, Souza PCT, Marrink SJ (2019) A practical view of the Martini force field. In: *Biomolecular simulations, Methods in molecular biology (methods and protocols)*. Springer, New York, pp 105–127
14. Souza PCT et al (2020) Martini 3, submitted
15. Alessandri R (2019) Multiscale modeling of organic materials: from the morphology up. Dissertation, University of Groningen
16. Monticelli L, Kandasamy SK, Periole X et al (2008) The MARTINI coarse-grained force field: extension to proteins. *J Chem Theory Comput* 4:819–834

17. Uusitalo JJ, Ingólfsson HI, Akhshi P et al (2015) Martini coarse-grained force field: extension to DNA. *J Chem Theory Comput* 11:3932–3945
18. Lo CA, Rzepiela AJ, De Vries AH et al (2009) Martini coarse-grained force field: extension to carbohydrates. *J Chem Theory Comput* 5:3195–3210
19. Rossi G, Monticelli L, Puisto SR et al (2011) Coarse-graining polymers with the MARTINI force-field: polystyrene as a benchmark case. *Soft Matter* 7:698–708
20. Panizon E, Bochicchio D, Monticelli L et al (2015) MARTINI coarse-grained models of polyethylene and polypropylene. *J Phys Chem B* 119:8209–8216
21. Grunewald F, Rossi G, de Vries AH et al (2018) Transferable MARTINI model of poly(ethylene oxide). *J Phys Chem B* 122:7436–7449
22. Monticelli L (2012) On atomistic and coarse-grained models for C₆₀ fullerene. *J Chem Theory Comput* 8:1370–1378
23. Ramezanghorbani F, Lin P, and Colina CM (2018) Optimizing protein–polymer interactions in a poly(ethylene glycol) coarse-grained model. *J Phys Chem B* acs.jpcc.8b05359
24. Woo SY, Lee H (2014) Molecular dynamics studies of PEGylated α -helical coiled coils and their self-assembled micelles. *Langmuir* 30:8848–8855
25. Zaghmi A, Mendez-Villuendas E, Greschner AA et al (2019) Mechanisms of activity loss for a multi-PEGylated protein by experiment and simulation. *Mater Today Chem* 12:121–131
26. Alessandri R, Souza PCT, Thallmair S et al (2019) Pitfalls of the Martini model. *J Chem Theory Comput* 15:5448–5460
27. Souza PCT, Thallmair S, Marrink SJ et al (2019) An allosteric pathway in copper, zinc superoxide dismutase unravels the molecular mechanism of the G93A amyotrophic lateral sclerosis-linked mutation. *J Phys Chem Lett* 10:7740–7744
28. Liu J, Qiu L, Alessandri R et al (2018) Enhancing molecular n-type doping of donor–acceptor copolymers by tailoring side chains. *Adv Mater* 30:1–9
29. Abraham MJ, Murtola T, Schulz R et al (2015) Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2:19–25
30. Kroon PC (2020) Automate, aggregate, assemble. Dissertation, University of Groningen
31. Touw WG, Baakman C, Black J et al (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res* 43:D364–D368
32. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
33. Periolo X, Cavalli M, Marrink S-J et al (2009) Combining an elastic network with a coarse-grained molecular force field: structure, dynamics, and intermolecular recognition. *J Chem Theory Comput* 5:2531–2543
34. Poma AB, Cieplak M, Theodorakis PE (2017) Combining the MARTINI and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins. *J Chem Theory Comput* 13:1366–1374
35. Herzog FA, Braun L, Schoen I et al (2016) Improved side chain dynamics in MARTINI simulations of protein–lipid interfaces. *J Chem Theory Comput* 12:2446–2458
36. Kuehner DE, Engmann J, Fergg F et al (1999) Lysozyme net charge and ion binding in concentrated aqueous electrolyte solutions. *J Phys Chem B* 103:1368–1374
37. Bartik K, Redfield C, Dobson CM (1994) Measurement of the individual pKa values of acidic residues of hen and turkey lysozymes by two-dimensional ¹H NMR. *Biophys J* 66:1180–1184
38. Stroet M, Caron B, Visscher KM et al (2018) Automated topology builder version 3.0: prediction of solvation free enthalpies in water and hexane. *J Chem Theory Comput* 14:5834–5845
39. Malde AK, Zuo L, Breeze M et al (2011) An automated force field topology builder (ATB) and repository: version 1.0. *J Chem Theory Comput* 7:4026–4037
40. Berendsen HJC, Postma JPM, van Gunsteren WF et al (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81:3684–3690
41. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: a new molecular dynamics method. *J Appl Phys* 52:7182–7190
42. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual Molecular Dynamics. *J Mol Graph* 14:33–38
43. Rossi G, Barnoud J, Monticelli L (2014) Polystyrene nanoparticles perturb lipid membranes. *J Phys Chem Lett* 5:241–246
44. Grunewald F, Souza PCT, Abdizadeh H, et al (2020) Titratable Martini Model for Constant pH Simulations. *J Chem Phys* 153:024118

45. Donnini S, Tegeler F, Groenhof G et al (2011) Constant pH molecular dynamics in explicit solvent with λ -dynamics. *J Chem Theory Comput* 7:1962–1978
46. Colby RH, Rubinstein M (2003) *Polymer physics*. Oxford University Press, New York
47. Alessandri R, Uusitalo JJ, De Vries AH et al (2017) Bulk heterojunction morphologies with atomistic resolution from coarse-grain solvent evaporation simulations. *J Am Chem Soc* 139:3697–3705



Chapter 19

Molecular Data Visualization on Mobile Devices: A Quick Starter's Guide

Chin-Pang Benu Yiu and Yu Wai Chen

Abstract

With the rise of tablets, truly portable molecular graphics are now available for wide use by scientists to share structural information in real time. We have surveyed the existing software available on Apple iPads and on Android tablets in order to make a recommendation to potential users, primarily based on the product features. Among the three apps for high-quality 3-D display, *iMolview* (available on both platforms) stands out to be our choice, with *PyMOL* app (iOS) a close alternative and *NDKmol* (Android) offering some uniquely useful functions. Hence we include a tutorial on how to get started using *iMolview* to do some simple visualization in 10 min.

Key words Protein structure, RCSB PDB, Protein Data Bank, Macromolecular graphics, Tablets, Mobile devices, Protein structure

1 Introduction

Molecular graphics is the language of structural biologists. In the past few years, the portable computers world witnessed the rise of the thin and light-weight handheld tablets. These are portable computers in every sense, without keyboard or mouse, thanks to a touch-sensitive screen. For instance, the Apple iPads have a large screen of very high sensitivity and resolution (ranging from 9.7-in. models of 2048×1536 pixels to 12.9-in. iPad Pro of 2732×2048 pixels, at 264 pixels per inch (ppi)). Since their inception, iPads have been well received by consumers, which encouraged software development on the iOS and the latest iPadOS (the operating systems on Apple mobile devices) platform. On the other hand, many rivals to iPads have been developed; these devices mostly adopt the Google Android operating system, which is based on Linux. Together, these mobile devices completely revolutionized how users interact with computers, in more intuitive ways using finger gestures.

In this article, we shall compare the currently available molecular graphics products on the iPads and the Android tablets that can be used to visualize protein structures deposited at the RCSB Protein Data Bank (PDB). Among these, we shall recommend the best all-round graphics software. Next, we shall discuss how to set up and perform some very basic visualization tasks. We aim to get people who are not familiar with molecular graphics to start using it on their mobile devices.

2 Graphics Software

2.1 Hardware Used for Testing

The iOS apps (application software on mobile devices) were tested on iPad Air 2 (A8X CPU, 128 GB, iOS 12.3.1) and iPad Pro (A12X CPU, 256 GB, iOS 12.3). For the Android platform, we tested with an average phone (quad-core CPU, 6 GB RAM + 128 GB ROM, customized Android 9.0) with a 5.5-in. display of 1080 × 1920 pixels, at 401 ppi pixel density.

2.2 Comparison of Graphics Software

In the first edition of this book chapter (2012), we identified 12 mobile graphics apps on the market. At 2019, the fierce competition is over, and only a few survived. Here, we shall give an updated account of the three products: *iMolview*, *PyMOL* and *NDKmol* (Table 1), which we noted previously for their better and/or unique product features (restated below).

We performed a comparison of the essential functions offered by the three molecular graphics apps (Table 2). Note that this is a features comparison and computing performance was not vigorously tested. All apps offer the basic control operations (rotate, translate, zoom and clip). We used the crystal and NMR structures of the p53 tetramerization domain (PDB ID 1AIE and 2J0Z, *see Note 1*), a small protein of 31 residues (monomer) or 124 residues (tetramer) for testing.

From Table 2, *iMolview* and *PyMOL* compare similarly, and both offer the full set of features to satisfy most structural

Table 1
Basic information of standalone mobile apps for macromolecular graphics

App	iOS	And.	Price (\$)	Developer	Version (updated)
<i>iMolview</i>	●	●	Free (lite) 0.99 (full)	Molsoft	1.9.4 (2019) 1.9.5 (3/2019)
<i>PyMOL</i>	●		Free	Schrödinger	1.7.6.5 (2016, halted)
<i>NDKmol</i>		●	Free	biochem_fan	0.97 (3/2018)

The respective versions reviewed are the latest at the time of writing (August 2019). “And.” stands for Android

Table 2
A comparison of main features of four molecular graphics apps

Feature		<i>iMolview</i> ^a	<i>PyMOL</i>	<i>NDKmol</i>	<i>Miew</i> ^b
Structural object styles	Ball-and-stick	●	●	●	●
	Space-filling	●	●	●	●
	Ribbon/cartoon	●	●	●	●
	Wire/stick	●	●	●	●
	Surface	●	●		●
	B-factor putty		●	●	
Custom color	Background	●	●		○
	Graphical object	●	●	●	●
Label		●			○
Selection	To act on a subset	★	○		●
Sequence view		★			
Biological assembly				★	★
Measure	Distance, angle	●	●		
View and render	Center on atom	●	●		●
	Stereo	●	●		
	Ray trace		★		
	Fog/clip	●	●	●	●
	Rock/spin	●	●		
Load/import	PDB	●	●	●	●
	Local import	●	●	●	●

Filled circle: feature available; *filled circle with a star*: unique feature; *open-circle*: a feature that is partially available or problematic. Note that each software has additional advanced features (e.g., transparency, molecules alignment, scripting) that are not included here, please refer to the respective developer's webpage

^a*iMolview* full version (iOS)

^b*Miew* was tested on a desktop computer. On a mobile device (iPad Air 2), it caused many unrecoverable error messages (e.g., loading an NMR structure, turning on ambient occlusion)

biologists' needs. We found *iMolview* easier to use and it offers a convenient "sequence view" which enables quick access to any residue in the structure. *PyMOL* on desktop computers is one of the most popular molecular graphics software, and its app excels in producing ray-traced photorealistic scenes. However, the app does not inherit the intuitive way of selection of subsets of atoms for rendering. Unfortunately, the development of *PyMOL* app has halted since 2016. Users can still download it but without support. *NDKmol* is unique in being able to display the biological assembly of a crystal structure. This is best illustrated with PDB ID 1AIE. While the other two apps show only the 31-residue monomer in the crystallographic asymmetric unit, with a few clicks, *NDKmol* displays this as a tetramer (Fig. 1), correctly taking crystallographic symmetry and oligomerization information into account.

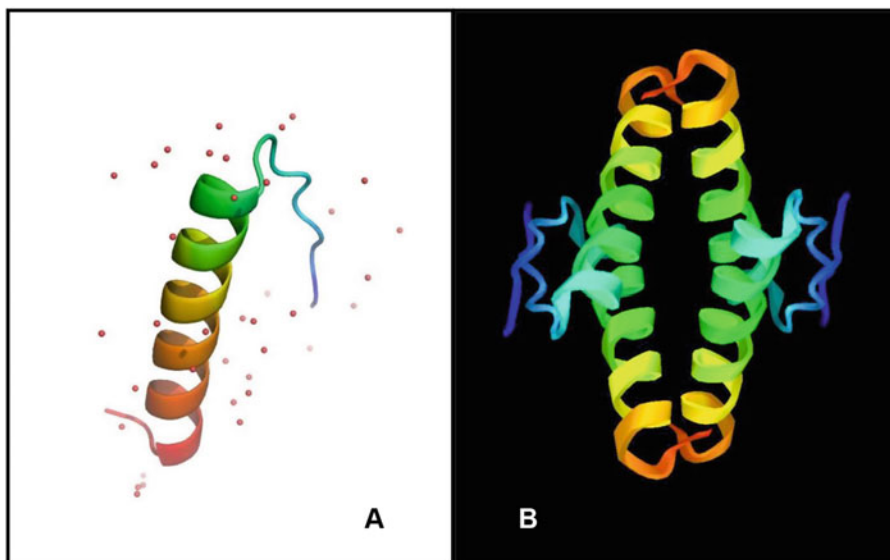


Fig. 1 *PyMOL* app and *NDKmol* compared. The crystal structure of p53 tetramerisation domain (PDB ID 1AIE) rendered in (a) *PyMOL* app (ray traced) and (b) *NDKmol*. *NDKmol* is the only utility which allows convenient viewing of the biologically functional tetrameric arrangement instead of just the monomer in the crystallographic asymmetric unit

The quality of the graphic images produced by various apps is ranked in descending order, as follows: *PyMOL* (ray tracing, Fig. 1a), *iMolview*, *NDKmol* (Fig. 1b).

If one needs a handy tool to import a PDB file and get an overall view of the protein fold with cartoon or ribbon style, then any of these apps can serve the purpose. We are also aware that the RCSB PDB website has already implemented its own server-side graphics utility to provide a quick view of a deposited structure (*see* later, in Subheading 4). The deciding factor of a good structural biologist's mobile tool is whether it allows the user to select a subset of atoms for rendering. For this, only *iMolview* is suitable.

3 Methods

iMolview can be used with or without the internet. An active connection is required to import structures from the PDB. After that, structure viewing, analysis and rendering can be performed offline (without internet).

3.1 Installation

iMolview is available (\$0.99) in the Apple App Store for iPhones and iPads. A "Lite" (free) version is made available in Google Play for Android devices. We did not test the iOS Lite (free) version.

3.2 Importing a PDB Entry

1. Make sure there is an active internet connection (Wi-Fi or mobile data).
2. Tap the top search bar, and enter some search criteria into it. For this tutorial, type “p53 tetramerization.” As the text is typed in, a dropdown menu appears listing all the entries that satisfy the search text string. Tap on the entry starting with “2J0Z”, which is the PDB ID for this structure. Alternatively, type “2J0Z” directly if the PDB ID is known. This is the solution NMR structure of the tetramerization domain of the p53 tumor suppressor.
3. A representation of the ensemble structure appears on the screen (Fig. 2). The default style is the Richardson protein secondary-structure cartoon (*see Note 2*). At the bottom of the screen, the single-letter-protein sequence is shown, with residue numbers, and color-coded according to secondary structures (strands: green, helices: red). If there are multiple protein chains in the crystal structure, each chain is represented



Fig. 2 Default *iMolview* display. The default display in *iMolview* of the p53 tetramerization domain (PDB ID 2J0Z) as Richardson secondary-structure cartoon. At the bottom of the display screen, the protein sequence is shown and color-coded by secondary structure (strands: green; helices: red), with residue numbers and chain tabs (“a,” “b,” “c,” “d”)

by a tab with a unique chain identifier (e.g., “a,” “b,” “c,” “d,” ... here) at the very bottom. This NMR structure has 30 models, each has A, B, C, and D chains. Hence there are altogether 120 chain tabs. Chains from each NMR model are indicated with the model number at the end (e.g., “a2” is the A chain of model 2). One can tap on these chain tabs to quickly show or hide a chain.

4. We want to show only the most representative model (model 1 of the PDB file). There is a quick way to hide all chains. Click on the button at the lower right corner which has an icon like a numbered list. This opens a menu as shown in Fig. 2.

There is a line showing the PDB entry 2J0Z with a gray-outlined blue square which has a white dot at the center. This means all chains of 2J0Z are shown. Tap on this square to turn it off. This immediately blanks the screen (hide all chains).

5. Click on the “a” tab to the far left of the bottom to select it, click again to show the chain A of model 1. A blue square with white dot will show in the A chain (“a”) tab to indicate that this chain is now visible. Do the same to show the B, C, and D chains of model 1 (Fig. 3).

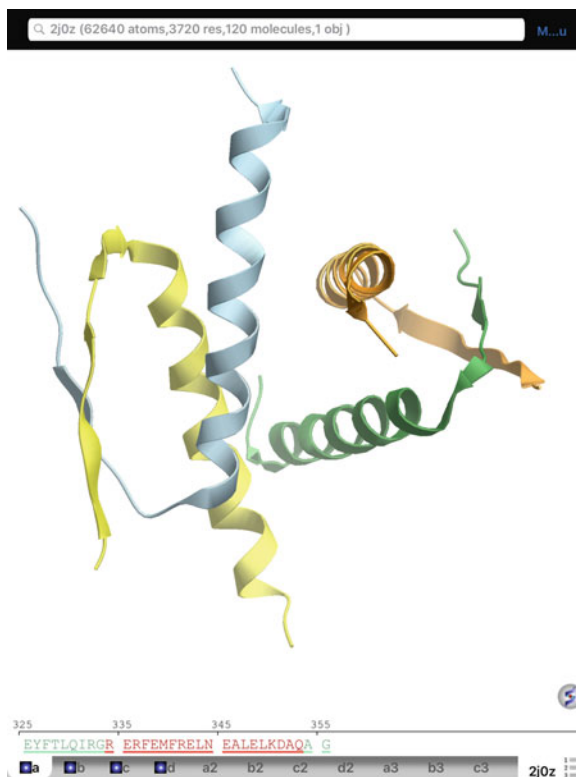


Fig. 3 Model 1 of an NMR ensemble is shown in *iMolview*

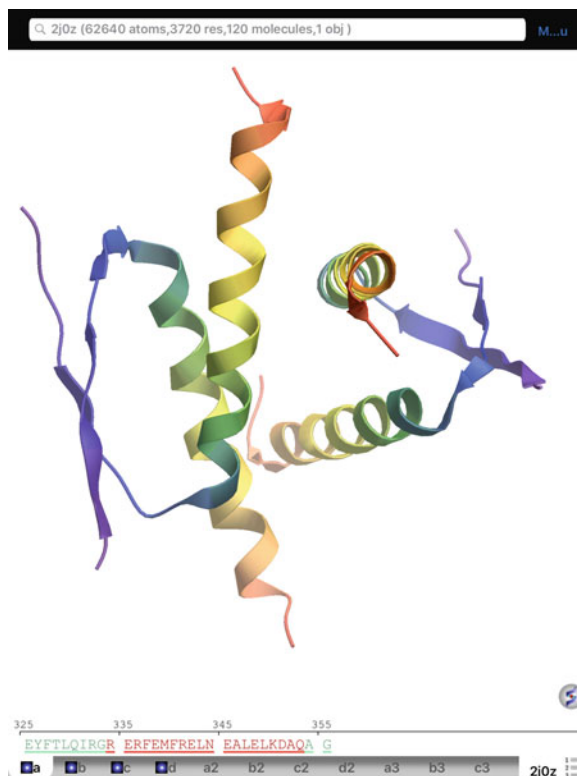


Fig. 4 Rainbow color display in *iMolview*

3.3 Viewing with Different Styles

3.3.1 Rainbow Coloring (Blue to Red) from N- to C-Termini

Tap **Menu** button (top right), tap “**Color Object/Selection by >**”; on the next menu, tap the first item “**NtoC**”. It is necessary to turn off all chains and show A, B, C, D chains (Fig. 4) of model 1 again (Subheading 3.2, steps 4 and 5).

3.3.2 Transparent Items

Tap the **Menu**. Make sure you are at the top level of the menu (you will see “**Display**” as the first item in this menu). If you followed the tutorial strictly up to Subheading 3.3.1, you would find yourself at an inner menu level, then you need to tap the **Back** button at the top to return to the top level (“Main Menu”). Tap “**Settings >**”; on the next menu, slide the “Transparent Ribbon” to **ON** (default is OFF; Fig. 5).

3.3.3 Molecular Surface

1. At the bottom of the screen, tap and hold the D chain (“d”) tab. This selects the whole D chain, and the selected atoms are represented by small green crosses.
2. Tap the **Menu**. Again, you may need to tap “**Back**” to get to top level.

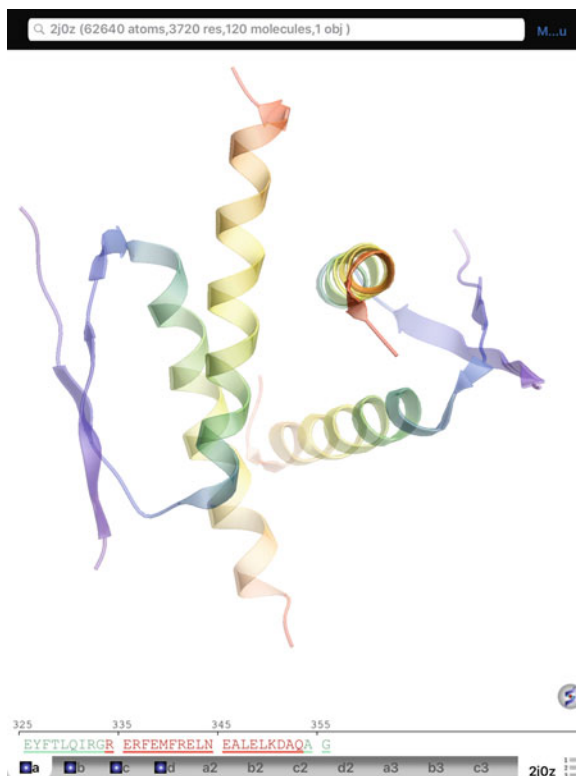


Fig. 5 Transparent ribbon display in *iMolview*

3. To the right of the “Display,” tap the fifth icon (*see Note 3*) for the surface representation. Tab and hold the bottom “d” tab to unselect the atoms (Fig. 6).

3.4 Exporting an Image

On the iPad, this is very easy. Just press on/off button and the main button together, a screenshot will be saved to the iPad’s photos storage. The image can then be shared with other mobile devices (*see Note 4*).

4 Conclusion

Portable molecular graphics has now entered into a mature phase. Finally, scientists can carry molecular models around and show these to their colleagues. The models can be examined in real time, using natural hand and finger manipulations. Among the software available, the low-cost *iMolview* tops the list because of its user-friendliness and it offers the complete set of functions for visual communication. *PyMOL* for iPad is still a useful tool, but it is no longer under development. *NDKmol* for Android is currently still in alpha testing.

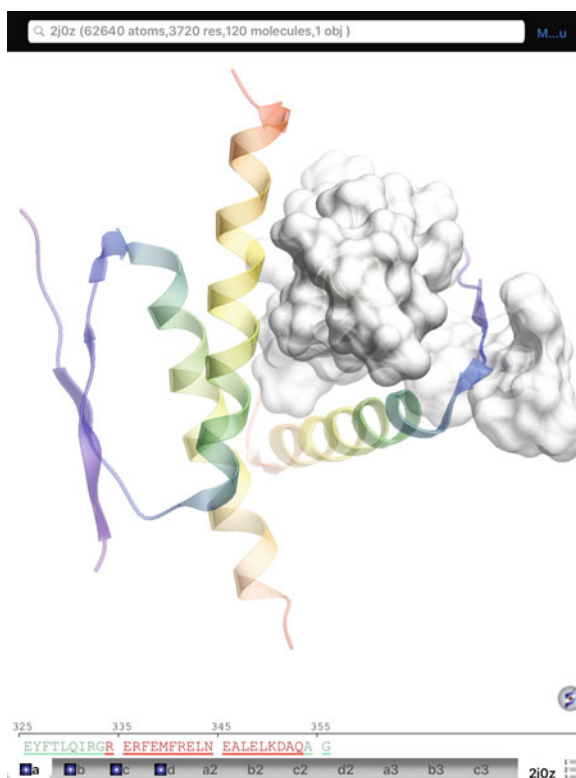


Fig. 6 Composite objects in *iMolview*. Molecular surface display of the D chain of 2j0z is shown. The D chain is accessed by the “d” tab at the bottom

We hope with the primer we have demonstrated how easy it is to use *iMolview* to create a molecular scene of mixed styles, and it can help some colleagues to start using their tablets in visualizing and communicating structures.

Many projects that started before 2012 have now been taken off the shelf. At 2019, the current trend is for graphic viewing utilities to be built into a web server. A good example is the RCSB PDB web site, which incorporates the NGL viewer into its user interface. With that, the previous *RCSB PDB Mobile* app has been discontinued since 2016 (*see Note 5*). Another excellent utility that is under development is the *Miew* viewer (*see Notes 6 and 7*), which is a free and open-source project. We included its current features in Table 2 for comparison to show that it can rival the best standalone app.

5 Notes

1. PDB ID is a unique four-alphanumeric character combination that is assigned to each deposited structure in the Protein Data Bank (www.rcsb.org). This ID is usually found in the manuscript that describes that particular structure.

- Richardson cartoon style is a representation of the overall backbone structure of the protein, with secondary-structural elements α -helices shown as coiled ribbons, β -strands shown as flat arrows, and coils/loops shown as thin tubes.
- The complete “user manual” of *iMolview* is accessed by the **Menu**, then “**Help**” inside the app.
- The users can transfer PDB files or images between mobile devices using Bluetooth-based Apps such as *iShareFiles* (free), without internet, or *AirDrop* in the iOS and macOS provided by Apple, with both Wi-Fi and Bluetooth. The files can also be sent via email.
- <https://www.rcsb.org/pdb/static.do?p=mobile/RCSBapp.html>
- Developers’ websites:
iMolview: www.molsoft.com/iMolview.html
PyMOL app: pymol.org/mobile
NDKmol: webglmol.osdn.jp/android-en.html
Miew: epa.ms/miew
- The Miew viewer is fully functional: <https://miew.opensource.epam.com/>

Acknowledgments and Declaration

The authors share no commercial interests in the software or hardware described in this article. Y.W.C. acknowledges support from the Innovation and Technology Commission of Hong Kong, the Hong Kong Polytechnic University and the Life Science Area of Strategic Fund 1-ZVH9.



Molecular Data Visualization with Augmented Reality (AR) on Mobile Devices

Chin-Pang Benu Yiu and Yu Wai Chen

Abstract

Augmented reality (AR) allows a computer-generated 3D model to be superimposed onto a real-world environment in real time. The model can then be manipulated or probed interactively as if it is part of the real world. The application of AR in visualizing macromolecular structures is growing, primarily in showing preset collections of scenes for education purpose. Here, our emphasis is, however, on exploiting AR as a tool to facilitate scientific communication on the go. We have searched for freely available mobile software and custom-built tools which allow the display of user-specified protein structures. We provide step-by-step guides on a standalone app *Ollomol* (iOS and Android), as well as an in-browser web app, *WebAR-PDB*. Both of them allow users to specify entries from the Protein Data Bank (PDB) for an elementary AR experience. The application of AR enhances interactivity and imaginativity in macromolecular visualization.

Key words RCSB PDB, Protein Data Bank, Molecular graphics, Tablets, Mobile devices, iPad, iPhone, iOS, Android, Augmented reality

1 Introduction

Since molecular graphics is the language of structural biologists, it is most communicative when it can be used anywhere in an interactive way. In the past few years, augmented reality (AR) has been developed to allow computer-generated molecular models to overlay with objects in the real-world environment [1–3]. This expands the imaginary space by offering an opportunity to align and compare virtual objects with real ones, of sizes orders of magnitude apart, in real time. With mobile devices, we can now carry the macromolecules around and show them against different real backgrounds.

In this article, we shall demonstrate two AR products on the iPads and Android tablets that use Protein Data Bank (PDB) files as input for visualizing macromolecular structures. The first one is a standalone app (application software). The other one is an in-browser web app (an application hosted by a web server). Both allow

Table 1
Basic information of mobile apps for macromolecular graphics in AR

App	iOS	And.	Price (\$)	Developer	Version (updated)
<i>Ollomol</i>	●	● ^a	Free	MD.USE Innovative Solutions SL	iOS—1.01 (3/2019) And.—1.0 (2/2019)
<i>CRISPR-3D</i>	●	●	Free	Innovative Genomics Institute	iOS—1.5.2 (8/2019) And.—1.5.3 (8/2019)
<i>BioChemAR</i>	●	●	Free	Carleton College	iOS—1.2 (1/2019) And.—1.2 (7/2019)

The respective versions reviewed are the latest at the time of writing (Summer 2019). “And.” stands for Android

^aDownloadable but only works occasionally. We contacted the app developer but have not received their response by the time of submission of this chapter

users to see real-time models on real-life backgrounds. We aim to encourage people who have no prior knowledge of AR to experience this emerging technology.

While searching for AR apps, we found three items in the Apple App Store or Google Play as shown in Table 1. *CRISPR-3D* and *BioChemAR* are good at displaying specific molecular structures already defined in the apps. In *CRISPR-3D*, users may select from the 14 preset structures (protein, RNA, or DNA) for AR viewing. Structures can be displayed, translated, rotated, and scaled easily and intuitively with finger gestures and shown around in different places along with the mobile device. With a pre-generated in-app QR code, *BioChemAR* displays the digital 3D model of the *KscA* potassium channel structure and its moiety. By contrast, *Ollomol* (see Note 1) permits users to import structures of their choice from the PDB (see Note 2) and generate AR models in a few simple steps. The app is, therefore, our recommended standalone app (Subheading 3).

Web apps provide an alternative way to display protein structures readily in AR from PDB files [4]. They are used within an internet browser, such as *Safari* or *Firefox*, without the need to install any software. The *WebAR-PDB* web app [5], developed by Luciano Abriata at École Polytechnique Fédérale de Lausanne, Switzerland, is freely accessible on the Web-based Augmented Reality for Chemistry and Structural Biology (*WARCSB*) portal (see Note 3). Users need to first download the PDB file of choice to the mobile device. Direct download by PDB ID is not available. The guide to using the *WebAR-PDB* web app is presented in Subheading 4.

AR is a powerful tool for enhanced immersiveness and interactivity [1–5]. At the time of writing (September 2019), several apps are available for viewing protein structures. While some apps do offer limited display style options, none of them possesses those subset rendering features of dedicated molecular graphics software

as we described in another chapter [6]. Hence this chapter serves as an early introduction to an emerging technology. There ought to be dramatic improvements in the near future.

2 Hardware and Software Used for Testing

The iOS apps were tested on iPad Air 2 (A8X CPU, 128 GB, iOS 12.4.1) and iPad Pro (A12X CPU, 256 GB, iOS 13.0). Most of the software written for the iOS should be able to run on an iPhone. For the Android platform, we used a OnePlus A3003 phone (6 GB, Android 9 OxygenOS 9.0.3). *Ollomol* (iOS version 1.01, released on 3/2019; Android version 1.0, released on 2/2019, MD.USE Innovative Solutions SL) and *WebAR-PDB* (version last tested last updated in June, 2019) were described herein.

3 A Tutorial of *Ollomol*, a Standalone App

Ollomol can be used with or without the internet. An active connection is required to download structures from the PDB. After the PDB file has been imported, structure viewing, examination and manipulation can be performed offline (without internet).

3.1 *Ollomol* Installation

Ollomol is available for free in the Apple App Store for iPhones and iPads and in Google Play for Android devices. At the first use, a request will pop up for granting access to the device's camera and storage. Answer "Yes" to authorize.

3.2 Generating a Tracker Pattern for a PDB Entry

1. Make sure there is an active internet connection (Wi-Fi or mobile data).
2. Tap on "AR" to select the AR mode.
3. Tap on "OPEN MENU" on the top left corner. A list of options should show up on the left side of the screen.
4. Tap on "CREATE QR" in the middle of the list to bring up the Tracker generator. Another list of options should become available on the left.
5. Now tap on "...PDB..." to enter a PDB ID. In this tutorial, we shall use the PDB entry 1SN4. Note that it is necessary to delete the characters "...PDB..." which are present by default and type in "1SN4" (without the double quotes) (*see Note 4*).
6. Then tap on "1 - Create tracker" to generate a QR-based "tracker" for the PDB entry (*see Note 5*). As shown in Fig. 1a, a tracker will appear on the right side of the screen.
7. To save, export (e.g., by email) or print the tracker, tap on "2-Save tracker" (*see Note 6*).
8. Tap on "Return to AR" at the lower left to exit the Tracker generator.



Fig. 1 Screenshots of *Ollomol* displaying PDB ID 1SN4. (a) The correct way to enter the PDB ID. The generated tracker pattern (jelly fish cartoon and QR code) is shown on the right side of the screen. (b) Cartoon displays of 1SN4 with the helix as red coil, strands as blue arrows and loops are in white. The tracker was placed on a lawn background

3.3 Viewing the Molecular Structure with AR in Ollomol

1. Display the tracker (saved in **step 7** of Subheading 3.2) on a second device, or print it on paper to use it.
2. Place the tracker on a physical surface at where the macromolecule will be shown.
3. On the device where *Ollomol* is running, from the main AR screen (where it should be, following the last step in the previous section), tap on “OPEN MENU” and choose “SCAN QR.”
4. Now, one can see the real-life environment through the device’s camera. At the center of the screen, there are corner marks defining a square scannable area. A top bar appears showing the words “Code Find: ...” Move the device and orient the camera to scan the entire tracker of **step 2** (see **Note 5**).
5. Once successful, the top bar reads “Code Find: 1SN4” and the “Load molecule” bar simultaneously shows up at the lower-left corner.
6. Tap on “Load molecule” and wait. The 3D model of the scorpion neurotoxin, BmK M4, will appear in AR against the real background (Fig. 1b).
7. To re-orientate this digital model, rotate the tracker or the mobile device (see **Note 7**).
8. Four rendering options are available for the molecule: Cartoon, Ball, Stick and Ball&Sticks. By default, all styles are shown. Select or deselect by clicking on the options on the right to alter the display.

4 A Tutorial of *WebAR-PDB*, a Web-Server Application

This is a server-side application that requires no installation. All tasks are performed via a web browser. A paper cube marker needs to be constructed for manipulation of the AR model.

4.1 Make the Cube Marker

1. Make sure there is an active internet connection (Wi-Fi or mobile data).
2. Open the *Safari* web browser (*see Note 8*).
3. Go to the WARCSB portal using the following URL link (*see Note 3*): <https://lucianoabriata.altervista.org/jsinscience/arjs/armodeling/>
4. Download the marker provided by the site by clicking the “cube marker” hotlink. This will lead to a *GitHub* page. Click the “Download” button therein. A cube marker image (*cube.png*) will now appear in the browser. Print this.
5. Alternatively, access the following link and download the cube marker file for printing: <https://lucianoabriata.altervista.org/jsinscience/arjs/markers.docx>
6. Cut out and fold up the printed image to construct the 3D cube marker as shown in Fig. 2a.

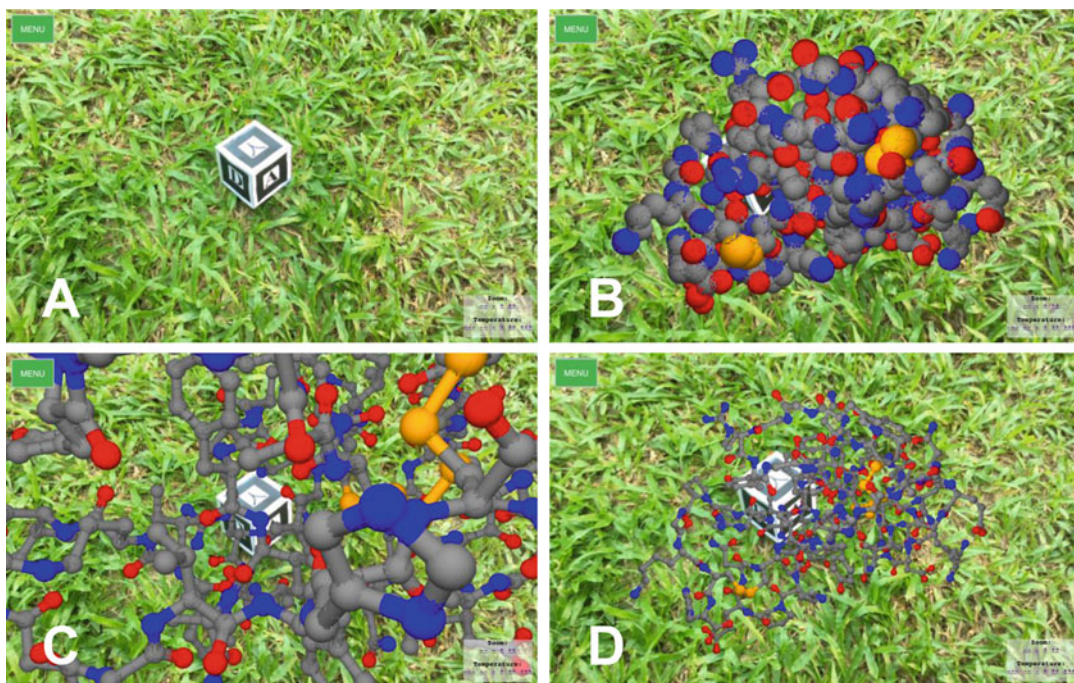


Fig. 2 Screenshots of *WebAR-PDB* displaying PDB ID 1SN4. (a) The tracker cube placed on a lawn background. (b) Spheres display of 1SN4. (c, d) Balls-and-Sticks display of 1SN4 at different zoom levels. (b–d) are all from the same viewing angle. The water oxygen atoms and ligand atoms in the PDB file were removed by editing

4.2 PDB File Download

1. *WebAR-PDB* does not support direct loading from the RCSB PDB. First, download the PDB file to the local device (*see Notes 6 and 9*).
2. On the mobile device, with a web browser, access the RCSB PDB website: <https://www.rcsb.org>
3. On the RCSB PDB search bar, type the name or PDB ID (*see Note 2*) of the molecule that is desired for display. For this tutorial, type “1SN4” (without the double quotes) and click “return.”
4. On the top right, click “Download Files” and select “PDB Format” from the pull-down menu. The PDB file will be automatically downloaded to the device’s default location.

4.3 Viewing the Molecular Structure with AR

1. On the mobile device, use the browser (*Safari* on the iOS) to visit the *WARCSB* portal (Subheading 4.1, step 3).
2. Go to the *WebAR-PDB* page by clicking on “Open any PDB file and handle in 3D with a cube marker” or the following URL link: <https://lucianoabriata.altervista.org/jsinscience/arjs/jsartoolkit5/pdbloader5.html>
3. The browser will now ask for access to the device camera. Tap “Allow.” The back camera will be turned on.
4. On the top right, there is a window inset. Unless the inset is closed (*see Note 10*), tap the “Choose File” button and then navigate to the download location (Subheading 4.2, step 4). Select the downloaded PDB file, 1sn4.pdb (*see Note 9*). The text window will be filled with the top header lines of the 1sn4.pdb file. This indicates successful loading.
5. Now select a display option from between “Spheres” and “Sticks” (default) near the bottom of the window inset.
6. Place the cube marker from Subheading 4.1 against a real-life background of choice (Fig. 2a).
7. Tap on the “Start AR” button in the inset to generate the 3D model. The macromolecular structure will appear on top of the cube marker (Fig. 2b, c) (*see Note 11*).
8. The displayed structure can be viewed at different zooming levels. Use the four “Zoom” options at the bottom right or “Zoom in” and “Zoom out” options under the pull-down menu on the left of the screen for control (Fig. 2d).

5 Discussion

AR is an emerging technology which will undoubtedly find many applications in molecular science research. Some of its capabilities are demonstrated by a number of existing AR apps in displaying built-in collections of structures or specific molecular scenes.

Ollomol, however, allows the loading and viewing of protein structures deposited at the RCSB Protein Data Bank. This standalone app also provides limited rendering styles to the displayed macromolecule. While it is under active development, we encountered problems when NMR ensembles and very large structures were used.

At present, both software allow the user to display only the entire molecule, in a few graphic styles. There is not much customization available. *WebAR-PDB* requires the PDB file to be downloaded to the mobile device first, which is one more step compared with *Ollomol*. However, this becomes an advantage as it allows the user to edit the PDB file for customization (*see Note 9*). When the software matures, this should not be needed—graphic customization should be offered within the AR viewer.

The *WARCSB* portal amasses many examples of how AR can be exploited in chemistry (*see Note 3*). It is possible to bring in two (simpler) models, each handled by a 2D tracker, and perform interactive (real time) scientific studies between them. The user can print out the trackers and try out an AR experience in calculations involving molecular dynamics, small-angle X-ray scattering and electrostatics [5]. The use of handheld trackers for structural display is intuitive on one hand. On the other hand, the use of AR in this manner will allow more immersive, interactive, and even collaborative molecular modeling [5].

Notably, there are alternative ways of molecular structure visualization at a better quality, which also take structural information from PDB for use in AR applications. Crow described a method that essentially involves the using a molecular graphics tool, *PyMOL*, assisted by *Meshlab* and *Blender*, to produce a 3D model for viewing in AR using *Augment* (*see Note 12*) [7]. *PyMOL* is one of the most popular molecular graphics program. The user can produce a customized 3D model to very high quality. The trade-off of this method is that the preparation, involving multiple files and multiple tools, needs to be performed on a desktop or laptop. We followed this method to create a model of the test molecule, scorpion neurotoxin (PDB ID 1SN4), for comparison (Fig. 3).

Instead of *PyMOL*, Poh et al. [8] used a method that employs *UCSF Chimera* for 3D model generation and *APD AR Holistic Review* or *HP Reveal* for structure viewing (*see Note 13*). An AR image will appear nicely as the 3D projection of a publication figure when its graphic is used as a tracker [8].

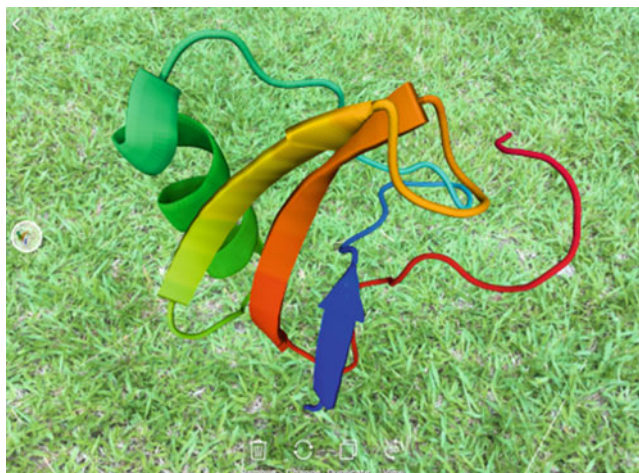


Fig. 3 Display of a model of 1SN4 in the *Augment* viewer on iPad. The model was prepared in *PyMOL* and represented in a secondary-structure cartoon style and colored in a spectrum from blue (N-terminus) to green (middle) to red (C-terminus). The view angle is similar to that in Fig. 1b for comparison

6 Notes

1. *Ollomol* developer's website: <http://ollomol.mduse.com/>
2. PDB ID is a unique four-alphanumeric character combination that is assigned to each deposited structure in the Protein Data Bank (www.rcsb.org). This ID is usually found in the manuscript that describes that particular structure.
3. The Web-based Augmented Reality for Chemistry and Structural Biology (*WARCSB*) portal is a hub hosting several AR web apps under development: <https://lucianoabriata.altervista.org/jsinscience/arjs/armodeling/>. These web apps are built using open web software, such as A-Frame, AR.js and Cannon.js, around the programming language, JavaScript [4, 5]. Luciano Abriata illustrates their uses via a YouTube channel that he maintains: <https://www.youtube.com/channel/UCdhOVimtNZmri967KdTEAKg>
4. If the PDB ID is entered incorrectly, e.g., if it contains extra characters, a dysfunctional pattern may be generated. No 3D model will be visible when this tracker is used. In such case, go back and make sure the PDB ID contains only the four alphanumeric characters.
5. The entire tracker image, not just the QR code, is required.
6. The users can transfer images e.g., trackers or PDB files between mobile devices using Bluetooth-based apps such as

iShareFiles (free), without internet, or *AirDrop* in the iOS and macOS provided by Apple, with both Wi-Fi and Bluetooth. The files can also be sent via email.

7. Users can probe the digital model interactively, in different sizes and orientations, by changing the relative positions (distances and angles) of the tracker and the viewing device.
8. Both *Safari* and *Chrome* were tested. At the time of writing (September 2019) only *Safari* worked. Some web browsers may need a tap on the screen for the webcam to start according to the *WARCSB* portal.
9. A PDB file needs to be downloaded onto the local device, which means that it can be edited for customization e.g., removal of waters, as for Fig. 2 or selecting one representative model out of an NMR ensemble, or display one out of multiple chains in an asymmetric unit.
10. To reload the inset window, tap on the “MENU” button and choose “Load PDB. . .” from the pull-down menu.
11. Sometimes errors in a browser may lead to recurring restarts of the *WebAR-PDB* page. A hard reset of the device may be needed for a cleanup.
12. PyMOL: <https://pymol.org>
 Meshlab: <http://www.meshlab.net>
 Blender: <https://www.blender.org>
 Augment: <http://www.augment.com> (14-day-free educational license)
 A tracker is anything that is marked with a unique pattern. The *Augment* app will recognize and keep track of it, via the device’s camera, for AR display.
13. UCSF Chimera: <https://www.cgl.ucsf.edu/chimera>
 APD AR Holistic Review: <https://www.apdskeg.com/Apps/Information>
 HP Reveal: <https://www.hpreveal.com>

Acknowledgments and Declaration

The authors share no commercial interests in the software or hardware described in this article. Y.W.C. acknowledges support from the Innovation and Technology Commission of Hong Kong, the Hong Kong Polytechnic University and the Life Science Area of Strategic Fund 1-ZVH9.

References

1. Berry C, Board J (2014) A Protein in the palm of your hand through augmented reality. *Biochem Mol Biol Educ* 42(5):446–449
2. Garcia-Bonete MJ, Katona MJG (2019) A practical guide to developing virtual and augmented reality exercises for teaching structural biology. *Biochem Mol Biol Educ* 47(1):16–24
3. Safadel P, White D (2019) Facilitating molecular biology teaching by using Augmented Reality (AR) and Protein Data Bank (PDB). *Tech-Trends* 63:188–193
4. Abriata LA (2017) Web apps come of age for molecular sciences. *Informatics* 4(28):1–18
5. Abriata LA (2018) Towards commodity, web-based augmented reality applications for research and education in chemistry and structural biology. arXiv:1806.08332 <https://arxiv.org/pdf/1806.08332.pdf>
6. Yiu CPB, Chen YW (2020) Molecular data visualization on mobile devices: a quick starter's guide, Chapter 19. In: *Structural genomics*. Humana, New York
7. Crow A (2018) How to display full colour proteins in augmented reality with the augment app. https://www.dropbox.com/s/s2tgk85luvacaIm/Colour_Augmented_Reality_Proteins.docx?dl=0
8. Poh JJ, Phua SX, Chan KF, Gan SKE (2018) Commentary: Augmented Reality Scientific Phone Apps - making the APD AR Holistic Review app and using existing AR apps for scientific publications. *Sci Phone Apps Mobile Devices* 4:4

INDEX

A

- Acid-stable proteases 162, 169, 170
- Active-site 13–21, 249, 289
- Active-site labelling 13–21
- Adherent cultures 79, 81, 102
- Agarose gels 27, 40, 42
- Aggregates 58, 59, 64, 65, 113, 123, 153, 184, 186, 192, 221, 225
- Aggregation 65, 153, 180, 184, 193
- Algal extract 139–141, 146
- Alignment, PIR format 244, 245
- Alternative conformations 266
- Amber force fields 268, 269
- AmberTools 258–260, 266, 268, 269, 273, 291
- Amino acid (residue)
- arginine 321
 - aromatic 129, 184
 - aspartic acid 141, 321
 - cysteine 141, 290, 320, 321
 - glutamic acid 321
 - histidine 137, 141
 - lysine 137, 321
 - methionine 137, 141, 258
 - perdeuterated 129, 137, 139
 - titratable 321, 332
 - tyrosine 137, 186
 - unnatural 260, 268, 269, 272
- Ammonium sulfate precipitation 136
- AmpC β -lactamase 14, 16, 17
- Ampicillin 42, 131, 134
- Amyloid fibrils 177, 184
- Android tablets 338, 347
- Antechamber* 259, 268, 290, 291
- AnteChamber PYthon Parser interface (ACPYPE) 290–294, 297, 298, 305, 307
- Antibiotics
- ampicillin 42
 - carbenicillin 41, 42
 - gentamycin 70, 75
 - kanamycin 28, 31, 36, 41, 47, 52, 53, 55, 70, 75, 77
 - penicillin 16, 71, 278
 - streptomycin 71, 90
- Antibodies 3, 24, 118, 122, 123, 159
- Antifoam 289 131
- Apple App Store 339, 348, 349
- Application programming interfaces (API) 216, 232
- Aromatic residues 184
- Assembly state 151
- Asymmetric unit 339, 340, 355
- Atomic charges 283, 289, 293, 301, 308
- Atomic coordinates 240, 279, 281, 282, 287
- Atomistic simulation 317
- Atom types 259, 262, 268, 269, 292, 294, 298, 301
- Augmented reality (AR) 347–355
- Autographa californica multiple nuclear polyhedrosis virus (AcMNPV) 68
- Automation 239
- Avogadro 280, 281

B

- BacMam system 96
- Bacmid 69–71, 75, 77, 78, 82, 89, 90, 97, 101
- Baculovirus expression vector system (BEVS) 47, 67, 68, 75, 96
- Barcodes 4, 5, 7, 9–12, 217, 224
- Barcoding 218
- Benchmark 265, 273
- Benzonase 48, 50, 51, 62, 72, 74
- Berendsen barostat 329
- β -lactam 14
- β -lactamase
- BlaC 290
 - AmpC 14–16, 18
- β -sheet 176, 177
- Binary checkpoint file (CPT) 270
- Binding affinity 221
- Bioconjugation 14
- Biological assembly 339
- Biological safety cabinet (BSC) 71, 77, 78, 80, 82, 83, 90
- Biotin 99, 105
- BlaC β -lactamase 290
- Bond angles 240, 289, 297
- Bonded model 257
- Bond length 240, 289, 295, 296
- Bond types 259, 291–293, 296, 307
- Bovine serum albumin (BSA) 31, 35, 36, 154, 155

C

- Cabbage looper, *Trichoplusia ni* 68
Camera 349, 350, 352, 355
Carbenicillin 31, 36, 41, 42
Cartoon or ribbon style 340
CAS number 216
CATH protein fold classification 182
CCP4 212
Cell disruptor 51, 56, 63, 74
Cell-free 129, 130, 132–134, 136–144, 146, 147
Center for Structural Genomics of Infectious Diseases
(CSGID) 212, 227
Chain length 183, 316
Chaotropic agents 128, 139
CHARMM22 force field 240, 241
CHARMM27 force field 290, 296
ChemDraw, Chem3D 280
Chiral center 291
CHO cells 118
Chromatograms 105
Chromatography
 fluorescence size exclusion chromatography
 (FSEC) 96, 99, 100, 107, 112
 gel filtration 21
 immobilized metal affinity chromatography
 (IMAC) 46, 57, 59, 63, 64, 88
 ion exchange chromatography 46, 51, 59, 74, 88
 size exclusion chromatography (SEC) 46, 51,
 57, 59, 64, 74, 75, 88
Chromatography column
 DEAE-cellulose (DE52) 63
 Econo- 57
 HiLoad Superdex 51, 75
 HiTrap Q, SP 51, 74
 Zenix-C SEC 300 99, 100
 Zenix-C SEC 300 guard 100
Circular dichroism (CD) 175–188
Clustering of sequences 240
Coarse-grained (CG) simulation 316
Column volume (CV) 57, 59, 64, 136, 152
Combinatorial genetics 3
Combinatorial mutations 7
Comma-separated value (CSV) files 215
Comparative modelling 279
Competent cells 4, 5, 7, 9, 27, 40, 41
Conformations 128, 151, 154, 159–171,
 175, 186, 250, 252, 266, 280, 315, 316, 326,
 330, 332
Conjugate gradient 240
Conservation 249
Contaminants 64, 65, 166
Continuous exchange cell-free (CECF) reaction 136
Continuous labeling 160–163
Coordinate bonds 261
Coordination 261, 282
Copolymers 322
Copper Cu^{2+} Cu^+ 258–261, 266
Critical micelle concentration (CMC) 162, 163
Cross-linking 14, 240
Cryo-electron microscopy (cryo-EM) 24, 46,
 63, 64, 151, 152, 154, 211
Crystallization screen 227
Crystallographic Information file (CIF) 291
Crystallographic Object-Oriented Toolkit
 (Coot) 279–282, 286, 287, 290,
 291, 298, 299, 307, 308
Crystallographic symmetry 339
Cut-off 65, 134, 136, 139, 146, 261, 287, 332
Cysteines 14, 19, 20, 137, 141, 290, 320
- D**
- Data acquisition 212, 214, 215
Databases 192–195, 198, 200, 202–204,
 212–215, 223–229, 231, 232, 234, 242, 248, 249
Data filters 191
Data identifier 191
Data management 210–234
Data mining analyses 225
Data record 193
DDM (n-Dodecyl β -D-maltoside) 46, 48, 50,
 98, 99, 104, 105, 162, 163
de novo modelling 239
Deoxynucleotide (dNTP) 24, 26, 27, 33, 40, 69
Desalting 21, 166
Deterministic 248
Dihedral angle
 improper 298, 299
 proper 298, 299
Dimers 59, 246
Dioxygen 258, 259, 261
Dioxygen binding metal 257–273
Discrete Optimized Protein Energy (DOPE)
 normalized 251
Disulfide bonds 67, 170, 272
Dithiothreitol (DTT) 21, 31, 35, 36, 40,
 48, 50, 72, 73, 98, 99, 131–134, 137, 138, 143
DMEM-F12 medium 101
DNA Ladders 26, 33, 38, 70, 79
DNA polymerase
 Herculase II Fusion 26
 MyTaq™ Red 26, 41, 69
 Q5 High-Fidelity 26, 39
 T4 24, 27, 35, 36, 40
DNA quantification 5
Domain boundaries 24, 31, 117
Domains 23, 24, 31, 117, 118, 140,
 147, 182, 183, 240, 278, 280, 340, 341

DSSP 183, 320
 Dynamic programming 242, 243, 250

E

E. coli S30 cellular extracts 127
E. coli strain BL21(DE3) 131
E. coli strain BL21(DE3) 35, 45, 47, 145
E. coli strain BL21(DE3)-R3-pRARE2 47, 52
E. coli strain DH10Bac 69, 75, 76
E. coli strain DH10EmBacY 75
E. coli strain Mach1 40
E. coli strain Rosetta pRARE 145
E. coli strain Rosetta2 47
 Elastic network 320, 332
 Electron density maps 280
 Electronic laboratory notebook (ELN) 229
 Elution volume 64
 End-group 322, 324
 Energy minimisation 305
 Energy profiles 244–246
 Enrichment ratio 10, 12
 Ensembles 250, 285, 305, 341, 342, 353, 355
 Enzyme Function Initiative (EFI) 212
 Enzyme kinetic assays 220
 Epistasis 3, 4, 11, 12
 E-value 243
 Expi293F cells 97, 101–104, 107
 Expi293F GnTI- cells 97–98, 101–104
 Expression screening 68, 96, 117–124
 Extended conformations 151
 Extinction coefficient difference 181

F

Fall army worm, *Spodoptera frugiperda* 68
 Fetal bovine serum (FBS) 70, 82, 83, 86, 92
 File format (type)
 chk 262, 263
 CIF 291
 cpt 265, 266, 306, 310
 CSV 215, 221
 FF (vermouth force field) 51, 59, 74
 frmod 260, 261, 263, 268, 269, 272
 gro 265, 266, 280, 282, 283, 285, 286, 293, 300, 303–308, 310, 311, 318, 319, 327–330
 inpcrd 264, 265
 itp 282, 284, 285, 287, 293–298, 300–302, 306, 307, 310, 318–327, 332
 mdp 265, 266, 285, 287, 303–306, 310, 311, 319, 328–330
 MOL2 291, 293, 307
 PIR 244, 245, 248
 prmtop 264, 265

top 265, 266, 285, 293–295, 300, 307, 308, 311, 312, 318–320, 327, 329, 330
 tpr 265, 266, 285, 305–307, 310, 311, 317, 328–330
 xtc 266
 Fingerprint files 261, 271
 Fluorescein 14, 20, 278, 281, 282, 287, 289
 Fluorescein-5-maleimide 14, 19–21
 Fluorescein-labeled 14, 17
 Fluorescence activated cell sorting (FACS) 9
 Fluorescence-based thermal shift assay (FBTSA) 221
 Fluorescence intensity 14, 135
 Fluorescence size exclusion chromatography (FSEC)
 screening 96, 99, 105
 Fluorescent label 289–309
 Fluorescent probe 14, 289, 290
 Fluorophores 14, 17, 19, 289, 298, 299
 Fold assignment 240, 250
 Folding-unfolding 151
 Fold prediction 175–187
 Force constant calculations 262, 271
 Force field
 Amber, ff14SB 264
 CHARMM22 240, 241
 CHARMM27 290, 293, 296
 general Amber force field (GAFF) 259, 268, 269
 Martini 316, 317
 OPLS, OPLS-AA/L 285, 287
 Formulatrix 218, 219
 FoXSDock 246
 Free modeling *ab initio* 239
 FreeStyle™ 293 expression medium 97
 Frozen stocks 79, 101, 102

G

GA341 score 245
 Gaussian03 270
 Gaussian09 270
 Gaussian16 261–263, 270
 Gel filtration 21, 64, 151, 152, 154
 Gene editing 3
 Gene ID 200, 203, 205
 Gene of interest (GOI) 24, 39
 General Amber force field (GAFF) 261, 268
 Gentamycin 70, 75, 77
 GitHub 12, 331, 351
 Glucose-thiamine solution 131, 134
 Glycosyltransferase 278
 GNOM 154
 Go approach 320, 332
 Google Play 339, 348, 349
 GROMACS 258, 265–267, 273, 277–287, 290, 291, 293, 295, 297, 298, 300, 301, 303–305, 307, 308, 318, 320, 328, 331, 332

H

H++ server 320
Haemocytometer 72, 118
Handheld 337, 353
Heat-shock 36, 75
HEK293 cells 123
HEK293S GnTI- (ATCC CRL-3022) 97
HEK293T cells 9
HEME group 266
Herculase II Fusion DNA Polymerase 26, 33
Heterogeneous proteins 151
High Five cells (BTI-Tn-5B1-4) 68, 70, 79
High-sensitivity DNA chip 5, 10
High-throughput 4, 192, 210, 212, 214, 220
His-tag 46, 65, 137, 143, 231
Histidine (HIS) residue (HID, HIP, HIE) 46, 137, 141, 258, 320
Histidine-tagged 56
HKL-2000/3000 212
Homogenization 46
Homology-derived restraints 240, 241
Hydrogen atoms 258, 259, 268, 271, 280, 282, 287, 291, 295, 301, 307
Hydrogen/deuterium exchange 164

I

Illumina library quantification 5, 10
Imaginary frequencies 262, 271
IMEx consortium 192, 193, 202
Imidazole 48, 50, 51, 57, 66, 72–74, 259
Immobilized metal affinity chromatography (IMAC) 46, 50, 57, 59, 63, 64, 73, 86, 88
Immobilized protease 164, 166, 170
IMolView 338–346
Impurity 65
In-browser web app 347
Inhibitors 13, 14, 24, 62, 74, 83, 140, 141, 278, 289
Insect cell line
 High Five cells (BTI-Tn-5B1-4) 68, 79
 Sf9 71, 76
 Sf21 68, 79
Insect-XPRESS serum-free and protein-free medium 73
Insertions 5, 7, 137, 250–252
Integral membrane proteins (IMPs) 24, 68, 95–113
Integrative Modeling Platform (IMP) 246
Integrative modeling 246
Interaction assay 117
Interaction network 191
Interaction records 194, 195, 197, 198, 200, 202, 203
Internet browser 348

Inverted light microscope 72
Ion channels 96, 151, 152
Ion exchange chromatography 46, 51, 59, 74, 88
Ionic strength 47, 63, 64, 138, 309
Ipads 337–339, 344, 347, 349, 354
Iphones 339, 349
IPTG 48, 53, 56, 62, 70, 75, 131, 134
iRefIndex 194–199
iRefWeb 191–205
Iron Fe³⁺ 267
Isotopes
 ¹⁵N 128–130, 137, 139, 140, 142–144
 ¹³C 128, 129, 137, 139–142
Isothermal titration calorimetry (ITC) 215, 221
Isotopic labeling 128, 142
Itp files 282, 284, 287, 293–296, 305, 318, 320, 322–327, 332

J

JLigand 290, 298

K

Kanamycin 31, 36, 41, 47, 52, 53, 55, 70, 75, 77

L

LabDB system 221
Laboratory information management systems (LIMS) 211–215, 223, 224, 226, 228, 229, 231–233
Lactate dehydrogenase (LDH) 242, 249
Large models 261–263, 267, 271
LB-agar 31, 36, 39
LEaP modeling 263
Least squares superposition 246, 247
Ligand parameterization 282–285
Ligand parameters 278
Light scattering 65, 180, 184, 187, 202
Link file 322, 324, 326
Linkers 4, 14, 278, 281, 282, 317, 325, 326, 328, 332
Literature curation 191
Local minimum 271
Long-read sequencing 4
Luria-Bertani (LB) medium 131
Lysozymes 48, 50, 132, 134, 155, 265, 273, 300, 304, 316, 318, 320, 324–326, 328, 330, 331

M

M9/²H₂O medium 129
Macromolecular structures 211, 347, 352
Macromolecule prep 220
Macropreps 214, 218, 221, 227

Malate dehydrogenases	242, 243
Maleimide	14, 21
Mammalian cells	5, 9, 10, 67, 68, 95, 96, 101, 103, 122
Mammalian cell line	
CHO	118
Expi293F™	97, 101–104, 107
HEK293	97, 101–104, 106, 107, 112, 118–124
HEK293S GnTI- (ATCC CRL-3022)	97
HEK293T	9
Martini 3	315–333
Martinize	318, 320, 326, 331
Martinizing	318
Martini parameters	317, 318, 321
Mass spectrometry (MS)	
in-line electrospray ionization time-of-flight analyzer	60, 88
Medium	
DMEM-F12	101
FreeStyle™ 293 expression	97, 102
Insect-XPRESS serum-free and protein-free	73
LB	42, 52, 75, 77, 131, 134
M9/ ² H ₂ O	129
Sf-900™ II SFM	73, 80, 82
SOC	52
TB	53, 55, 56
Melting curves	221
Membrane proteins	113, 128, 162, 163, 169, 177
Metadata	213, 215, 234
Metal binding proteins	151
Metal ions	63, 88, 257, 258, 261, 264, 266, 268–273
Metalloproteins	154, 258
Methionine (MET)	137, 141, 258
Methoxyethane	325, 326
Midwest Center for Structural Genomics (MCSG)	212, 227
Minimization	265, 273, 305, 329
Missing residues	290, 308, 331
Mobile devices	223, 337–355
Model assessment	
DOPE	244–246, 251
DOPE, normalized	245, 251
GA341 score	245
ModEval	245
Model protein	131, 135, 140
Modeller	239–252, 279, 280, 286
Modeling	
<i>ab initio</i>	240
comparative	239–242, 250
<i>de novo</i>	240
free	239
proteins	279
ModEval	245
Modularization	7
Moenomycin A	278, 281
Molecular docking	246
Molecular dynamics (MD)	258, 265–267, 273, 277–290, 299, 300, 303–305, 308, 316, 348, 349, 353
Molecular dynamics (MD) simulation	258, 277, 300, 316
Molecular graphics	280, 281, 337–339, 344, 347, 348, 353
Molecular interaction (MI)	192–194, 202, 203, 205
Molecular models	344, 347
Molecular surface	345
Molecular weights	14, 20, 21, 46, 57, 59, 65, 88, 128, 136, 161, 184, 316
Monomers	59, 65, 320, 322–324, 332, 338–340
MS/MS	57, 88, 167
MultiFit method	246
Multiple alignment	240
Multiplicity	262, 270–272, 298, 300
Multiplicity of infection (MOI)	9, 85, 86, 92, 93, 112
Mutagenesis	3–12
Mutants	3, 4, 10, 11, 19, 160, 168, 213, 251
MyTaq™ Red DNA polymerase	26, 69
N	
National Institutes of Health (NIH)	228, 234, 252
Native overlap	246
Natively disordered	175
NEBNext Library Quant Kit for Illumina	5, 10
Negative results	46
Neutralize, ions	285
New York Structural Genomics Research Consortium (NYSGRC)	212, 227
Next-generation sequencing (NGS)	5
Ni-IDA metal chelate resin	50, 73
96-well	24, 26, 27, 31, 33, 36, 37, 39, 46, 48–50, 52–55, 57, 61, 68, 70, 71, 73, 75, 77–79, 82, 83, 85, 89, 99, 105, 118, 120, 121, 123, 221, 223
Ni-NTA-agarose	50, 73
Non-bonded inter-atomic distance	240
NPT equilibration	265, 273
Nuclear magnetic resonance (NMR) spectroscopy	127, 239
NuPAGE	48, 50, 72, 73, 98, 99
NVT equilibration	265
NVT production (MD)	266

O

Objective function240, 241, 244, 251
Observation robot.....219
Oligomerisation339
Ollomol.....348–350, 353, 354
Open Babel.....280, 281, 290, 307
Open source168, 213, 246, 345
OPLS.....284, 291, 293
OPLS-AA/L force field285, 287
Organometallic.....258
Origin214, 218, 221
Overlap concentration326, 332

P

p10 and polyhedrin genes68
p53 tetramerization domain.....338
Pairwise interactions192, 203
PAP format244
Parameter (MD)
 atom type260, 263, 268, 269, 292, 294, 298, 301
 atomic charge283, 289, 293, 301, 308
 bond angle289
 bond length240, 289, 295
 bond type.....260, 291–293, 296, 306–307
 dihedral240, 273, 284, 287, 293, 295, 298, 299, 300,
 307, 312, 317, 320, 322
 improper293–295, 298, 300, 307
 non-bonded inter-atomic distance240
 proper ...60, 96, 161, 180, 181, 217, 277, 278, 284,
 293, 295, 298–300, 309
Parameterization257–273, 277–287, 291–293
Parameterize258, 262
Parrinello–Rahman pressure coupling330
Polymerase chain reaction (PCR)
 primers4, 5, 10
 colony24, 26, 36–38
 RT-221
 touchdown33, 39
Partial charges .262, 264, 268, 269, 271, 272, 282, 284
PDB IDs ...17, 230, 258, 278, 280, 281, 290, 338–341,
 345, 348–354
PDB: HETATM, LINK.....285
PEGylation315, 320
Penicillin70, 90
Peptide bonds.....175, 181, 295
Peptide sequence194
Peptidoglycan278
Perdeuterated amino acids.....129, 139, 141
Perdeuterated proteins.....127–147
Periodic boundary condition (PBC).....307
Phenix212
Phosphate buffer152, 179, 309
Phosphate-buffered saline (PBS)72, 87, 97, 98, 103, 152,
 153, 179
Phosphates...21, 63, 133, 137, 139, 161, 287, 292, 309
PIR format, sequence248
PIVEX-GFP plasmid131
PIVEX vectors137
Pka320, 332
Plasmid library.....9
Pluronic F68 reagent119–121
Polyethylene glycol (PEG)98, 103, 105, 315–317,
 320–326, 328, 332, 333
Polyethyleneimine (PEI)51, 56, 63, 112, 119–121, 123,
 132, 136, 143
Polymers63, 316, 317, 320, 322–324, 326, 328–333
Polyply318, 320, 322, 323, 326, 331–333
Positional restraints300, 302, 305, 328, 329
PostgreSQL database195, 213, 224
Post-translational modification (PTM)46, 60, 67, 88, 96,
 117
Potential energy305
Primer dimer39
PRIMUS.....154
Probability density function240
Probability distributions317, 325
Production run, MD.....266, 273, 330
Protease inhibitor cocktail48, 50, 51, 72, 74, 98
Protein aggregates.....65, 177, 184
Protein chains251, 341
Protein conformation159, 161, 167
Protein Data Bank (PDB)16, 17, 20, 31, 182, 183, 212,
 213, 230, 242–245, 248, 251, 258–261,
 264–266, 268, 269, 271, 272, 285, 290, 291,
 293, 294, 299–301, 307, 308, 318, 320, 326,
 331, 338–342, 345–349, 351–355
Protein dynamics320
Protein engineering3–12
Protein fold175–187, 340
Protein functions.....3, 159
Protein modelling280, 320, 322
Protein production 24, 45–96, 137, 141, 211–214, 218,
 231
Protein selection.....3
Protein sequences.....3, 21, 23, 39, 195, 239, 240, 341
Protein structures....20, 23, 24, 67, 177, 180, 183, 184,
 239–252, 280, 290, 295, 298, 308, 326, 332,
 338, 348, 353
Protein synthesis
 in vitro130, 139, 142–144
Protein variant characterization3
Protein-drug interaction159
Protein-ligand complex285
Protein-ligand interaction159
Protein-protein interactions (PPI)....191–194, 197, 198,
 200–203, 205

- Proteomics 118, 211
Proteomics Standards Initiative (PSI) 192
Protonation 128–130, 140–143, 145,
259, 266, 268, 280, 286, 287, 320, 329, 332
ProtParam 21, 186
PSI-MI standards 191
PubChem 215
PubMed IDs 195, 198, 202, 203, 205
PyMOL 17, 20, 280, 282, 318,
338–340, 344, 346, 353–355
Python 12, 232, 241, 242, 244, 245,
248, 285, 331
Python 3 318, 331
Python scripts 241, 248, 265, 278, 285
- Q**
- Q5 High-Fidelity DNA Polymerase 26
QR code 348, 354
Quantum calculations 261–263, 268–272
Quenching 161, 162, 164, 165, 169–171
- R**
- Radioactive isotope 151
Radius of gyration (R_g) 154, 155
Rainbow coloring 343
Real time 5, 10, 13, 14, 193, 225, 344, 347, 348
Real-world environment 347
Reference structure 320
Relaxation 128, 129, 305, 306
Rendering 339, 340, 348, 350, 353
Reproducibility 210–234
Research Collaboratory for Structural Bioinformatics
Protein Data Bank (RCSB PDB) 278, 280,
340, 345, 352
Reservoirs 15, 19, 27, 33, 49, 57, 71,
75, 91, 113, 218, 226
Resolution 64, 127, 128, 162, 164,
165, 182, 183, 226, 249, 251, 286, 337
Restrained electrostatic potential (RESP) charge
fitting 262
Restriction enzymes 5, 7, 24, 35
RNase 26, 98, 130, 131, 133, 134, 137
Root-mean-squared-deviation (RMSD) 246, 247,
251, 266
RT-PCR 221
- S**
- SacB gene 39
Salt concentrations 57, 63, 64, 88, 329
SCATTER 154, 155
Screening 3, 9, 24, 26, 36–38, 41,
42, 46, 47, 68, 78, 95–113, 118, 218–220, 289
SDS PAGE 20, 21, 50, 55, 57, 59,
61, 64, 73, 85, 99, 105, 112, 132, 136, 146, 154,
180, 184
Secondary structures 31, 175–188, 240,
243, 286, 315, 317, 320, 331–333, 341, 354
SELCON/CONTIN/CDSSTR methods 183
Seminaro method 262, 270
Sequence alignments 31, 286
Sequence identity 243, 245, 247, 249–251
Sequence mismatch 271
Sf-900™ II SFM medium 80–83
Sf9 and Sf21 cells 68, 79
Short-read sequencing 4
Simulated annealing 240
Simulation
atomistic 317
coarse-grained (CG) 320
molecular dynamics (MD) 277, 300, 316
Single spectrum analysis 181, 183
Site-directed mutagenesis 14, 19, 240, 242
Size exclusion chromatography (SEC) 57–59,
63–66, 99, 105
Small angle neutron scattering (SANS) 316
Small angle X-ray Scattering (SAXS) 151–156, 246
Small model 262, 267, 271, 272
SMILES representation 216
SOC media 31, 36, 39, 41, 48, 57
Software development kit (SDK) 216
Solvents 14, 99, 128, 129, 139,
142–144, 160, 162, 164–167, 171, 243, 249,
264, 302, 303, 305, 309
Sonication 46, 62, 63, 83, 93, 104, 136
Source databases 193–200, 202, 203, 205
Spatial restraints 240, 241, 251
Spreadsheets 210, 211, 215, 218, 229
Solid phase reversible immobilisation (SPRI)
paramagnetic beads 5
Steepest descents 304
Stereochemical restraints 240, 241
Stereoisomers 291
Stop codon 31
Strep-Tactin®XT resin 99, 100, 105, 106, 112
Streptomycin 71, 90, 133
Structural biology 130, 136, 147, 209–234
Structural Genomics Consortium (SGC) 23, 24,
39, 42, 45, 66, 93, 113
Structure-function research 211
Subunits 183, 203, 204, 246, 252
Sucrose 31, 36, 39–41, 48, 70
Suspension culture 67, 68, 79, 80, 91, 96, 101, 102
Swiss-Pdb Viewer 17, 20
SYBR Green PCR Mix 5
SYBYL MOL2 format 291

T

T4 DNA polymerase 24, 27, 35, 36, 40
T7 RNA polymerase 130–132, 134, 136, 139
TAE buffer 26, 69, 70
Target sequence 31, 146, 240–242,
245, 247, 249–251
TE Buffer 26, 34, 70, 78
Templates 26, 33, 34, 39–41, 130,
240, 242–247, 249–252, 268, 279, 280, 286
Template structures 240, 241, 245, 247,
249, 251, 280
Tertiary structure 320, 332
Tetramer 338, 339
TEV protease 46, 65
Text editor 291, 299, 309
Thermocycler 27, 33, 35–37, 41, 50, 71, 73, 78
Thiol 14, 20, 21, 320, 324
Three-dimensional (3D) structure 239
Titratable amino acids 320, 332
Topologies 182, 183, 259, 264–267,
272, 282, 284, 291, 293, 294, 297, 298,
300–308, 326, 329
Touchdown PCR 33, 39
Tracker pattern 349, 350
Trajectory 14, 266, 305, 307
Transaminase inhibitors 129
Transfection
transient 96
Transformation 4, 9, 24, 36, 37, 40, 41,
47–49, 52, 57, 61
Transmembrane domains 96
Transporters 96, 151–155
Transposable element 76
Transposition 69–71, 75, 97, 101
Transposition, bacmid production 68, 96, 101
Trapping 122, 162, 164, 166
Trichomonas vaginalis (TvLDH) 242
Tris(2-carboxyethyl)phosphine (TCEP) ... 21, 26, 48, 69,
72, 98, 133, 144, 161, 186
Trypan blue 70, 80, 90, 102, 111, 120

2D-(¹⁵N, ¹H) solid-state NMR 130, 143
2D-(¹H, ¹⁵N)-CRINEPT-HMQC-TROSY 142, 144
Type IIS restriction enzymes 4, 5, 7, 11

U

UniProt 195, 197, 200, 203, 205
Unnatural residue 260, 268, 269, 272
UV-spectrophotometer, Nanodrop 78

V

Vector

BEVS 37, 47, 67, 68, 75, 96
pHTBV1.1 (BacMam) 24, 27, 40, 43,
47, 94, 95, 113
pIVEX 133
pIVEX-GFP 131
Vermouth force-field format 322
Viability 42, 85, 90, 93, 102, 111, 123
Visual molecular dynamics (VMD) 259, 264,
272, 280, 286, 290, 305, 318, 330
Visualization 16, 17, 20, 272, 318, 330, 337–355

W

Water molecules 268, 303, 304, 328, 329
Web interfaces 194, 212, 221, 223
Web-based 212, 217, 223, 224, 229
Web-based Augmented Reality for Chemistry and
Structural (WARCSB) portal 348, 354

X

X-ray crystallography 24, 46, 239
X-ray diffraction 214

Z

Zinc Zn²⁺ 63
Z-matrix 270
Z-medium 131, 134, 145