METHODS IN MOLECULAR BIOLOGY™ 364

Macromolecular Crystallography Protocols

Volume 2
Structure Determination

Edited by

Sylvie Doublié

HUMANA PRESS

Macromolecular Crystallography Protocols

Volume 2, Structure Determination

METHODS IN MOLECULAR BIOLOGY™

John M. Walker, SERIES EDITOR

383. Cancer Genomics and Proteomics: *Methods and Protocols*, edited by *Paul B. Fisher, 2007*

382. Microarrays, Second Edition: *Volume 2, Applications and Data Analysis*, edited by *Jang B. Rampal, 2007*

381. Microarrays, Second Edition: *Volume 1, Synthesis Methods*, edited by *Jang B. Rampal, 2007*

380. Immunological Tolerance: *Methods and Protocols*, edited by *Paul J. Fairchild, 2007*

379. Glycovirology Protocols, edited by *Richard J. Sugrue, 2007*

378. Monoclonal Antibodies: *Methods and Protocols*, edited by *Maher Albitar, 2007*

377. Microarray Data Analysis: *Methods and Applications*, edited by *Michael J. Korenberg, 2007*

376. Linkage Disequilibrium and Association Mapping: *Analysis and Application*, edited by *Andrew R. Collins, 2007*

375. In Vitro Transcription and Translation Protocols: *Second Edition*, edited by *Guido Grandi, 2007*

374. Quantum Dots: *Methods and Protocols*, edited by *Charles Z. Hotz and Marcel Bruchez, 2007*

373. Pyrosequencing® Protocols, edited by *Sharon Marsh, 2007*

372. Mitochondrial Genomics and Proteomics Protocols, edited by *Dario Leister and Johannes Herrmann, 2007*

371. Biological Aging: *Methods and Protocols*, edited by *Trygve O. Tollefsbol, 2007*

370. Adhesion Protein Protocols, *Second Edition*, edited by *Amanda S. Coutts, 2007*

369. Electron Microscopy*: Methods and Protocols, Second Edition*, edited by *John Kuo, 2007*

368. Cryopreservation and Freeze-Drying Protocols, *Second Edition*, edited by *John G. Day and Glyn Stacey, 2007*

367. Mass Spectrometry Data Analysis in Proteomics, edited by *Rune Matthiesen, 2007*

366. Cardiac Gene Expression: *Methods and Protocols*, edited by *Jun Zhang and Gregg Rokosh, 2007*

365. Protein Phosphatase Protocols: edited by *Greg Moorhead, 2007*

364. Macromolecular Crystallography Protocols: *Volume 2, Structure Determination*, edited by *Sylvie Doublié, 2007*

363. Macromolecular Crystallography Protocols: *Volume 1, Preparation and Crystallization of Macromolecules*, edited by *Sylvie Doublié, 2007*

362. Circadian Rhythms: *Methods and Protocols*, edited by *Ezio Rosato, 2007*

361. Target Discovery and Validation Reviews and Protocols: *Emerging Molecular Targets and Treatment Options, Volume 2*, edited by *Mouldy Sioud, 2007*

360. Target Discovery and Validation Reviews and Protocols: *Emerging Strategies for Targets and Biomarker Discovery, Volume 1*, edited by *Mouldy Sioud, 2007*

359. Quantitative Proteomics by Mass Spectrometry, edited by *Salvatore Sechi, 2007*

358. Metabolomics: *Methods and Protocols*, edited by *Wolfram Weckwerth, 2007*

357. Cardiovascular Proteomics: *Methods and Protocols*, edited by *Fernando Vivanco, 2006*

356. High Content Screening: *A Powerful Approach to Systems Cell Biology and Drug Discovery*, edited by *D. Lansing Taylor, Jeffrey Haskins, and Ken Guiliano, 2007*

355. Plant Proteomics: *Methods and Protocols*, edited by *Hervé Thiellement, Michel Zivy, Catherine Damerval, and Valerie Mechin, 2006*

354. Plant–Pathogen Interactions: *Methods and Protocols*, edited by *Pamela C. Ronald, 2006*

353. DNA Analysis by Nonradioactive Probes: *Methods and Protocols*, edited by *Elena Hilario and John. F. MacKay, 2006*

352 Protein Engineering Protocols, edited by *Kristian Müller and Katja Arndt, 2006*

351 *C. elegans:* *Methods and Applications*, edited by *Kevin Strange, 2006*

350 Protein Folding Protocols, edited by *Yawen Bai and Ruth Nussinov 2007*

349 YAC Protocols, *Second Edition*, edited by *Alasdair MacKenzie, 2006*

348 Nuclear Transfer Protocols: *Cell Reprogramming and Transgenesis*, edited by *Paul J. Verma and Alan Trounson, 2006*

347 Glycobiology Protocols, edited by *Inka Brockhausen-Schutzbach, 2006*

346 *Dictyostelium discoideum* Protocols, edited by *Ludwig Eichinger and Francisco Rivero, 2006*

345 Diagnostic Bacteriology Protocols, *Second Edition*, edited by *Louise O'Connor, 2006*

344 *Agrobacterium* Protocols, *Second Edition: Volume 2*, edited by *Kan Wang, 2006*

343 *Agrobacterium* Protocols, *Second Edition: Volume 1*, edited by *Kan Wang, 2006*

342 MicroRNA Protocols, edited by *Shao-Yao Ying, 2006*

341 Cell–Cell Interactions: *Methods and Protocols*, edited by *Sean P. Colgan, 2006*

340 Protein Design: *Methods and Applications*, edited by *Raphael Guerois and Manuela López de la Paz, 2006*

339 Microchip Capillary Electrophoresis: *Methods and Protocols*, edited by *Charles S. Henry, 2006*

338 Gene Mapping, Discovery, and Expression: *Methods and Protocols*, edited by *M. Bina, 2006*

METHODS IN MOLECULAR BIOLOGY™

Macromolecular Crystallography Protocols

Volume 2: Structure Determination

Edited by

Sylvie Doublié

Department of Microbiology and Molecular Genetics, University of Vermont, Burlingon, VT

HUMANA PRESS TOTOWA, NEW JERSEY

© 2007 Humana Press Inc.
999 Riverview Drive, Suite 208
Totowa, New Jersey 07512

www.humanapress.com

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise without written permission from the Publisher. Methods in Molecular Biology™ is a trademark of The Humana Press Inc.

All papers, comments, opinions, conclusions, or recommendations are those of the author(s), and do not necessarily reflect the views of the publisher.

This publication is printed on acid-free paper. ♾
ANSI Z39.48-1984 (American Standards Institute)

Permanence of Paper for Printed Library Materials.
Cover illustration: Figure 4 from Chapter 7, Volume 1, "Screening and Optimization Methods for Nonautomated Crystallization Laboratories," by Terese Bergfors.

Cover design by Patricia F. Cleary

For additional copies, pricing for bulk purchases, and/or information about other Humana titles, contact Humana at the above address or at any of the following numbers: Tel.: 973-256-1699; Fax: 973-256-8341; E-mail: orders@humanapr.com; or visit our Website: www.humanapress.com

Photocopy Authorization Policy:
Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Humana Press Inc., provided that the base fee of US $30.00 per copy is paid directly to the Copyright Clearance Center at 222 Rosewood Drive, Danvers, MA 01923. For those organizations that have been granted a photocopy license from the CCC, a separate system of payment has been arranged and is acceptable to Humana Press Inc. The fee code for users of the Transactional Reporting Service is: [978-1-58829-902-4 • 1-58829-902-3/07 $30.00].

Printed in the United States of America. 10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging in Publication Data

eISBN: 1-59745-266-1 13ISBN 978-1-58829-902-4

ISSN: 1064-3745

Library of Congress Cataloging in Publication Data

Macromolecular crystallography protocols / edited by Sylvie Doublié.
p. cm. -- (Methods in molecular biology, ISSN 1064-3745 ; v. 363-364)
Includes bibliographical references and indexes.
Contents: v. 1. Preparation and crystallization of macromolecules -- v. 2. Structure determination.
ISBN-13: 978-1-58829-292-6 -- ISBN-13: 978-1-58829-902-4
ISBN-10: 1-58829-292-4 (v. 1 : acid-free paper) -- ISBN-10: 1-58829-902-3 (v. 2 : acid-free paper)
1. Macromolecules--Structure. 2. Crystallography. 3. Proteins--Structure. 4.
Membrane proteins. 5. X-Ray crystallography--Technique. I. Doublié, Sylvie. II. Series:
Methods in molecular biology (Clifton, N.J.) ; v. 363-364.
QP551.M332 2007
572'.633--dc22

2006043538

Preface

In the decade since the publication of the first edition of *Crystallographic Methods and Protocols*, the field has seen several major developments that have both accelerated the pace of structure determination and made crystallography accessible to a broader range of investigators. As evidence of this growth, this new work, *Macromolecular Crystallography Protocols*, encompasses two volumes: volume 1, *Preparation and Crystallization of Macromolecules*, and volume 2, *Structure Determination*.

Fanning the fire are the large number of synchrotron beamlines dedicated to macromolecular crystallography and the availability of inexpensive desktop supercomputers. Expression systems for proteins and nucleic acids have greatly improved as well. Several improvements come to mind: ligation-independent cloning, the development of N-terminally fused expression tags that help protein solubility, and the use of eukaryotic expression systems. In addition, structural genomics has increasingly changed the way we go about solving crystal structures, not only because of the shear increase in the number of deposited structures, but more importantly because of the new tools the structural genomics centers have developed and are making available to the community at large.

Following volume I, which is dedicated to the preparation and crystallization of macromolecules, this second volume, *Structure Determination*, covers both laboratory and computational methods for characterizing crystals and solving structures. The topic of crystal handling, characterization, and data collection is covered in Chapters 1–6. Most crystals are cryocooled in order to increase their lifetime in the X-ray beam. Garman and Owen share their practical experience in optimizing cryocooling techniques in Chapter 1. In Chapter 2, Chu outlines how reaction intermediates can be captured at ultra-low temperatures. Annealing techniques, which can dramatically improve crystal diffraction limits, are reviewed by Bunick and Hanson in Chapter 3. In Chapter 4, Jeruzalmi acquaints us with the first analysis of macromolecular crystals. Garman and Sweet go over the nitty-gritty of macromolecular crystal data collection in Chapter 5. In addition, in Chapter 6 Sawaya tells us everything we need to know about characterizing a crystal from an initial dataset.

Five chapters are dedicated to the subject of phasing. In Chapter 7, Toth presents the different ways to solve a structure by molecular replacement. Isomorphous replacement is covered in Chapter 8 by Dauter and Dauter, who describe the use of halide ions for phasing, and also in Chapter 9 by Rould, who extends the concept of isomorphous replacement to isomorphous difference

Fourier maps and their use. Location of heavy metal positions is the focus of Chapter 10 by Smith and collaborators, who describe the use of *Shake-and-Bake* with anomalous datasets, and Chapter 11 by Grosse-Kunstleve and Schneider, who present anomalous and isomorphous cases. Vonrhein and coworkers describe automated structure solution by autoSHARP in Chapter 12.

Structure refinement is detailed in Chapter 13 by Tronrud. Kleywegt focuses on quality control and validation in Chapter 14. Finally, Chapter 15 by Everse and Doublié surveys the available crystallographic software.

It is my sincere hope that students will find the two *Macromolecular Crystallography Protocols* volumes useful, as it was they that I had in mind when I put this book together. I essentially designed a book I wished I could have had available when I was a student. May these two volumes help all crystallographic apprentices obtain crystals and guide their steps along the all too often rugged path of structure determination.

I would like to express my sincere thanks to Anne MacLeod for her help during the final phase of manuscript handling. Finally, I am grateful to all the authors for carving out time to write their manuscripts, for being so cooperative, and for their patience throughout the different stages of the book production.

Sylvie Doublié

Contents

Contributors

ERIC BLANC • *European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom*
GÉRARD BRICOGNE • *Global Phasing Ltd., Sheraton House, Castle Park, Cambridge, United Kingdom*
GERARD J. BUNICK • *Department of Biochemistry, Cellular, and Molecular Biology and the Center of Excellence for Structural Biology, University of Tennessee, Knoxville, TN*
KELVIN CHU • *Physics Department, University of Vermont, Burlington, VT*
MIROSLAWA DAUTER • *Macromolecular Crystallography Laboratory, National Cancer Institute, Argonne National Laboratory, Argonne, IL*
ZBIGNIEW DAUTER • *Synchrotron Radiation Research Section, National Cancer Institute, Advanced Photon Source, Argonne, IL*
SYLVIE DOUBLIÉ • *Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, VT*
STEPHEN J. EVERSE • *Department of Biochemistry, University of Vermont, Burlington, VT*
ELSPETH GARMAN • *Department of Biochemistry, Laboratory of Molecular Biophysics, University of Oxford, Oxford, United Kingdom*
RALF W. GROSSE-KUNSTLEVE • *Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA*
B. LEIF HANSON • *College of Arts and Science Instrumentation Center, University of Toledo, Toledo, OH*
P. LYNNE HOWELL • *Structural Biology and Biochemistry, Hospital for Sick Children, Toronto, Canada and Biochemistry Department, University of Toronto, Toronto, Canada*
DAVID JERUZALMI • *Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA*
GERARD J. KLEYWEGT • *Department of Cell and Molecular Biology, Uppsala University, Biomedical Center, Uppsala, Sweden*
CHRISTOPHER T. LEMKE • *Structural Biology and Biochemistry, Hospital for Sick Children Toronto, Canada and Biochemistry Department, University of Toronto, Toronto, Canada*
ROBIN L. OWEN • *Department of Biochemistry, University of Oxford, Oxford, United Kingdom*
MARK A. ROULD • *Department of Molecular Physiology and Biophysics, University of Vermont, Burlington, VT*

PIETRO ROVERSI • *Department of Biochemistry, Laboratory of Molecular Biophysics, University of Oxford, Oxford, United Kingdom*
MICHAEL R. SAWAYA • *Department of Chemistry and Biochemistry, Molecular Biology Institute, University of California–Los Angeles, Los Angeles, CA*
THOMAS R. SCHNEIDER • *Italian Foundation for Cancer Research, Institute of Molecular Oncology, European Institute of Oncology, Milan, Italy*
G. DAVID SMITH • *Structural Biology and Biochemistry, Hospital for Sick Children, Toronto, Canada and Hauptman–Woodward Medical Research Institute, Buffalo, NY*
ROBERT M. SWEET • *Biology Department, Brookhaven National Laboratory, Upton, NY*
ERIC A. TOTH • *Department of Biochemistry and Molecular Biology, Marlene and Stewart Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, MD*
DALE E. TRONRUD • *Institute of Molecular Biology, Howard Hughes Medical Institute, University of Oregon, Eugene, OR*
CLEMENS VONRHEIN • *Global Phasing Ltd., Sheraton House, Castle Park, Cambridge, United Kingdom*

Contents of Volume 1

1

Cryocrystallography of Macromolecules

Practice and Optimization

Elspeth Garman and Robin L. Owen

Summary

Techniques for flash-cooling protein crystals to around 100K (–173°C) for data collection have developed enormously in the last decade, to the extent that cryocrystallography is now standard practice. The main advantage of these methods is the vastly reduced rate of radiation damage to protein crystals in the X-ray beam at cryogenic temperatures over room temperature, extending their lifetimes so that complete datasets can be collected from a single crystal. The practical application of the techniques has become somewhat anecdotal and rather fixed within individual laboratories.

This chapter gives step-by-step guidelines for flash-cooling crystals and some of the rationale for the recommended procedures. Optimization of the entire cryoprotocol can give substantial improvements to both the resolution and quality of the data, often resulting in more straightforward structure solution and subsequent model refinement. Attention to seemingly insignificant details can have a real impact on the usefulness of the final dataset, and are thus worth addressing.

Key Words: Cryocrystallography; cryoprotection; cryostat; cryogen; cryoprotectant; ice; mosaicity; radiation damage.

1. Introduction

Cryocrystallographic techniques are now an essential part of macromolecular X-ray data collection, and have had a major impact on the quality and number of structures being obtained by structural biologists. The basic methods are well covered in the literature ***(1–6)*** and should be consulted for more detailed information, especially for troubleshooting. This chapter will concentrate on the basic practical application of the techniques, which often become modified and non-optimal with time as they are handed down, largely anecdotally, to new researchers. The recommended procedures have a rationale based on finding the

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

cryoprotocol that will ultimately give the optimum diffraction data quality. Their application ensures the best chemical conditions for the crystal (e.g., cryoprotectant composition and concentration, minimum osmotic shock) and the least damaging fishing and flash-cooling procedure in order to retain the physical state of the crystal. These two factors together can significantly improve the quality of a dataset and thus make subsequent structure solution and refinement more straightforward.

The compelling advantage of collecting data at cryotemperatures (below 136K, the glass-transition temperature of pure water) is the vastly reduced mobility of the secondary radiation products induced by primary radicals, which are formed when the X-ray beam loses energy in the crystal. This increases the useful lifetime of the crystal in the beam and enables a complete dataset to be collected from a single crystal, or several datasets at different incident X-ray wavelengths if a multiwavelength anomalous dispersion structure determination is being attempted. This, in turn, reduces the experimental systematic errors, which arise when data from several crystals have to be merged, and thus improves the overall quality of the data.

It is well known that protein crystals have a limited lifetime in the X-ray beam even at cryotemperatures, with a calculated dose limit of 2×10^7 Gy (=J/kg) *(7)*, which is equivalent to approx a 5-min exposure on a third generator undulator beamline, 24 h on a wiggler beamline, and 2.5 mo on a modern in-house rotating anode generator equipped with confocal optics. Radiation damage has been found to occur at specific structural sites, with disulfide bonds breaking first *(8–10)*. Experiments are currently underway to find conditions that will minimize this damage at cryotemperatures (reviewed in **ref.** *11*).

An additional benefit of cryotechniques is that the prevailing mounting method of suspending the crystal by surface tension in a vitrified liquid film held in a small loop *(12)* of fiber (nylon, rayon, or mylar), is much gentler and easier than the room temperature glass or quartz thin-walled capillary mounting methods. Small, fragile crystals are much more conveniently mounted in a loop.

Lastly, crystals can be flash cooled into liquid cryogen while in peak condition for later data collection, or can be screened in-house for data collection at a synchrotron source.

2. Materials

The items listed next can be obtained as follows: Hampton Research (Aliso Viejo, CA) (2–5,7–11), Molecular Dimensions (Cambridgeshire, UK) (2–10), Rikagu (MSC, The Woodlands, TX) (1), Nalgene (Rochester, NY) (7–10) Oxford Cryosystems (Oxfordshire, UK) (1–3,5–12), and Taylor-Wharton (Husum, Germay) (12).

1. A reliable cryostat delivering a stream of gaseous nitrogen at around 100 ± 1.0K with an outer sheath of room temperature dry air (dew point ~ –60°C) or nitrogen.

2. A magnet that fits into the 3-mm diameter hole in the top of a goniometer head.
3. Top-hats or cryocaps.
4. Pins and rayon/nylon loops.
5. Handling tools: cryotongs, lid undoer, crystal wand, vial holder (*see also* **refs. *1–3***).
6. Removable arc.
7. Cryovials.
8. Cryocanes.
9. Cryosleeves.
10. Cryomarkers.
11. Small working Dewar, a 2-L Dewar, a storage Dewar.
12. Dry transport Dewar.
13. Acupunture needles (www.acumedic.com).
14. Gloves: nitrile or latex.
15. Thermally insulating gloves (KT1, KT2; www.marigoldindustrial.com).
16. Safety goggles.
17. Cryogen: liquid or gaseous nitrogen (*see* **Note 1**).

3. Methods

3.1. Cryoprotection

In order to prevent ordered crystalline ice forming during flash cooling, a crystal must be cooled so fast that any water in it is vitrified. For pure water this time is on the order of 10^{-5} s. The addition of cryoprotectant agents ("antifreeze") increases the time to 1–2 s for the cooling process to take place without the formation of crystalline ice (**Fig. 1**). There are two types of cryoprotectant agents commonly used: (1) cryoprotectant agents, such as glycerol, which penetrate into solvent channels increasing the time required for crystalline ice formation within the crystal. **Figure 2** shows that glycerol is the most commonly used cryoprotectant, although this by no means implies glycerol is always best; it is often just the most convenient and no subsequent optimization takes place because it yields adequate results ***(13)***. (2) Cryoprotectant agents, such as oil, which do not penetrate into the crystal but provide a barrier between the surface of the crystal and the air, are most useful for crystals with narrow solvent channels.

When deciding which cryoprotectant agent to try, the components of the mother liquor should first be considered. If a cryoprotectant agent is already present at low concentration, the concentration can often be increased without damaging the crystal or giving it too large an osmotic shock. **Table 1** gives an idea of where to start in selecting a cryoprotectant agent. In approximately two-thirds of cases the addition (by replacement of water in the mother liquor) of 15–25% (v/v) glycerol will give satisfactory, though not optimal, cryoprotection. A benign agent will neither attack the crystal surface (this ultimately results in the crystal dissolving), nor cause its surface to increasingly resemble crazed pottery.

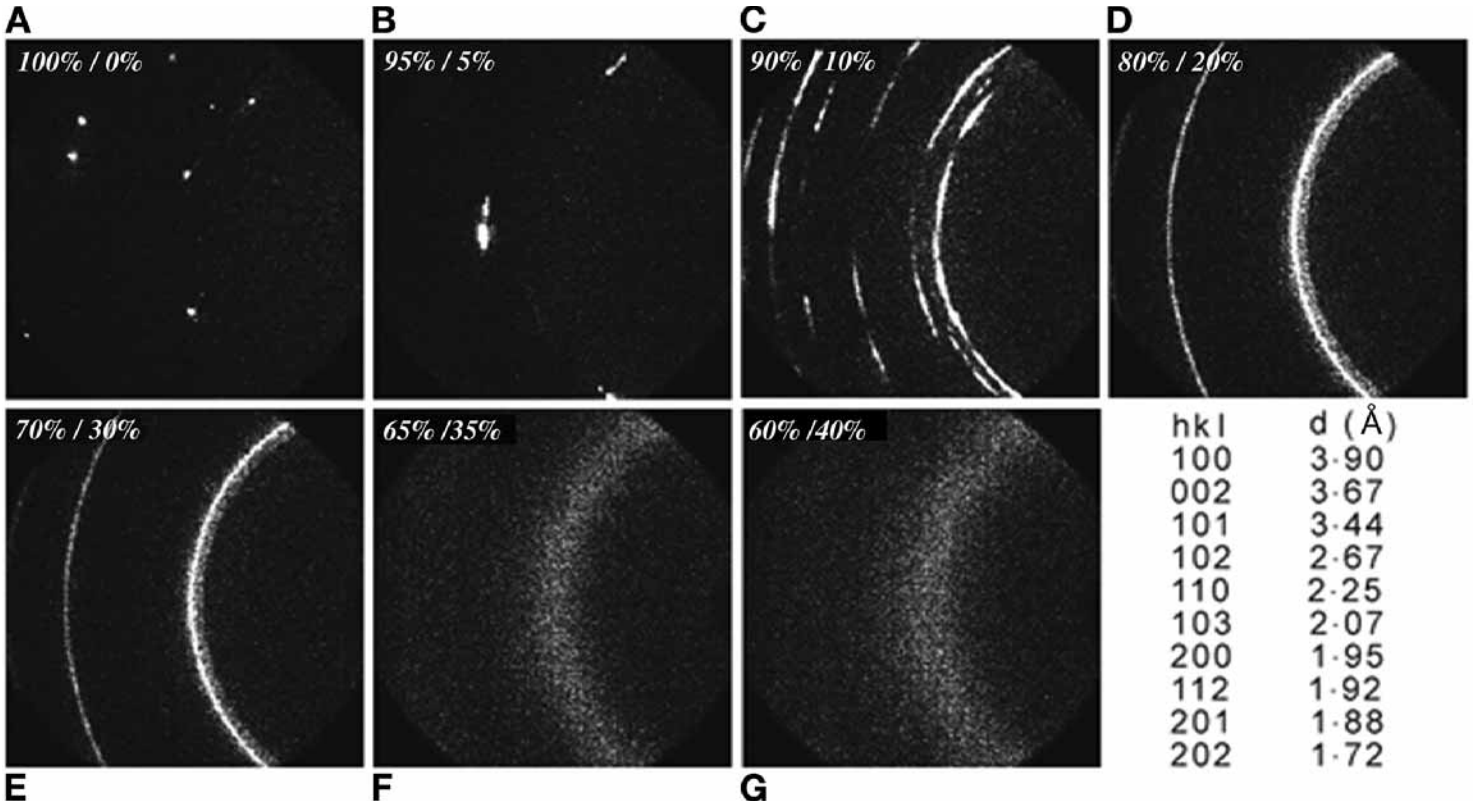


Fig. 1. Diffraction patterns from water/glycerol mixtures collected at 2θ = 10°C: **(A)** 100/0%, **(B)** 95/5%, **(C)** 90/10%, **(D)** 80/20%, **(E)** 70/30%, **(F)** 65/35%, and **(G)** 60/40%. The resolutions of the main diffraction rings seen from crystalline ice are also shown. (Reprinted from **ref.** *15* with permission from the IUCr.)

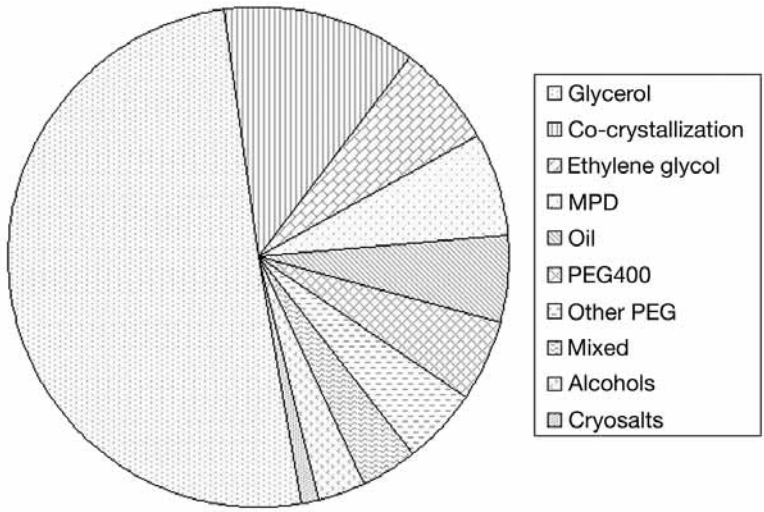


Fig. 2. Pie chart showing the most commonly used cryoprotectant agents. Co-crystallization refers to crystallization conditions that do not require the addition of a cryoprotective agent. Cryobuffers containing more than one cryoprotectant were grouped into the "mixed" category. Data were compiled from a survey of all papers published in *Acta Crystallographica D*, January 2000 to May 2003.

The putative cryosolution is first mounted in the loop and flash frozen for a diffraction test. A glassy-looking result is a necessary but not sufficient condition for adequate cryoprotectant agent concentration; it can be seen in **Fig. 1** that a water/glycerol mixture of 60/40% gives a diffuse scattering diffraction ring. However, this ring has a sharper low-resolution edge than high-resolution edge, and the addition of a further 5% of glycerol ensures equal slopes on both sides and gives better results. Also, the cryobuffer concentration is often slightly diluted by the addition of the crystal and its surrounding mother liquor; a safe strategy is to increase the initial concentration by 2–5%.

Once a cryoprotectant has been chosen, its concentration should be optimized as well as the cryocooling protocol. The flow diagram in **Fig. 3** outlines a possible strategy to follow when selecting and optimizing a cryoprotocol. In cases where difficulty is experienced in finding suitable cryoconditions, the effects of temperature and osmotic shock should be investigated. For instance, a 4°C soak and/or *in situ* serial transfer (*see* **Subheading 3.2.**) could be tried, or more exotic cryoprotectants, such as mixtures of light and heavier sugars, could be used, e.g., 15% trehalose + 15% sucrose is particularly good for salt-based crystal buffers.

Table 1

Major component of mother liquid	Suggested cryoprotectant agents/strategy
All components	Add 15–25% glycerol (works in around two-thirds of cases)
Polyethylene glycol (PEG) < 4K	Increase PEG, add small molecular weight PEGs
PEG > 4K	Add small molecular weight PEGs
Crystal screens solution I *(14)*	Glycerol concentrations required given in **ref.** *15*
Crystal screens solutions I and II *(14)*	PEG400, ethylene glycol, glycerol, and 1,2-propanediol concentrations required given in **ref.** *16*
Methylpentanediol (MPD)	Increase MPD concentration
Salt (low salt concentration requires a greater concentration of cryoprotectant than high salt)	Add MPD and/or ethylene glycol or glycerol Increase concentration/add salt *(17)* Exchange salt *(18)*

If initial attempts at cryocooling result in no diffraction, it is always worth testing a crystal at room temperature (*see* **Note 2**) before trying different cryoprotocols.

3.2. Crystal Handling and Transfer

As a general rule, the less a crystal is handled the better. Handling can cause damage to the crystal surface as well as dehydration, especially if the crystal is transferred in the loop through air several times. These traumas invariably result in an increase in the mosaicity of the crystal. Data quality is best when the mosaicity is minimized *(19)*, as the signal-to-noise ratio is higher and weak reflections can be measured more accurately.

The crystals may require gentle surgery to remove the "skin" that has developed on the surface of the crystallization drop, or to separate clusters before being mounted in a loop. Acupuncture needles are very useful for this purpose as they are thin, slightly flexible, and better than syringe needles because there is no capillary action and thus no liquid removal, which might result in dehydration of the crystal. The needles can be reused many times.

There are broadly three ways to transfer a crystal from its growth drop into the cryosolution.

1. Co-crystallization. The best scenario is when the crystal mother liquor already contains a high enough concentration of cryoprotecting agent for flash cooling and that

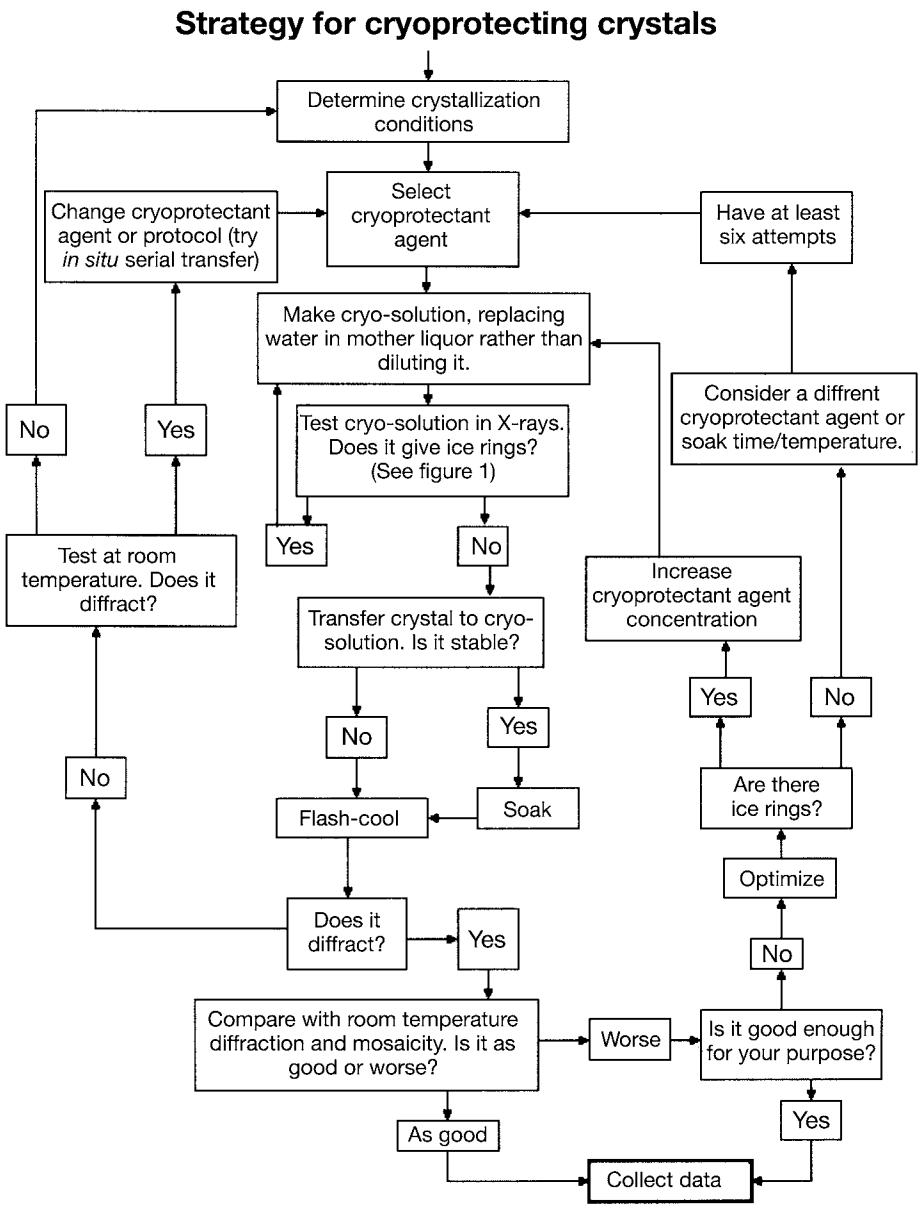


Fig. 3. Flow chart showing a possible strategy to follow when selecting a suitable cryoprotocol. If an initial flash-cooling attempt proves unsuccessful, there are several alternative routes to try. For example, changing the time and temperature of the soak or dragging the crystal through oil following a soak in cryosolution. (Reproduced from **ref. 20** with permission from the IUCr.)

no more need be added. The crystal can then be fished directly from the mother liquor for cooling. Co-crystallization with cryoprotecting agents is an increasingly used method (12.5% of cases; *see* **Fig. 2**). There are now several commercial crystallization screens available in which the solutions are already fully cryoprotected.

2. Soaking. The crystal is transferred from its growth drop with a loop (*see* **Notes 3–5**) straight into the final concentration of cryosolution. For fragile crystals, a suction device made from a pulled Pasteur pipet, a piece of flexible tube, and a small syringe can be used. The crystal is left in the cryosolution for anything from 1 s ("quick dip") to several days, depending on whether or not it is stable. A kinder strategy for the crystal is to transfer it first into a small volume of mother liquor, and then to pipet a volume of the cryosolution onto it, mix with the pipet tip, remove an equal volume, and repeat. This method can also be used for serial increase of the cryosolution concentration, starting with 10% (v/v) added, mixed, and removed twice, moving onto 20% and so on until the required concentration is reached; 40% can be reached in about 3 min using this method, and in many cases this strategy has been found to produce better results than quick dips or serial transfers where the crystal is moved between drops of increasing concentration. It involves a lot less crystal manipulation and a gentler gradient to the cryoagent concentration increase *(4)*.

   For crystallization drops containing only one crystal, the cryoprotecting agent can be added directly into the drop. This method can also be used for drops containing many crystals if no benign artificial mother liquor can be found, though often at the expense of the crystals remaining in the drop after removal of the selected crystal.
3. Dialysis. In pathological cases, the cryosolution can be dialyzed slowly into the crystal. This method is not often used, but should be attempted when a benign cryosolution cannot be found, and the use of oil fails.

3.3. Crystal Storage and Retrieval

Experimentally, the storage and retrieval of flash-cooled crystals is fraught with pitfalls. The success rate is lower than desirable, and ice is often a problem on retrieved crystals. In fact, ice formation on, in, and around the crystal can be a limiting problem in cryocrystallography, but in all cases can be avoided (*see* **Notes 6** and **7**). The storage and retrieval hardware is currently under development to accommodate crystal-mounting robots, and thus success rates and ease of use should improve before too long.

Below we give step-by-step instructions for the more common operations. The tools referred to are a wand (a stick with a magnet with its plane perpendicular or at around 60° to the stick), a vial holder (a stick with an insulated handle having a delrin block on the end that has the shape of the vial machined out of it), self-opposing tweezers, vials (which may or may not have a magnetic strip around the top to hold the top-hat in), cryocanes, cryosleeves, removable arc, cryotongs, and a lid undoer (a plastic rod with the end machined so that it fits precisely into the inside of a vial lid).

3.3.1. General Rules

1. Safety goggles should be worn while handling liquid nitrogen.
2. Gloves should also be worn. There are no satisfactory commercially available gloves that allow movement and protection. A possible solution is to wear thermally insulating gloves (such as KT1 and KT2, available from www.marigoldindustrial.com) with laboratory gloves (latex or nitrile) over the top. Laboratory gloves are too thin on their own, and the thermal gloves must not be used on their own as they soak up liquid nitrogen and hold it next to the skin for efficient burning.
3. A standard pin length should be established; this pin length will keep crystals in the center of the cryostat nitrogen stream at all times, even when rotated on the removable arc (*see* **Subheading 3.3.4.1.**). Ideally all of the pins and top-hats in the laboratory should conform to the standard length. This will save great aggravation in the long run.
4. In all cases a tall cylindrical 2-L Dewar (or a Dewar at least as deep as the length of a cryocane) should be filled with liquid nitrogen and used for holding cryocanes ready for use. In addition to this, a small working Dewar should be filled before starting work and used to precool vial holders (or vial-holding tongs) before use.
5. If more than four or five crystals are being cooled in succession, the small working Dewar should be emptied, warmed, dried, and refilled with fresh liquid nitrogen to avoid the incorporation into the vials of small snow balls, which accumulate in the Dewar when it is open to the moist room air.
6. All vials and canes should be labeled before cooling (it is impossible to label them when they are cold). It is advisable to do a "dry run" without a crystal in the loop before carrying out any of the procedures described with a crystal.
7. Before transport in the dry shipper (*see* **Note 8**), canes should be enclosed in plastic cryosleeves to retain the vials on the cane during rough journeys. The dry Dewar should be filled with liquid nitrogen and cooled down the day before it is required: six or seven fills at regular intervals are usually required to cool down the material inside to 77K.
8. The crystal must be kept below the glass-transition temperature of the cryosolution at around 155K ***(21)*** (increased from the 136K of pure water by the mother liquor constituents) at all times (except if annealing, *see* **Note 9**).

3.3.2. Flash Cooling Into Liquid for Storage or Transport

1. Place the small working Dewar of liquid nitrogen as close to the microscope as possible. Hold the top-hat and pin on the magnetic wand.
2. Fish the crystal (*see* **Notes 3–5**) from its cryosolution and plunge it immediately into the liquid nitrogen. The shorter the time the crystal is traveling through the air and dehydrating, the better. Make sure that the Dewar is filled to the top with liquid nitrogen.
3. Wait until the liquid nitrogen surface is calm.
4. Bring the vial holder and vial up underneath the pin and top-hat (*see* **Note 10**). Locate the top-hat into the top of the vial.
5. Move the wand *sideways* toward the handle of the vial holder to break the magnetic hold between the top-hat and wand.

3.3.3. Flash Cooling in a Gas Stream

1. Pre-center the loop in the nitrogen stream. Check the centering of the cryostream nozzle. Place the microscope as close as possible to the goniometer; inside the generator radiation enclosure is ideal.
2. Fish the crystal by holding the top-hat in a pair of self-opposing tweezers; those with bent tips are most convenient. Move immediately toward the goniometer head, and with your free hand, use a small narrow piece of flexible card to block the nitrogen stream.
3. When the crystal is safely on the goniometer head and your tweezers hand is out of the way, remove the card quickly. Use of the card is essential for optimum results because it ensures the crystal is not cooled and then warmed up by being moved in and out of the stream and that cooling is fast. Very importantly, it also prevents the crystal dehydrating in the dry air outer sheath.

3.3.4. Retrieval of a Crystal on the Goniometer Head Into a Dewar

3.3.4.1. Using a Removable Arc (*see* **Fig. 4**)

1. Rotate the goniometer so that the arc attachment screw hole is upwards. Attach the removable arc (**Fig. 4B**). If the cryostream nozzle is in the way, rack the nozzle back a little.
2. Slide the top-hat round the arc so that it is pointing downward, by pushing the horizontal metal bar round gently (**Fig. 4C**).
3. Put on protective gloves.
4. Bring the small Dewar, vial holder, and vial as near to the goniometer as possible; inside the generator radiation enclosure is ideal.
5. Bring the vial holder and the vial full of liquid nitrogen up under the pin and carefully enclose the top-hat within the vial. Move the vial holder *sideways* to break the magnetic hold of the magnet on the hat. If necessary use the forefinger of your free hand to nudge the top-hat sideways from the other side of the holding magnet. Try not to panic at this stage. You have approx 25 s before the liquid nitrogen will boil off to below the position of your crystal in the vial (**Fig. 4D**).
6. Immerse the vial holder and vial into the small Dewar as soon as possible.
7. If putting the vial lid on (advisable if using canes with lateral extrusions to rest the vial base on, otherwise optional), make sure there is a "blow hole" drilled out of the top of the lid beforehand. Bring the vial holder up so that the top of the vial just breaks the surface of the liquid nitrogen. Put the lid on the end of the white delrin rod ("lid undoer") and screw it on the vial. *Do not* screw it on too tightly as the lids can sometimes freeze on and be very hard to get off. Immerse the vial and lid in the liquid nitrogen.
8. When the vial is at equilibrium with the nitrogen (the surface of the liquid becomes calm), lift out the vial holder and remove the vial (you should still be wearing gloves) with your thumb on the top of it, your palm *upward*, and use your third and fourth fingers to hold the vial near the bottom on each side. Put the vial holder down. With your free hand lift the prelabeled cane out of the 2-L Dewar and put the vial on it at the next free position, working from the bottom of the cane upwards.

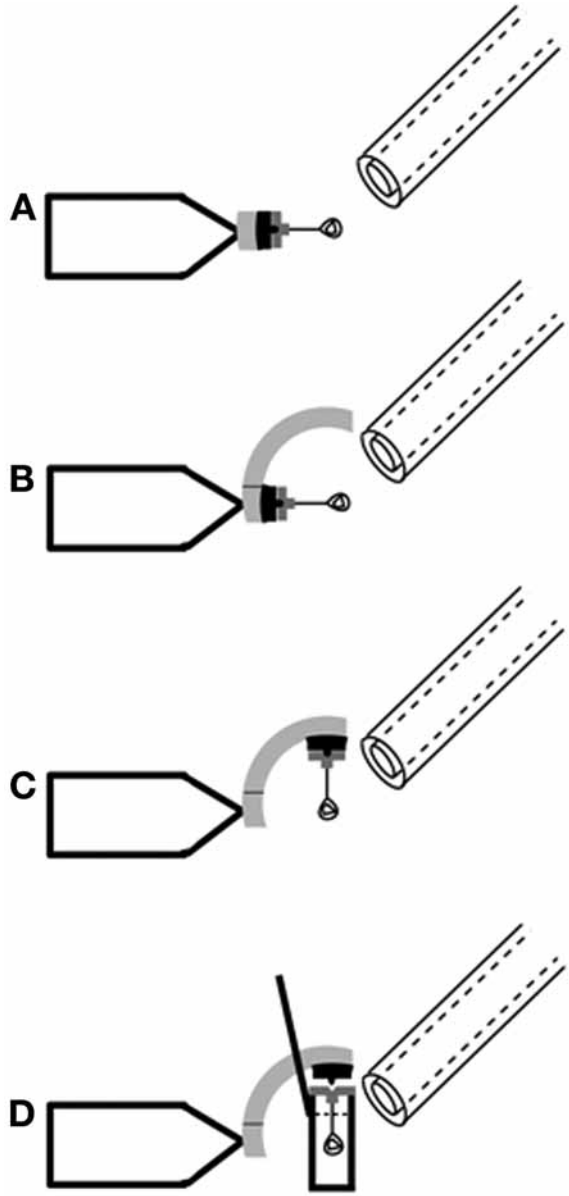


Fig. 4. Storage of a crystal using a removable arc goniometer. **(A)** A modified goniometer is shown and the removable arc has been attached to this **(B)**. **(C)** The crystal has been moved to the vertical position and in **(D)** a cryovial is brought up for crystal removal and subsequent storage. (Reprinted from **ref.** *2* with permission from the IUCr.)

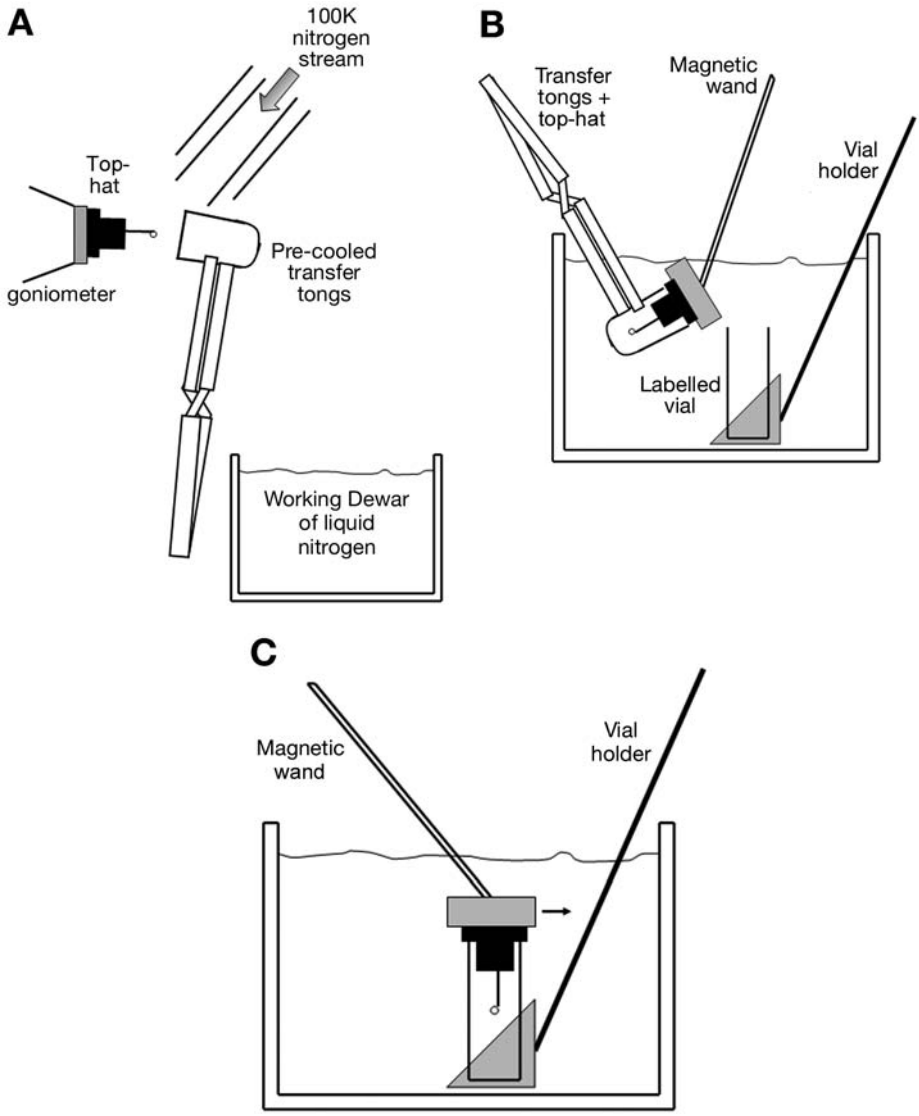


Fig. 5. Storage of a crystal using cooled transfer tongs. **(A)** A crystal has been flash cooled into gaseous nitrogen and is about to be grasped with precooled transfer tongs. The stream of 100K nitrogen must not be blocked by the approach of the tongs. **(B)** The tongs are plunged into a small working Dewar of liquid nitrogen positioned nearby and a magnetic "wand" is brought into contact with the top-hat, still held in the tongs. **(C)** The tongs are opened and removed, and the crystal is steered into a cryovial for storage on a cryocane. For retrieval from storage, the process is carried out in reverse. (Reprinted from **ref.** *13* with permission from Elsevier.)

9. Make a note of the crystal shape, the marking on the vial, the marking on the cane, and the date.
10. When the canes are full, transfer them to the main storage Dewar all at the same time. This avoids opening and closing storage Dewar unnecessarily, preventing ingress of moisture from the air. Make sure someone is responsible for topping up the storage Dewar regularly.

3.3.4.2. Using Cryotongs (**Fig. 5**)

1. Place the warm dry cryotongs in the small Dewar and put the Dewar as close to the goniometer as possible; inside the generator housing is ideal. Wait for the tongs to cool down.
2. Bring the cold tongs up to the crystal orientated such that the tongs never come between the stream and the crystal. This action will need some pre-experimentation so that you know exactly where to place the tongs. Separate the tongs and clasp the top-hat and crystal in them. Remove them sideways to break the magnetic hold between the goniometer and top-hat (**Fig. 5A**).
3. Immerse the tongs well into the small Dewar as soon as possible (for a noncoaxial cryostream you have only 1–2 s before the crystal warms up above 155K).
4. Transfer the crystal to the magnetic wand by placing the wand in the Dewar and letting it cool down. Bring the wand up to the base of the top-hat and release the tongs carefully. Remove the tongs from the Dewar (**Fig. 5B**).
5. Bring the vial holder and vial up underneath the pin and top-hat (*see* **Note 10**). Locate the top-hat into the top of the vial.
6. Move the wand *sideways* to break the magnetic hold between the top-hat and wand (**Fig. 5C**).
7. Continue as when using the removable arc from **Subheading 3.3.4.1.**, **step 7**.

Note that the tongs should be warmed up and dried after each use to stop them from freezing together with ice because of water condensing from the air.

3.3.5. Retrieval From Transport/Storage Dewar Back Onto the Goniometer

3.3.5.1. Using the Removable Arc

1. Put the removable arc on the goniometer and position the magnet on it so that the top-hat and pin will point downward when on the magnet.
2. Check that you can gain access to the magnet with the vial holder by doing a "dry run" with an empty vial in the holder.
3. Transfer the cane that is holding the desired crystal from the transport/storage Dewar or dry Dewar to the 2-L Dewar.
4. Put the vial holder into the small Dewar to cool down.
5. Put on protective gloves.
6. Lift the cane out of the 2-L Dewar and take off the required vial with the other hand or with the vial-holding tongs.

7. Place the vial in the vial holder (or the vial-holding tongs) and put both into the small Dewar.
8. If there is a lid on the vial, remove it using the delrin lid undoer.
9. Move the small Dewar as close as possible to the goniometer.
10. Bring the vial holder up underneath the magnet until the top-hat is located on the magnet. Carefully withdraw the vial to leave the crystal in the nitrogen stream.
11. Move the crystal round the arc to the data collection position by pushing on the metal bar on the sliding piece.
12. Unscrew the removable part of the arc and proceed with the final crystal alignment.

Note that if the sliding part of the arc is too loose, it can be adjusted by tightening the small screws in the underside of the sliding magnet.

3.3.5.2. Using the Tongs

1. Check that you can gain access to the magnet with the tongs by doing a "dry run" with an empty top-hat in the tongs. Orientate the tongs such that the nitrogen stream can reach the crystal as soon as the tongs are opened, i.e., an uninterrupted line can be drawn between the cryonozzle and the crystal through the open tongs.
2. Put the tongs and wand in the small Dewar to cool down.
3. Transfer the cane that is holding the desired crystal from the storage Dewar or dry Dewar to the 2-L Dewar.
4. Put the vial holder into the small Dewar to cool down.
5. Put on protective gloves.
6. Lift the cane out of the 2-L Dewar and take off the required vial with the other hand or with the vial-holding tongs.
7. Place the vial in the vial holder (or the vial-holding tongs) and put both into the small Dewar.
8. If there is a lid on the vial, remove it using the lid undoer.
9. Put the wand on the top-hat and withdraw it from the vial keeping the hat under the liquid nitrogen.
10. Take the vial holder and empty the vial out of the small Dewar.
11. When the tongs are cold (surface of liquid nitrogen calm), open them and grasp the top-hat. Withdraw the wand sideways and put it down.
12. Move the small Dewar as near the goniometer as possible.
13. Lift the tongs to the goniometer magnet and locate the top-hat base on the magnet. Release tongs so as not to block the nitrogen stream and carefully withdraw them.
14. Proceed with crystal alignment.

4. Notes

1. Propane has become increasingly unpopular because of the safety implications for its transport. Its use will not be covered in this chapter. A recent theoretical study modeling cooling rates for the flash cooling of protein crystals in loops *(22)* concluded that the choice of cryogen was of relatively low importance to successful cryocooling. The crystal solvent content and solvent composition came at the top of the list, followed by the crystal size and shape (crystals with large surface-to-volume [s/v] values cool faster and more uniformly than those

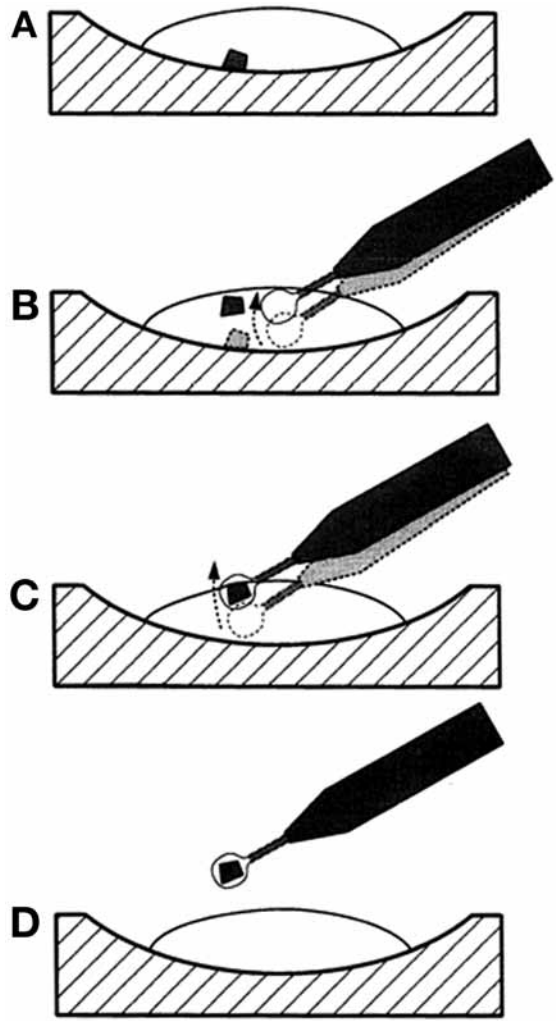


Fig. 6. Steps in lassoing a crystal in a fiber loop for flash cooling. Physical contact between the crystal and the loop is minimized. The plane of the loop is kept perpendicular to the surface during removal from the drop, to equalize the surface tension forces on the two sides of the loop. (Reprinted from **ref.** *25* with permission from the IOS Press.)

with small s/v), amount of residual liquid around the crystal (which should be minimized), and the cooling method (liquid or stream cooling). This theoretical study has given some rationale to procedures that have been empirically determined over the last 10 yr.

2. Room temperature testing of crystals can conveniently be carried out by mounting the crystal in mother liquor in a cryoloop and then covering it in a glass capillary

sealed with plasticine to the brim of the top-hat *(23)*. Alternatively, with more difficulty it can be mounted on the inner wall of a sealed glass capillary, a procedure for which is presented in **ref.** *24*.

3. A fishing procedure is shown in **Fig. 6**: the liquid is first agitated to lift the crystal from the bottom of the drop. If it is stuck, an acupuncture needle can be gently used to free it. The loop is brought up to the crystal from the side with its plane vertical, so that it encloses the crystal. The loop is then pulled upward so that its edge breaks the surface tension of the drop. Leaving the plane of the loop perpendicular to the surface of the liquid minimizes the forces on the crystal and also minimizes the thickness of the liquid film.
4. When fishing for crystals in the cryosolution (**Fig. 6**), have the microscope on fairly low magnification so that the whole drop is visible and the depth of field is larger than on high magnification.
5. When fishing, rest the fourth finger of the fishing hand on the edge of the crystallization tray to steady your hand and act as an anchor point.
6. Adjusting the geometry of the experiment can help to avoid ice. The stream of nitrogen must *not* hit the pin first, and use the cryostat alignment nozzle *(20)* to ensure that the cryostream is centered. Position the end of the nozzle 5–8 mm from the crystal. Shielding the crystal position from draughts helps ensure the crystal is kept at cryotemperatures the whole time. There is absolutely no need for ice to be a problem. For a full trouble-shooting guide, *see* **ref.** *2*.
7. Ice can sometimes gradually accumulate on the crystal during data collection. It can be removed using an acupuncture needle or an artist's brush. Alternatively. a small volume of liquid nitrogen can be poured over it, but if the crystal-viewing camera is directly below the crystal, beware of cracking the lens!
8. Dry shipping Dewar care (*see also* manufacturer's instructions): the dry shipper should be properly dried out after each use, especially if it was opened and closed many times at the synchrotron. Moisture can get into the absorption material and seriously compromise its cooling capacity; the next time the Dewar is cooled with liquid nitrogen, the water freezes to ice in the material, which hinders nitrogen absorption.
9. If the crystal has cooled poorly, annealing can be attempted to reduce the mosaicity and increase the resolution. Annealing can be carried out in two ways, either by blocking the gas stream temporarily *(26)* or by putting the crystal back in the cryoprotectant and then flash cooling again *(27)* (*see* Chapter 3). An understanding of the annealing process is now starting to emerge *(28,29)*.
10. When manipulating crystals and vials under liquid nitrogen, wait until the liquid has stopped bubbling and is calm before trying to see where the objects are under the liquid. Also, it is easier to move only one hand at a time and rest the side of the stationary hand on the edge of the Dewar top.

References

1. Rodgers, D. W. (1997) Practical cryocrystallography. In: *Methods in Enzymology, vol. 276*, (Carter, C. W., Jr., and Sweet, R. M., eds.), Academic Press, San Diego, CA, pp. 183–202.

2. Garman, E. and Schneider, T. (1997) Macromolecular cryocrystallography. *J. Appl. Cryst.* **27,** 211–237.
3. Parkin, S. and Hope, H. (1998) Macromolecular cryocrystallography: cooling, mounting, storage and transportation of crystals. *J. Appl. Cryst.* **31,** 945–953.
4. Garman, E. (1999) Cool data: quality and quantity. *Acta Cryst*. **D55,** 1641–1653.
5. Hope, H. (2001) Introduction to cryocrystallography. In: *International Tables for Crystallography: Volume F, Crystallography of Biological Macromolecules,* (Rossmann, M., Arnold, G., and Dordrecht, E., eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 197–201.
6. Rodgers, D. W. (2001) Cryocrystallography techniques and devices. In: *International Tables for Crystallography*: *Volume F, Crystallography of Biological Macromolecules,* (Rossmann, M. G., and Arnold, E. eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 202–208.
7. Henderson, R. (1990) Cryo-protection of protein crystals against radiation damage in electron and X-ray diffraction. *Proc. R. Soc. Lond. B* **241,** 6–8.
8. Weik, M., Ravelli, R. G. B., Kryger, G., et al. (2000) Specific chemical and structural damage to proteins produced by synchrotron radiation. *PNAS* **97,** 623–628.
9. Burmeister, W. P. (2000) Structural changes in a cryo-cooled protein crystal owing to radiation damage. *Acta Cryst.* **D56,** 328–341.
10. Ravelli, R. G. B. and McSweeney, S. (2000) The 'fingerprint' that X-rays can leave on structures. *Structure* **8,** 315–328.
11. Garman, E. (2003) 'Cool' crystals: cryocrystallography and radiation damage. *Curr. Opin. Struct. Biol.* **13,** 545–551.
12. Teng T-Y. (1990) Mounting of crystals for macromolecular crystallography in a free-standing thin-film. *J. Appl. Cryst*. **23,** 387–391.
13. Garman, E. and Doublié, S. (2003) Cryocooling of Macromolecular Crystals: Optimisation methods. *Methods in Enzymology, vol. 368,* (Carter, C. W., Jr. and Sweet, R. M., eds.), Academic Press, San Diego, CA, pp. 190–217.
14. Jancarik, J. and Kim, S. -H. (1991) Sparse matrix smpling: a screening method for crystallisation of proteins. *J. Appl. Cryst.* **24,** 409–411.
15. Garman, E. F. and Mitchell, E. M. (1996) Glycerol concentrations required for cryoprotection of 50 typical protein crystallisation solutions. *J. Appl. Cryst.* **29,** 584–587.
16. McFerrin, M. and Snell, E. (2002) The development and application of a method to quantify the quality of cryoprotectant solutions using standard area-detector X-ray images. *J. Appl. Cryst.* **35,** 538–545.
17. Rubinson, K. A., Ladner, J. E., Tordova, M., and Gilliland, G. L. (2000) Cryosalts: suppression of ice formation in macromolecular crystallography. *Acta Cryst.* **D56,** 996–1001.
18. Wierenga, R. K., Zeelan, J. P., and Noble, M. E. (1992) Crystal transfer experiments carried out with crystals of triosephosphate isomerase (TIM) *J. Cryst. Growth* **122,** 231–234.
19. Mitchell, E. M. and Garman, E. F. (1994) Flash freezing of protein crystals: investigation of mosaic spread diffraction limit with cryoprotectant concentration. *J. Appl. Cryst* **27,** 1070–1074.

20. Garman, E. F. and Owen, R. L. (2006) Cryocooling and radiation damage in macromolecular crystallography. *Acta Cryst.* **D62,** 32–47.
21. Weik, M., Kryger, G., Schreurs, A. M. M., et al. (2001) Solvent behaviour in flash-cooled protein crystals at cryogenic temperatures. *Acta Cryst.* **D57,** 566–573.
22. Kriminski, S., Kazmierczak, M., and Thorne, R. E. (2003) Heat transfer from protein crystals: implication for flash-cooling and X-ray beam heating. *Acta Cryst.* **D59,** 697–708.
23. Skrzypczak-Jankun, E., Bianchet, M. A., Amzel, L. M., and Funk, M.O., Jr. (1996) Flash-freezing causes a stress-induced modulation in a crystal structure of soybean lipoxygenase L3. *Acta Cryst.* **D52,** 959–965.
24. Garman, E. F. (1999) Crystallisation for cryo-data collection. In: *Protein Crystallisation: Techniques, Strategies and Tips,* (Bergfors, T. M., ed.), International University Line, La Jolla, CA, pp. 183–195.
25. Garman, E. F. (2001) Macromolecular cryo-crystallography. In: *Methods in Macromolecular Crystallography* (Turk, D. and Johnson, L., eds.), IOS Press, Amsterdam, The Netherlands, pp. 15–25.
26. Yeh, J. I. and Hol, W. G. J. (1998) A flash-annealing technique to improve diffraction limits and lower mosaicity in crystals of glycerol kinase. *Acta Cryst*. **D54,** 479–480.
27. Hanson, B. L., Schall, C. A., and Bunick, G. J. (2003) New techniques in macromolecular cryocrystallography: macromolecular crystal annealing and cryogenic helium. *J. Struct. Biol.* **142,** 77–87.
28. Kriminski, S., Caylor, C. L., Nonato, M. C., Finkelstein, K. D., and Thorne, R. E. (2003) Flash-cooling and annealing of protein crystals. *Acta Cryst.* **D58,** 459–471.
29. Juers, D. H. and Mathews, B. W. (2004) The role of solvent transport in cryo-annealing of macromolecular crystals. *Acta Cryst.* **D60,** 412–421.

2

Determination of Reaction Intermediate Structures in Heme Proteins

Kelvin Chu

Summary

Developments in structural biology and molecular biology have allowed increasingly detailed investigations of structure–function relationships. Although atomic-resolution structures of proteins are becoming more common, a growing number of structural studies have focused on the role played by dynamics and have sought to determine the structure of intermediates in protein reactions. These experiments have revealed the first atomic-level pictures of enzyme catalysis and the conformational motions required for biological function. This chapter uses the cryotrapping of reaction intermediates in horse heart myoglobin (Mb) to illustrate the methods utilized in determining the structures of reaction intermediates in protein systems. The techniques described here are applicable to a wide variety of heme proteins including Mb, hemoglobin, photosynthetic reaction centers, and cytochrome p450cam.

Key Words: Myoglobin; cytochrome p450; heme proteins; reaction intermediates; cryotrapping.

1. Introduction

Developments in structural biology and molecular biology have allowed increasingly detailed investigations of structure–function relationships. A growing number of structural studies have focused on the role played by dynamics and have sought to determine the structure of intermediates in protein reactions. These experiments have revealed the first atomic-level pictures of enzyme catalysis and the conformational motions required for biological function *(1–13)*. Here, we use the cryotrapping of reaction intermediates in horse heart myoglobin (Mb) to illustrate the methods utilized in determining the structures of reaction intermediates in protein systems. The techniques described here are applicable to a wide variety of heme proteins including Mb *(14–17)*, hemoglobin *(18)*, photosynthetic reaction centers *(19)*, and cytochrome p450cam *(20)*.

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

Examples of studies of reaction intermediates using structural techniques include the rapid trapping of intermediates in isocitrate dehydrogenase from *Escherichia coli* (21), the self-cleavage reaction in RNA catalysis in the hammerhead ribozyme (22), intermediates in the photocycles of photoactive yellow protein (23–25) and bacteriorhodopsin (8–13,26,27), cryotrapping and isolation of the intermediates along the catalytic pathway of cytochrome p450cam from *Pseudomonas putida* (20), ligand migration through Mb (16,20,28), oxygen activation in cytochrome *cd1* nitrite reductase (14), the nucleotidyl transferase pathway of DNA polymerase β (29), and deacylation in a serine protease (30).

Other techniques have been used to investigate systems that are inaccessible to X-ray studies. Electron crystallography has been used to determine the structure of the N intermediate in the bacteriorhodopsin photocycle (27) and the mechanism of proton translocation (26). Three-dimensional nuclear magnetic resonance has been used to examine the structure of the aspartyl-phosphorylation switch in the bacterial enhancer-binding protein NtrC under steady-state phosophorylation conditions (31).

These studies provide the foundation for interpreting biochemical and biophysical data and have been used to elucidate mechanisms of enzyme action (32), laying the groundwork for further study by mutagenesis and computation. We describe the materials and methods for characterization of the reaction intermediates in Mb. Because experimental protocols are crucial in the kinetic characterization of intermediates, we present a detailed rationale for sample treatment and preparation.

2. Materials

2.1. Crystallization of Mb

1. Horse heart Mb (Sigma, St. Louis, MO).
2. 1.7–1.8 *M* Ammonium sulfate solution.
3. 0.1 *M* Tris-HCl solution, pH 7.5.
4. 3.4–3.6 *M* Ammonium sulfate solution.
5. 0.1 *M* Tris-HCl solution, pH 7.4.
6. 50 m*M* Sodium dithionite solution.
7. 70 m*M* Potassium phosphate solution.
8. Hanging-drop crystallization trays (Hampton Research, Alisa Viejo, CA).
9. Carbon monoxide gas (Merriam-Graves, Springfield, MA).
10. Liquid nitrogen.

2.2. Experimental Equipment

1. Microspectrophotometer (32).
2. Open flow helium cryostream capable with temperature control (*see* **Subheading 3.2.**).
3. 500 mW Argon ion laser (National Laser Company, Salt Lake City, UT).

4. CryoCap system and crystal handling tools (Hampton Research).
5. Fiber optic illuminator (Oriel, Stratford, CT).
6. O_2 pressure cell for crystallography *(33)*.

3. Methods

The methods following the outline: (1) the crystallization of horse heart Mb, (2) characterization of the protein, (3) the preparation of ultra-low temperature cryocrystallography, and (4) data collection strategies.

3.1. Crystallization of Mb and Derivatives

Horse heart Mb is crystallized at room temperature by equilibrating 10-mL drops of 5 mg/mL protein in 1.7–1.8 *M* ammonium sulfate and 0.1 *M* Tris-HCl, pH 7.5, against 1 mL of 3.4–3.6 *M* ammonium sulfate and 0.1 *M* Tris-HCl, pH 7.4, using the hanging-drop geometry. Crystals should appear within several days with rosette-shaped crystals within 2 wk. Leaflets from the crystals can be harvested from the rosettes using Hampton Microtools. Typical sizes of crystals are 0.01 × 0.07 × 0.3 mm. These crystals are high-spin iron (*met*, Fe^{III}) and should appear brown. Crystals are typically P21 with characteristic unit cell dimensions of a = 63.6 Å, b = 28.8 Å, c = 35.6 Å, and β = 106.5 Å *(28,34)*. These crystals are the starting point for preparation of ligand derivatives:

1. Ferrous unligated Mb crystals are obtained by soaking *met* crystals in a nitrogenated solution containing 50 m*M* sodium dithionite, 70 m*M* potassium phosphate, and 70% saturated ammonium sulfate. The color of the crystals should change from brown to bright red. Deoxy Mb crystals should be flash frozen quickly in liquid nitrogen to prevent O_2 substitution from the atmosphere.
2. Ferrous CO-bound crystals are obtained by soaking *met* crystals in CO-saturated mother liquor supplemented by 8 mg/mol dithionite, 5 mL 2 *M* NaOH/mL, and 7.5% glycerol. The color of the crystals should change from brown to raspberry red over the period of 30 min. MbCO crystals should be flash frozen immediately in liquid nitrogen to prevent auto-oxidation.
3. Ferrous O_2-bound crystals are obtained by soaking *met* crystals in a solution of 50 m*M* potassium phosphate at pH 7.0, 70% saturated ammonium sulfate, 10% glucose (w/v), and 10% sucrose (w/v). Crystals are transferred into a pressure chamber and exposed to 100 bar of O_2 for 30 min at 4°C. Once ligated, pressure should be released from the chamber slowly (over tens of seconds) to prevent flash cooling from rapid depressurization. MbO_2 crystals should be flash frozen immediately in liquid nitrogen to prevent auto-oxidation.

3.2. Characterization of Intermediates

Structural characterization of reaction intermediates relies on independent confirmation of the nature of the intermediate, often by spectroscopy *(35)*. In addition, the nature of this identification is often *kinetic* as opposed to purely

spectroscopic. Because protein chromophores are often not sensitive to the ligand position in the protein, many studies have therefore relied on infrared spectroscopy in addition to UV/Vis spectroscopy, using the dipole absorption of the ligand in the infrared. The transient nature of intermediates in a protein reaction means that two techniques must be added to conventional structure determination: initiation of the reaction and sufficient accumulation of the reaction in crystals *(35,36)*. These techniques must be efficient, not overly damage the sample or radically change the kinetic properties of the system, and must be rapid compared with the time-scale of the reaction being studied *(36)*.

3.2.1. Reaction Initiation

For Mb, the initiation of the reaction is accomplished by photolysis. Carbon monoxide is used as a ligand because although it similar in many respects to O_2, the quantum yield for photolysis is unity. Time-resolved spectroscopic measurements and molecular dynamics simulations show that ligand rebinding in MbCO occurs in two sequential intermediates resulting from both movement of the ligand within the protein matrix and conformational relaxation of the protein.

Spectroscopic studies indicate that following dissociation, the spectrum of free CO appears in approx 0.5 ps and persists for hundreds of nanoseconds *(37)*. The extent of geminate, or internal, recombination from this site depends on the reactivity of the ligand with the heme iron and its ability to diffuse away from the active site to a secondary docking site (D) or the solvent (S). The latter process requires relatively large anharmonic protein fluctuations on a microsecond time-scale that are frozen out below the glass-transition temperature, around 180K *(38)*. The general scheme is shown in **Table 1**.

3.2.2. Accumulation of Intermediates

To observe intermediates crystallographically, one must accumulate roughly 30% occupancy in the experiment. Separate protocols that can distinguish between the B and D intermediates must be designed. Temperature-derivative spectroscopy (TDS), a technique adapted from low-temperature solid-state physics *(39)* is used to design the protocol for characterization of intermediates. In a typical experiment, the reaction is initiated and the temperature is ramped up linearly in time while spectra are collected for each Kelvin temperature increase. This procedure effectively sweeps thermally activated processes through the window of observable rates of the instrument.

The structural heterogeneity in the sample results in a distribution of rates. At low temperatures, only processes with low barriers for recombination can rebind within the time window of the spectrometer. Differences between spectra from consecutive temperatures are plotted as a contour plot of population or change in absorbance as a function of wavenumber and rebinding enthalpy (**Fig. 1**). As the

Table 1
Reaction Scheme for Carbonmonoxymyoglobin[a]

Reaction	MbCO	⇔	Mb*CO	⇔	Mb**CO	⇔	Mb + CO
State:	A		B		D		S
Name:	(bound)		(photolyzed)		(photorelaxed)		(deoxy)
Ligand position	Heme		1^o site		2^o site		solvent

[a]Photolysis protocols can drive the ligand to either state B (photolyzed) or state D (photorelaxed), depending on the temperature profile of the sample. Rebinding from B⇒A or B⇒D occurs sequentially. Enthalpies for rebinding are determined using temperature-derivative spectroscopy. H_{BA} ~12 kJ/mol, H_{DA} ~30 kJ/mol.

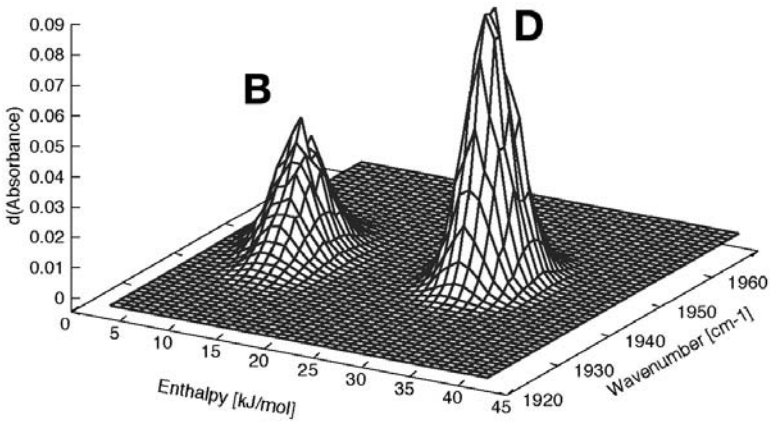


Fig. 1. Temperature-derivative spectroscopy plot for hMbCO. This plot reveals the rationale for design of the experiment. Samples photolyzed at 10K will generate only proteins in the B state. Samples cooled under illumination will generate both states B and D. However, collecting data at 90K allows rebinding of the B state, leaving the only nonligated CO molecules in the D position.

temperature is increased, individual ligand molecules gain sufficient energy to overcome higher barriers and recombine. Thus, TDS sorts different rate processes according to their activation enthalpy. The relation between enthalpy and peak rebinding temperature are related if we assume a first-order rate process,

$$\frac{dN}{dt} = -kN \tag{1}$$

and an Arrhenius-like rate,

$$k = A\left(\frac{T}{T_0}\right)\exp\left(-\frac{H}{RT}\right) \tag{2}$$

where A is the pre-exponential at the arbitrary reference temperature T_0, H is the height of the enthalpic barrier, R is the gas constant, and T is the temperature (Kelvin). The measured TDS signal can be then written as

$$\frac{dN}{dT} = -\frac{N_0}{\beta}\int_0^\infty k\exp\left(-\int_{T_0}^{T}\frac{k}{\beta}dT'\right)g(H)dH \tag{3}$$

The characteristic time for the measurement, τ_c, is given by

$$\tau_c = \frac{RT_p^3}{\beta T_p\left(H + RT_p\right)} \tag{4}$$

where the peak enthalpy is

$$H_p = T_p R\ln\left(A\tau_c\right) \tag{5}$$

Laser irradiation is used to enhance a particular photoproduct species, a process called pumping. Samples cooled under illumination undergo continuous photolysis while the temperature is ramped from 160 to 10K. The range of accessible conformational motions is frozen out as the sample cools, trapping long-lived states.

A TDS plot for hMbCO is shown in **Fig. 1**. Data are presented as a surface plot (or a contour plot) of population as a function of wavelength and rebinding enthalpy. Large signals (peaks) indicate increases in rebinding population for a particular wavelength and a particular rebinding enthalpy. Flat areas indicate no rebinding occurs.

3.3. Ultra-Low Temperature Crystallography

Crystallographic determination of structures of reaction intermediates relies on either rapid collection via the Laue diffraction or stabilization of intermediates on time-scales suitable for monochromatic data collection ***(40–42)***. The latter can be done chemically (e.g., by modification of the macromolecule, substrate, cofactor, solvent, or pH) or physically (by temperature) ***(43,44)***. Because most single-protein crystals are commonly flash cooled for stabilization during data collection to slow radiation damage, trapping intermediates by freeze quenching a reaction is very common.

Metalloporphyrin proteins often have a relatively large heme pocket designed to accommodate ligands, as in the case of hemoglobin or Mb, or large substrates, as in cytochrome P450cam. This results typically in a fast-rebinding geminate phase in flash-photolysis studies of heme protein kinetics, corresponding to low enthalpic barriers for rebinding (~4 kJ/mol) following photolysis. To adequately trap suitable populations of ligands in these low enthalpic barrier states, correspondingly low temperatures must be used. For Mb and hemoglobin, liquid helium temperatures are employed to slow recombination.

We have designed a helium cryostream that consists of a liquid helium stream for cooling the crystal and an exterior, concentric flow of warm, room-temperature gas (*see* **Note 1**). The gas flow rate of the inner stream is controlled by the needle valve of a liquid helium transfer line (Janis) (*see* **Note 2**). The regulator from a cylinder of helium gas controls the gas flow rate of the outer stream (*see* **Note 3**). Recently, cryostreams from commercial suppliers have become available. These ultra-low temperature cryostreams employ displex refrigerators for cooling, and need no cryogenic liquids.

3.4. Data Collection

*3.4.1. Mb*CO (B) Structure*

To determine the structure of the Mb*CO complex, crystals are photolyzed at helium temperatures to slow the rebinding reaction to experimentally convenient time-scales for monochromatic data collection. Experiments are aimed at optimizing the population of the D intermediate. Cryocooling prevents the escape of the photolyzed CO to the solvent.

Steps in data collection are:

1. Prepare MbCO crystals as described in **Subheading 3.1.** and mount crystals directly in the loop of a CryoCap.
2. Freeze the crystals directly in the helium stream. Note that prefreezing in liquid nitrogen is not possible.
3. Illuminate the crystal with low-intensity light, such as a fiber-optic illuminator (*see* **Note 4**). Collection of data should be performed with the illumination light on. Rebinding is non-negligible at helium temperatures and molecular tunneling of the CO can occur even at 4.2K, so samples should be under continuous illumination at low power during data collection. Simultaneously, too high an illumination will result in significant crystal heating and increased recombination *(17)*.

*3.4.2. Mb**CO (D) Structure*

The D intermediate requires relaxation of the protein. This occurs thermally at room temperature for native Mb with around 10% occupancy. Cryotrapping of the D state is required to accumulate sufficient amounts of the intermediate for structural determination by X-ray crystallography. For data collection of the photorelaxed hMb**CO complex, the crystal should be illuminated with an argon ion laser at a relatively high temperature (160–180K) and then cooled under illumination to 90–100K for trapping at a rate of 10K/h.

Steps in data collection are:

1. Prepare MbCO crystals as described in **Subheading 3.1.** and mount crystals directly in the loop of a CryoCap.
2. Freeze the crystals directly in a nitrogen stream.

3. Illuminate crystals with an argon ion laser (*see* **Note 5**). To ensure homogeneous illumination conditions, illuminate both sides of the flat surface of the leaflet for 6 h at 160K with an intensity of roughly 5 mW/mm^2.
4. Cool the crystal under illumination at 10K/h to 85K. Excessive illumination will cause heating of the sample. Temperature fluctuations on the order of 5K will result in significant rebinding during the experiment.

4. Notes

1. Despite the best alignment of cold and warm gas streams, turbulence and pressure regulation problems may eventually generate sufficiently nonlaminar flow so that moisture from ambient air will reach the cold helium stream and condense. This icing can be diminished by construction of small Kapton shields that extend over the CryoCap pin.
2. We have measured the temperature profile of the helium stream using a silicon diode mounted on a goniometer loop and found that crystallography at liquid helium temperatures requires much more precision in positioning of cryostream position with respect to the goniometer geometry. This is because the specific heat of helium is much lower than that of nitrogen, and, thus, the cooling power of the helium cryostream is greatly reduced. The typical protocol for cryostream positioning is to move the cryostream tip as close to the sample as possible without eclipsing high-resolution data. We use a very narrow cryostream tip (1.5-mm diameter) and typically have distances between the cryostream tip and the sample of less than 3 mm.
3. Obviously, samples should not be directly manipulated by hand when they are in the helium cryostream. When using prefrozen crystals, we have had good results with cryotongs that have been modified for the tight helium cryostream geometry. If, when mounting, nitrogen from LN2 freezes on the sample, it can be gently removed by direct manipulation with an unused cryoloop.
4. Heating from the photolysis and illumination beams can be significant. For photolysis, low-intensity illumination protocols should be used; heating in excess of 10–20°C occurs with direct illumination with a laser. If monochromatic illumination must be used, a standard trick is to choose a long wavelength where the crystal appears optically less thick. This strategy has been successfully employed by time-resolved studies.
5. Steps should be taken to ensure that photoselection does not occur during illumination. If a polarized light source is used for illumination and photolysis, it should not rotate with the crystal.

Acknowledgments

Thanks to Ilme Schlichting and Hans Frauenfelder for stimulating discussions. This work was supported by grants from the National Science Foundation, the Research Corporation, and the Petroleum Research Fund.

References

1. Stoddard, B. L. (1998) New results using Laue diffraction and time-resolved crystallography. *Curr. Opin. Struct. Biol.* **8,** 612–618.
2. Ridder, I. S., Rozeboom, H. J., Kalk, K. H., and Dijkstra, B. W. (1999) Crystal structures of intermediates in the dehalogenation of haloalkanoates by L-2-haloacid dehalogenase. *J. Biol. Chem.* **274,** 30,672–30,678.
3. Pannifer, A. D., Flint, A. J., Tonks, N. K., and Barford, D. (1998) Visualization of the cysteinyl-phosphate intermediate of a protein-tyrosine phosphatase by x-ray crystallography. *J. Biol. Chem.* **273,** 10,454–10,462.
4. Burzlaff, N. I., Rutledge, P. J., Clifton, I. J., et al. (1999) The reaction cycle of isopenicillin N synthase observed by X-ray diffraction. *Nature* **401,** 721–724.
5. Ogle, J. M., Clifton, I. J., Rutledge, P. J., et al. (2001) Alternative oxidation by isopenicillin N synthase observed by X-ray diffraction. *Chem. Biol.* **8,** 1231–1237.
6. Wilmot, C. M., Hajdu, J., McPherson, M. J., Knowles, P. F., and Phillips, S. E. (1999) Visualization of dioxygen bound to copper during enzyme catalysis. *Science* **286,** 1724–1728.
7. Murray, J. B., Szoke, H., Szoke, A., and Scott, W. G. (2000) Capture and visualization of a catalytic RNA enzyme-product complex using crystal lattice trapping and X-ray holographic reconstruction. *Mol. Cell.* **5,** 279–287.
8. Kuhlbrandt, W. (2000) Bacteriorhodopsin—the movie. *Nature* **406,** 569–570.
9. Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P., and Lanyi, J. K. (1999) Structural changes in bacteriorhodopsin during ion transport at 2 angstrom resolution. *Science* **286,** 255–261.
10. Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P., and Lanyi, J. K. (1999) Structure of bacteriorhodopsin at 1.55 A resolution. *J. Mol. Biol.* **291,** 899–911.
11. Edman, K., Nollert, P., Royant, A., et al. (1999) High-resolution X-ray structure of an early intermediate in the bacteriorhodopsin photocycle. *Nature* **401,** 822–826.
12. Royant, A., Edman, K., Ursby, T., Pebay-Peyroula, E., Landau, E. M., and Neutze, R. (2000) Helix deformation is coupled to vectorial proton transport in the photocycle of bacteriorhodopsin. *Nature* **406,** 645–648.
13. Sass, H. J., Buldt, G., Gessenich, R., et al. (2000) Structural alterations for proton translocation in the M state of wild-type bacteriorhodopsin. *Nature* **406,** 649–653.
14. Chu, K., Vojtchovsky, J., McMahon, B. H., Sweet, R. M., Berendzen, J., and Schlichting, I. (2000) Structure of a new ligand-binding intermediate in wildtype carbonmonoxymyoglobin. *Nature* **403,** 921–923.
15. Schlichting, I., Berendzen, J., Phillips, G. N., Jr., and Sweet, R. M. (1994) Crystal structure of photolyzed myoglobin. *Nature* **371,** 808–812.
16. Ostermann, A., Waschipky, R., Parak, F. G., and Nienhaus, G. U. (2000) Ligand binding and conformational motions in myoglobin. *Nature* **404,** 205–208.
17. Teng, T. Y., Srajer, V., and Moffat, K.(1994) Photolysis-induced structural changes in single crystals of carbonmonoxymyoglobin at 40K. *Nat. Struct. Biol.* **1,** 701–705.

18. Adachi, S., Park, S. Y., Tame, J. R., Shiro, Y., and Shibayama, N. (2003) Direct observation of photolysis-induced tertiary structural changes in hemoglobin. *Proc. Natl. Acad. Sci. USA* **100,** 7039–7044.
19. Stowell, M. H., McPhillips, T. M., Rees, D. C., Soltis, S. M., Abresch, E., and Feher, G. (1997) Light-induced structural changes in photosynthetic reaction center: implications for mechanism of electron-proton transfer. *Science* **276,** 812–816.
20. Schlichting, I., Berendzen, J., Chu, K., et al. (2000) The catalytic pathway of cytochrome P450cam at atomic resolution. *Science* **287,** 1615–1622.
21. Stoddard, B. L. (1999) Visualizing enzyme intermediates using fast diffraction and reaction trapping methods: isocitrate dehydrogenase. *Biochem. Soc. Trans.* **27,** 42–48.
22. Scott, W.G. (1999) Biophysical and biochemical investigations of RNA catalysis in the hammerhead ribozyme. *Q. Rev. Biophys.* **32,** 241–284.
23. Genick, U. K., Borgstahl, G. E., Ng, K., et al. (1997) Structure of a protein photocycle intermediate by millisecond time-resolved crystallography. *Science* **275,** 1471–1475.
24. Genick, U. K., Soltis, S. M., Kuhn, P., Canestrelli, I. L., and Getzoff, E. D. (1998) Structure at 0.85 A resolution of an early protein photocycle intermediate. *Nature* **392,** 206–209.
25. Perman, B., Srajer, V., Ren, Z., et al. (1998) Energy transduction on the nanosecond time scale: early structural events in a xanthopsin photocycle. *Science* **279,** 1946–1950.
26. Subramaniam, S. and R. Henderson (2000) Molecular mechanism of vectorial proton translocation by bacteriorhodopsin. *Nature* **406,** 653–657.
27. Vonck, J. (2000) Structure of the bacteriorhodopsin mutant F219L N intermediate revealed by electron crystallography. *Embo. J.* **19,** 2152–2160.
28. Brunori, M., Vallone, B., Cutruzzola, F., et al. (2000) The role of cavities in protein dynamics: crystal structure of a photolytic intermediate of a mutant myoglobin. *Proc. Natl. Acad. Sci. USA* **97,** 2058–2063.
29. Sjogren, T. and Hajdu, J. (2001) Structure of the bound dioxygen species in the cytochrome oxidase reaction of cytochrome cd1 nitrite reductase. *J. Biol. Chem.* **276,** 13,072–13,076.
30. Arndt, J. W., Gong, W., Zhong, X., et al. (2001) Insight into the catalytic mechanism of DNA polymerase beta: structures of intermediate complexes. *Biochemistry* **40,** 5368–5375.
31. Wilmouth, R. C., Edman, K., Neutze, R., et al. (2001) X-ray snapshots of serine protease catalysis reveal a tetrahedral intermediate. *Nat. Struct. Biol.* **8,** 689–694.
32. Kern, D., Volkman, B. F., Luginbuhl, P., Nohaile, M. J., Kustu, S., and Wemmer, D. E. (1999) Structure of a transiently phosphorylated switch in bacterial signal transduction. *Nature* **402,** 894–898.
33. Hadfield, A. and Hajdu, J. (1994) On the photochemical release of phosphate from 3,5-dinitrophenyl phosphate in a protein crystal. *J. Mol. Biol.* **236,** 995–1000.
34. Urayama, P., Phillips, G. N., and Gruner, S. M. (2002) Probing substates in sperm whale myoglobin using high-pressure crystallography. *Structure* **10,** 51–60.

35. Vojtechovsky, J., Chu, K., Berendzen, J., Sweet, R. M., and Schlichting, I. (1999) Crystal structures of myoglobin-ligand complexes at near-atomic resolution. *Biophys. J.* **77,** 2153–2174.
36. Schlichting, I. and Chu, K. (2000) Trapping intermediates in the crystal: ligand binding to myoglobin. *Curr. Opin. Struct. Biol*. **10,** 744–752.
37. Wilmot, C. M. and Pearson, A. R. (2002) Cryocrystallography of metalloprotein reaction intermediates. *Curr. Opin. Chem. Biol*. **6,** 202–207.
38. Schlichting, I. and Goody, R. S. (1997) Triggering Methods in crystallographic enzyme kinetics. *Meth. Enzymol.* **277,** 467–490.
39. Lim, M., Jackson, T. A., and Anfinrud, P. A. (1997) Ultrafast rotation and trapping of carbon monoxide dissociated from myoglobin. *Nat. Struct. Biol.* **4,** 209–214.
40. Chu, K., Ernst, R.M., Frauenfelder, H., Mourant, J. R., Nienhaus, G. U., and Philipp, R. (1995) Light-induced and thermal relaxation in a protein. *Phys. Rev. Lett.* **74,** 2607–2610.
41. Berendzen, J. and Braunstein, D. (1990) Temperature-derivative spectroscopy: a tool for protein dynamics. *Proc. Natl. Acad. Sci. USA* **87,** 1–5.
42. Petsko, G. A. and Ringe, D. (2000) Observation of unstable species in enzyme-catalyzed transformations using protein crystallography. *Curr. Opin. Chem. Biol.* **4,** 89–94.
43. Schlichting, I. (2000) Crystallographic structure determination of unstable species. *Acc. Chem. Res.* **33,** 532–538.
44. Ursby, T., Weik, M., Fioravanti, E., Delarue, M., Goeldner, M., and Bourgeois, D. (2002) Cryophotolysis of caged compounds: a technique for trapping intermediate states in protein crystals. *Acta Crystallogr. D. Biol. Crystallogr.* **58,** 607–614.

3

Annealing Macromolecular Crystals

B. Leif Hanson and Gerard J. Bunick

Summary

The process of crystal annealing has been used to improve the quality of diffraction from crystals that would otherwise be discarded for displaying unsatisfactory diffraction after flash cooling. Although techniques and protocols vary, macromolecular crystals are annealed by warming the flash-cooled crystal, then flash cooling it again. To apply macromolecular crystal annealing, a flash-cooled crystal displaying unacceptably high mosaicity or diffraction from ice is removed from the goniometer and immediately placed in cryoprotectant buffer. The crystal is incubated in the buffer at either room temperature or the temperature at which the crystal was grown. After about 3 min, the crystal is remounted in the loop and flash cooled. *In situ* annealing techniques, where the cold stream is diverted and the crystal allowed to warm on the loop prior to flash cooling, are variations of annealing that appears to work best when large solvent channels are not present in the crystal lattice or the solvent content of the crystal is relatively low.

Key Words: Macromolecular crystal annealing; cryocrystallography; flash cooling.

1. Introduction

Flash cooling protein or other macromolecular crystals for diffraction studies can often result in undesirable features that make data measurement and analysis troublesome. These features can include greatly increased mosaicity and ice rings in the diffraction pattern (*see* **Note 1**). To overcome the effects of crystal flash cooling, the techniques of crystal annealing, especially macromolecular crystal annealing (MCA) have been developed *(1,2)*. These techniques involving the heating and cooling of a crystal can be referred to as annealing or tempering the macromolecular crystal. As these terms are borrowed from metallurgy, neither completely matches the process involved. The term annealing implies a slow or gradual cooling of material after heating. In metals, tempering refers to the process of softening a material by heating to a certain temperature and then cooling. Our preferred terminology is annealing. Implicit in this

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

term is the removal of stresses and the production of definite microstructure. We believe this is the case when a crystal is warmed and flash cooled using macromolecular annealing techniques.

The annealing techniques in current use involve a variety of crystal treatments and warming strategies (**Fig. 1**). MCA refers to the specific protocol that introduced the concept of annealing, and which was developed originally with crystals of chromatin structural elements (*see* **Subheading 3.1.**). It now has been used to improve the diffraction data for a variety of macromolecules. Annealing a crystal without damage to the diffraction quality can simplify the handling of macromolecular crystals (*see* **Note 2**). In addition to flash-cooled crystals that display unsatisfactory mosaicity, MCA has been used to salvage crystals that became iced during cryogenic storage or because of problems with the cold stream. Crystal-mounting accidents have been resolved with MCA as well. For example, crystals have been recovered that have been flash cooled in a position outside of the loop, having been flicked out of the loop when the magnetic cap engaged the goniometer head too forcefully.

Several annealing protocols are described next; we recommend these in the specific circumstances detailed. We refer to annealing techniques where crystals are not removed from the goniometer *(3)* as *in situ* methods. A correlation has been noted between crystal solvent content and the outcome of *in situ* annealing (*[2,4]*; *see* **Note 3**). Despite the preference for and recommendation of MCA as primary annealing methodology, in circumstances where crystals are very thin or fragile, or if the crystal disintegrates when returned to the cryoprotectant solution, it may be necessary to consider an *in situ* annealing method.

As a general rule, we find that crystals with lower solvent content respond better to *in situ* annealing, whereas crystals with higher solvent content do not, and require MCA. The more salient aspect of the crystal solvent content may be the size of the solvent channels within the crystal *(5)*. Larger solvent channels can permit ice nucleation and more widespread crystal lattice disorder. Thus, the longer incubation time and lattice reformation that occur during MCA are needed to return useful diffraction from the crystal. Because solvent channel size and solvent content are usually established after diffraction data are measured, it may be safest to assume that a crystal of an as yet undetermined protein needing annealing should be treated with MCA.

2. Materials

1. Cryostat providing an open-flow of cryogen.
2. Macromolecular crystal.
3. Cryoprotectant buffer in which crystal is stable.
4. Cryocrystallographic-mounting loops (Hampton Research, Aliso Viejo, CA).

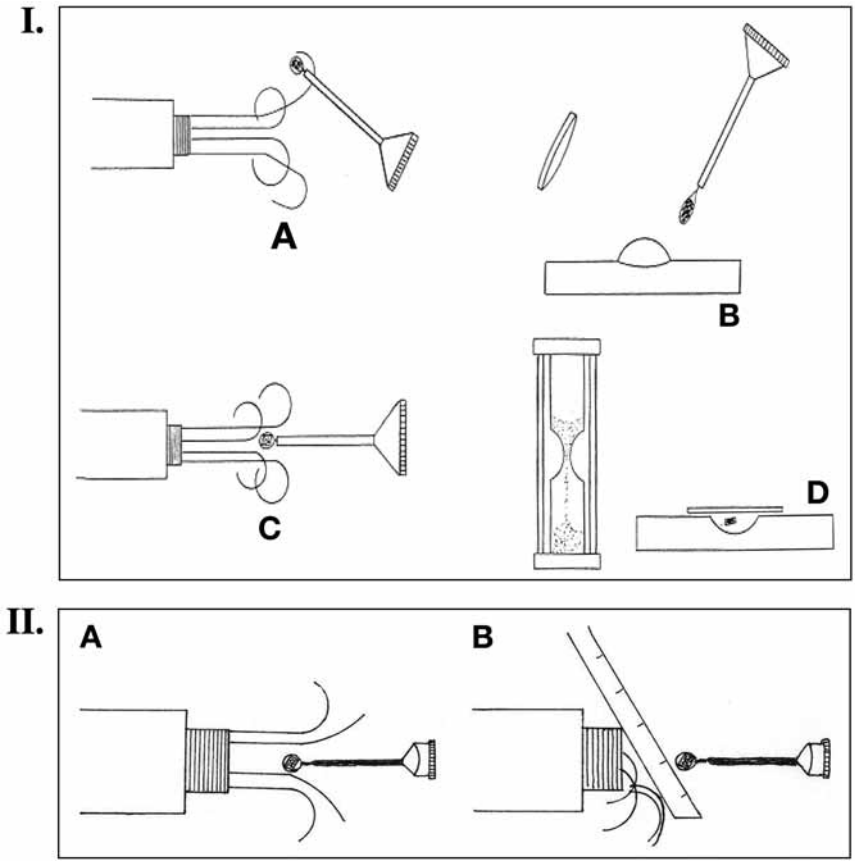


Fig. 1. Protocols for annealing macromolecular crystals. I. Macromolecular crystal annealing (MCA): **(A)** flash-cooled crystal is removed from the cold stream; **(B)** crystal is quickly transferred to a droplet of cryoprotectant solution; **(C)** crystal in cryoprotectant droplet is covered to prevent solvent change and allowed to incubate for a period of time; and **(D)** crystal is repositioned on loop, and flash cooled. One application of MCA is usually sufficient to restore the crystal diffraction to acceptable levels. II. *In situ* annealing: **(A)** flash-cooled crystal displays unacceptable diffraction; **(B)** cold gas stream is diverted from crystal allowing the crystal to rewarm. Removal of the diverter flash cools the crystal. This process can be repeated multiple times, but *see* **Notes 3** and **5**.

5. Well slide (Fisher, Hampton, NH) sufficient to hold 300 µL (for MCA).
6. Silanized-glass cover slips (Hampton Research) (for MCA).
7. 3-min Timer.
8. Cold gas stream diverter, usually a flexible plastic ruler or credit card (for *in situ* annealing).

9. Crystallization start buffer for nucleosome core particle (NCP): 50 m*M* KCl, 10 m*M* K-cacodylate, 1 m*M* phenylmethylsulfonyl fluoride (PMSF), and 80–100 m*M* $MnCl_2$, pH 6.0.
10. Crystallization end buffer for NCP: 50 m*M* KCl, 10 m*M* K-cacodylate, 1 m*M* PMSF, and 40 m*M* $MnCl_2$, pH 6.0.
11. Crystal stabilization starting buffer for NCP: 30 m*M* KCl, 30 m*M* $MnCl_2$, 10 m*M* K-cacodylate, pH 6.0, 2% 2-methyl-2,4-pentanediol (MPD).
12. Crystal stabilization stock buffer for NCP: 30 m*M* KCl, 30 m*M* $MnCl_2$, 10 m*M* K-cacodylate, pH 6.0, and 50% MPD.
13. Crystal stabilization ending and cryoprotection buffer for NCP: 30 m*M* KCl, 30 m*M* $MnCl_2$, 10 m*M* K-cacodylate, pH 6.0, and 22.5% MPD.
14. XI precipitant material: 50% (w/v) $(NH_4)_2SO_4$.
15. XI solubilizing buffer: 25 m*M* Tris-HCl, pH 8.0.
16. XI metal stripping buffer: 25 m*M* Tris-HCl and 10 m*M* EDTA, pH 8.0.
17. XI crystallization buffer: 100 m*M* Tris-HCl and 2 m*M* $MgCl_2$, pH 8.0.
18. XI "adequate" cryoprotection buffer: 100 m*M* Tris-HCl, 2 m*M* $MgCl_2$, 30% (w/v) $(NH_4)_2SO_4$, serial passage through solutions of this material with 5–30% glycerol (w/v).
19. XI "inadequate" cryoprotection buffer: 100 m*M* Tris-HCl, 6 m*M* $MgCl_2$, 1.5 *M* xylose, and 40% (w/v) $(NH_4)_2SO_4$.
20. Leukemia inhibitor factor (LIF)/gp130 crystallization buffer: 0.1 *M* imidazole, 0.2 *M* sodium iodide, pH 7.5, 8–10% polyethylene glycol 3350.
21. LIF/gp130 cryoprotection buffer: 0.1 *M* imidazole, 0.2 *M* sodium iodide, pH 7.5, 8–10% polyethylene glycol 3350, serial passage through solutions of this material with 5 and 20% glycerol (w/v).

3. Methods

Detailed next are three methods for annealing crystals that resulted in two successes and one failure. The unsuccessful outcome is included to assist in determining the conditions when any annealing protocol will not result in improved diffraction characteristics (*see* **Note 4**). In general, the process for initially flash cooling a crystal is irrelevant to the outcome of crystal annealing. MCA was developed using a cold nitrogen gas stream ***(1)***, however, other flash-cooling techniques such as plunging into cryogenic liquids also can be used.

3.1. The Nucleosome Core Particle

MCA was developed with the NCP, a 210-kDa macromolecule of approximately equal masses of DNA and protein ***(6)***. Crystals of this macromolecule (space group $P2_12_12_1$) are typically grown in buffer containing 50 m*M* KCl, 10 m*M* K-cacodylate, and 1 m*M* PMSF. Crystallization is induced by lowering the concentration of $MnCl_2$ from around 80–100 m*M* to near 40 m*M* $MnCl_2$. Crystals are harvested into stabilization buffer containing 30 m*M* KCl, 30 m*M* $MnCl_2$, 10 m*M* K-cacodylate, pH 6.0, and 2% MPD. The concentration of MPD

is increased in steps by addition of buffer containing 30 m*M* KCl, 30 m*M* $MnCl_2$, 10 m*M* K-cacodylate, pH 6.0, and 50% MPD. The amount of buffer containing 50% MPD added at each step is adjusted to provide a 2% increase in MPD concentration and a minimum of 12 h allowed for equilibration of the crystals before the next MPD addition. The concentration of MPD for data measurement is 22.5%, adjusted using refractometry to verify the final value. Addition of MPD to the crystals affects the *c*-axis length of the unit cell as well as the resolution limit. The target concentration of MPD (22.5%) provides the highest resolution limit for diffraction data. Because the MPD used in the unit cell condensation process for optimal diffraction does not provide optimal cryoprotection, crystal diffraction from NCP after initial flash cooling in a cold nitrogen stream is often unacceptable.

If there is a problem with the crystal after the initial flash cooling, the crystal is removed from the cold gas stream and quickly transferred into a 0.3-mL droplet of cryoprotectant (for NCP this is a buffer containing 30 m*M* KCl, 30 m*M* $MnCl_2$, 10 m*M* K-cacodylate, pH 6.0 with 22.5% MPD) at the crystal growth temperature on a siliconized glass well slide or depression plate. This slide or plate is placed as close as is practical to the crystal in the cold gas stream. After removal, the crystal is allowed to drift off the loop to minimize mechanical stresses on the crystal. Once the crystal is off the loop, the drop is covered using a siliconized cover slip to prevent evaporation or hydration of the cryoprotectant solution. Similar care should be taken with the incubation solution prior to crystal placement in it. A hydrophobic surface is important to hold the incubation solution to the well and to keep the crystal from migrating with the cover slip when it is lifted at the end of the incubation period. The crystal is allowed to warm in the incubation solution for 3 min after which time the cover slip is removed, the crystal is remounted on a loop, and flash cooled again. Images of the initial diffraction pattern and the diffraction from the annealed NCP crystal are shown in **Fig. 2**. MCA experiments with different proteins and incubation times established the value of 3 min for incubation. Longer duration did not improve the outcome. Reduction of the incubation time to about 2 min or less produced results that were inferior to the 3-min incubation. However, exceptions abound. In one case, the crystal was soaked in 10 µL of cryoprotectant solution for 10 s with outstanding results *(7)*.

3.2. D-Xylose Isomerase

The 43-kDa metalloenzyme D-xylose isomerase (XI) catalyzes the interconversion of D-xylose to D-xyulose and D-glucose to D-fructose by transferring a hydrogen atom between C1 and C2 of the substrate. The pH of optimal activity is approx 7.8. In the I222 form, the crystal solvent content is approx 50%,

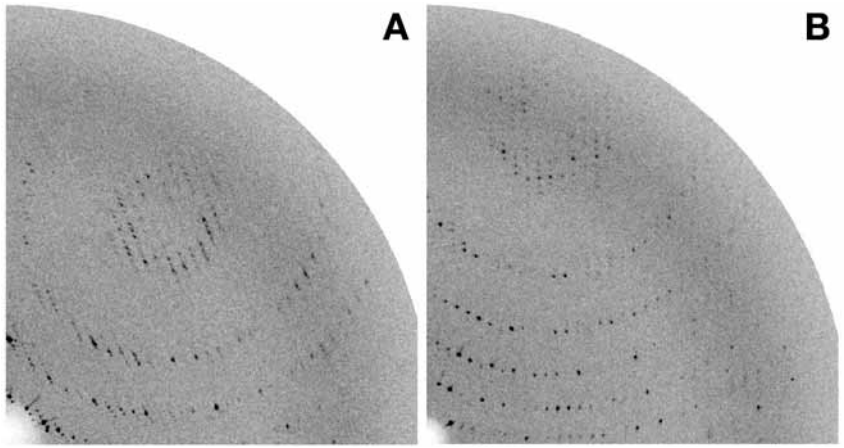


Fig. 2. A comparison of diffraction images from a nucleosome core particle crystal illustrating the improvement in mosaicity observed after macromolecular crystal annealing. **(A)** Shows typical diffraction after flash cooling. A significant improvement in diffraction characteristics is observed after annealing in image **(B)**.

and it readily forms well-ordered crystals that diffract X-rays to ultra-high resolution (<1 Å).

XI was purified from a commercial enzyme formulation of a genetically modified strain of *Streptomyces rubiginosus* (Gensweet SGI) that was kindly provided by Genencor International (Rochester, NY). The starting material consisted of a 2-mL aliquot containing approx 1 g of protein. The protein was separated from inhibitor (sorbitol) by precipitation in 50% (w/v) ammonium sulfate, followed by repeated washing of the precipitate with 50% (w/v) solution of ammonium sulfate. The purification of the sample can be followed by monitoring the color as the precipitate changes from brown to light tan. The precipitate was solubilized in 25 m*M* Tris, pH 8.0 and dialyzed against 25 m*M* Tris and 10 m*M* EDTA, pH 8.0. After removing the EDTA by dialysis, the sample was brought to crystallization trial conditions in 100 m*M* Tris, 2 m*M* $MgCl_2$, pH 8.0, and concentrated to approx 100 mg/mL.

XI can be crystallized by several techniques, ranging from batch to hanging-drop vapor diffusion experiments with ammonium sulfate as a precipitant in concentrations ranging from 10 to 45% (w/v). Cryoprotection methods for XI have been reported in **ref.** *8*. In short, this involves the serial passage of the crystal through wells of mother liquor with increasing concentrations (5–30% [w/v]) of glycerol. Diffraction to the edge of the detector is seen in these cryoprotected crystals (**Fig. 3A**). It has since been determined that glycerol enters

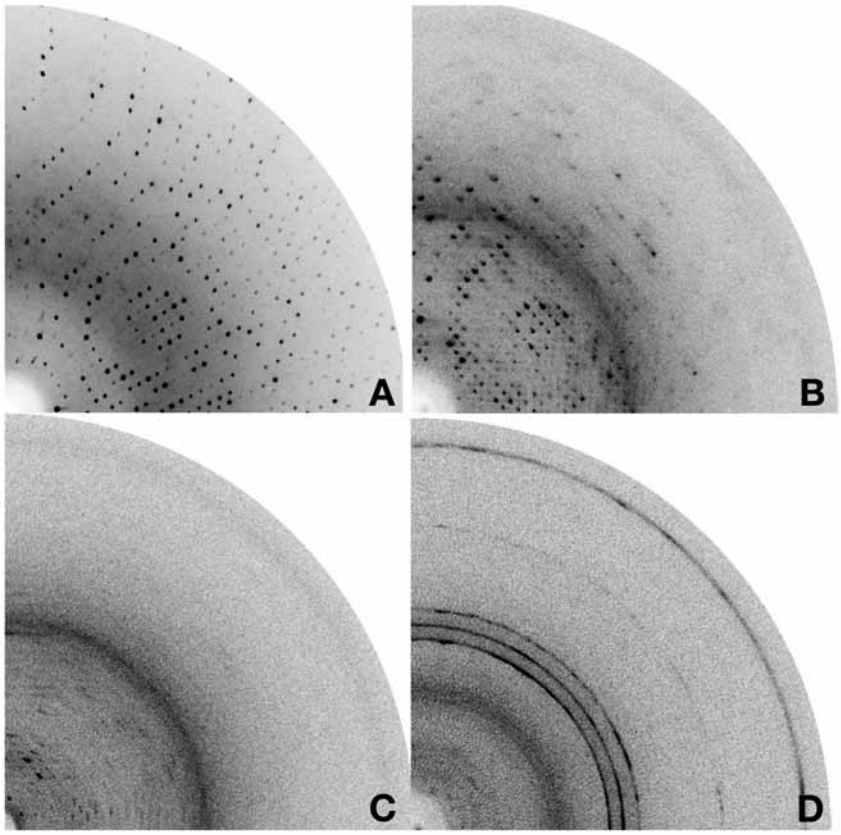


Fig. 3. Diffraction images from crystals of D-xylose isomerase. **(A)** Diffraction from an XI crystal that was stable in cryoprotectant showing diffraction to the edge of the detector. **(B)** Diffraction from a poorly cryoprotected XI crystal. Diffraction such as this is usually a good indication for annealing. **(C)** Diffraction after macromolecular crystal annealing. The diffraction has worsened indicating that crystal lattice is perturbed, most likely owing to crystal instability in the cryoprotectant. **(D)** Similar crystal as **B** after *in situ* annealing. In addition to crystal instability in the cryoprotectant as in **C**, ice rings have appeared, perhaps due to high local humidity in the room.

the active site of the enzyme. Consequently, a series of studies were undertaken to assess alternative cryoprotectants for XI.

A variety of similarly formulated cryoprotectant solutions were tested, one of which was composed of 100 m*M* Tris, 6 m*M* $MgCl_2$, 1.5 *M* xylose, and 40% (w/v) ammonium sulfate. This cryoprotectant was first tested to determine if it flash cooled into a vitreous state, which was found to be the case. Next, the

cryoprotectant was used to flash cool crystals of XI. For each attempt, the crystal was dragged through the cryoprotectant solution and flash cooled in the <100K nitrogen gas stream. This method was used because leaving a crystal in the cryoprotectant for more than 30–60 s resulted in its dissolution. By all physical appearance the crystal flash cooled well; however, the diffraction (**Fig. 3B**) was found to be deficient. A cycle of MCA was tried, and the annealed crystal flash cooled. The diffraction became worse (**Fig. 3C**). Additional crystals were tried with similar diffraction results following the initial flash cooling. *In situ* annealing was also attempted after flash cooling diffraction was no longer apparent (**Fig. 3D**).

These results mirror similar outcomes with Concanavalin A (con A) described briefly in **ref.** *2*. Con A is initially stable in its cryoprotectant buffer, flash cools well, and subsequently diffracts well. However, diffraction is severely degraded or lost after MCA or *in situ* annealing. In both cases, the failure of annealing is traced to instability of the protein crystals in their cryoprotectant buffers. In the case of annealing con A using MCA, degradation of the protein crystal was evident after 30 s, with cracks and pitting of the crystal the most obvious sign of the instability. Similarly, crystals of XI could not be soaked for any length of time in cryoprotectant solutions at room temperature based on xylose. These examples reinforce the importance of making an effort to develop a cryoprotectant solution that not only vitrifies when flash cooled but is also one in which the macromolecule is stable. Annealing is not a substitute for a poor cryoprotectant (*see* **Note 2**).

3.3. LIF/gp130

Very thin, fragile crystals sometimes do not survive transfer to and from the MCA incubation buffer. The only practical annealing is an *in situ* method (*see* **Note 5**). Our preference for *in situ* annealing calls for the cold gas stream to be diverted (not blocked, because restricting the flow might increase the temperature of the gas or shut down the cryogenics apparatus) until the crystal warms completely to room temperature as indicated by the melting of the frost that forms on the crystal. The time it takes for the crystal to warm completely varies with the volume of the crystal and the amount of solution present. The cold gas stream diverter is then removed and the crystal is flash cooled again. The best results in our laboratory have been seen by applying the procedure once, not multiple times (*see* **Notes 3** and **6**).

The structure of the cytokine LIF and the shared signaling receptor gp130 is an example of *in situ* annealing with a very thin crystal, where without annealing the project would not be completed *(9)*. Crystals of this complex were grown in 0.1 *M* imidazole, 0.2 *M* sodium iodide, pH 7.5, with 8–10% polyethylene glycol 3350 as a precipitant. The molecular weight of the complex is

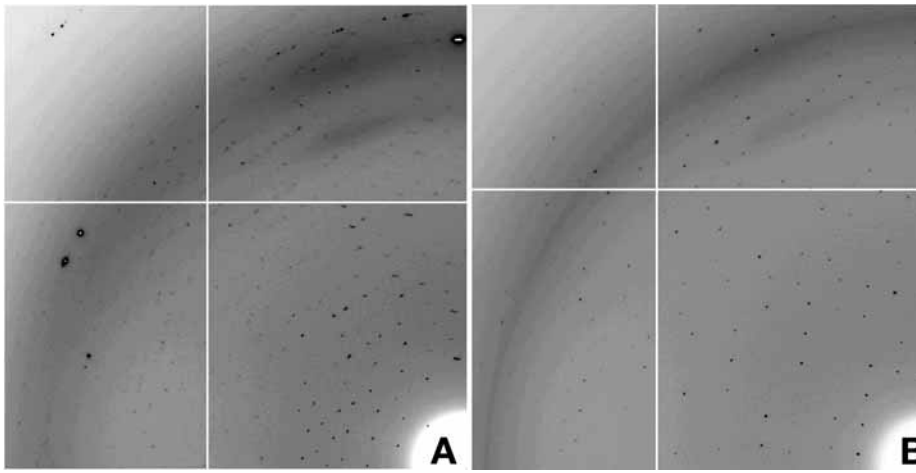


Fig. 4. Diffraction from a small crystal of the LIF/gp130 complex. **(A)** Initial diffraction pattern from the crystal. The reflections show significant smearing. **(B)** After one cycle of *in situ* annealing, the mosaicity is reduced and the diffraction data can be measured. (Images courtesy of Martin Boulanger.)

approx 42 kDa with two complexes in the asymmetric unit. The crystals (space group $P2_12_12_1$) were small plates of dimensions $0.05 \times 0.01 \times 0.01$ mm, with a solvent content of 50–55%. Cryoprotection was provided by serial transfer (5 s in each passage) of the crystal through mother liquor containing 5 and 20% glycerol (w/v). To obtain quality diffraction images that could be processed, the crystals required annealing; however, removing the crystals from the cryostream and putting them back in mother liquor destroyed the crystals. *In situ* annealing was performed by diverting the cryostream until the crystal completely thawed (~1 s) and then flash cooling the crystal again. The annealing resulted in sharpened diffraction spots and the measured data could be processed as is seen in **Fig. 4**. Similar results with thin crystals of the F41 fragment of flagellin have been reported ***(10)***.

4. Notes

1. Good cryocrystallographic practice is essential for successful annealing. Factors that affect the crystal-cooling rate from most to least important are detailed in **ref.** ***11***. These factors are (1) crystal solvent content, (2) cryoprotectant concentration, (3) crystal size and shape, and the amount of residual liquid around the crystal, (4) cooling method (liquid plunge vs gas stream), (5) choice of cryogen (nitrogen, helium, propane), and (6) relative speed between cooling fluid and the crystal surface. Annealing the crystal will often address problems with mosaicity that are a function of the crystal solvent content and cryoprotectant concentration, especially

when the cryoprotection is marginal. To minimize the effects of factors five and six, the loop should be precentered in the cold gas stream and as close to the nozzle as possible. The cold gas should be at the maximum flow rate; typically 10 L/min for N_2 cold streams. After the crystal has been flash cooled, the nozzle can be moved away for data measurement so as to not shadow the detector.

2. To perform annealing the crystal must be stable in cryoprotectant whether the cryoprotectant is a hydrocarbon or silicone-based oil, an organic additive, or a saturated salt solution. *See* **ref.** ***12*** or Chapter 1 for a discussion of functional cryoprotection for macromolecular crystals.
3. A body of anecdotal reports, plus the published account ***(3)***, has shown that *in situ* annealing seems particularly successful at facilities in relatively arid regions. The relative humidity of a specific locale may affect the composition of the cryoprotectant solution around a crystal during *in situ* annealing because these solutions are usually hygroscopic. Annealing experiments that produce "popsicles" of frost in our home laboratory can be frost free when performed in arid climates or in midwinter in heated facilities. Relative humidity can be a difficult variable to control. The MCA technique should minimize the impact of relative humidity during annealing.
4. Annealing techniques should not be expected to improve diffraction resolution. A distinction must be made between real and functional resolution improvements. Real resolution improvements are the result of increased molecular order within the crystal mosaic; functional improvements are the result of increased ordering of the mosaic blocks in the crystal. Resolution improvements after annealing are, in many cases, a consequence of diffraction peak sharpening. In this case, integration of a reflection may result in similar counts before and after annealing, but after annealing, a particular reflection may have a higher signal-to-noise ratio and emerge from the background because the photon counts are concentrated in a smaller area of the detector. The net effect from the annealed crystal is increased resolution; however, the molecular order within the crystal has not changed.
5. If a crystal can be flash cooled and is stable in cryoprotectant, the size of the crystal is irrelevant for MCA; successful annealing has been performed on crystals of XI that were used in neutron diffraction experiments (1.3 mm/edge). However, as seen in **Subheading 3.3.**, very thin crystals that would likely be physically damaged during transfer might best be annealed *in situ*. When one is annealing a crystal on the loop it is important to allow the crystal to thaw completely before flash cooling.
6. MCA will produce a maximum reduction in mosaicity after one cycle (with a 3-min incubation period). Multiple cycles are unnecessary. Multiple cycles of *in situ* annealing may be required to approach the mosaicity improvement seen after one cycle of MCA. However, multiple *in situ* annealing cycles can change the solvent composition of a crystal (hydration or dehydration), resulting in unpredictable results.

Acknowledgments

Research sponsored by grants from National Institutes of Health (GM-29818), NASA (NAG8-1568 and NAG8-1826), the Office of Biological and

Environmental Research, US Department of Energy and the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy under contract no. DE-AC05-00OR22725.

References

1. Harp, J. M., Timm, D. E., and Bunick G. J. (1998) Macromolecular crystal annealing: overcoming increased mosaicity associated with cryocrystallography. *Acta Cryst.* **D54,** 622–628.
2. Harp, J. M., Hanson, B. L., Timm, D. E., and Bunick, G. J. (1999) Macromolecular crystal annealing: evaluation of techniques and variables. *Acta Cryst.* **D55,** 1329–1334.
3. Yeh, J. L. and Hol, W. G. (1998) A flash-annealing technique to improve diffraction limits and lower mosaicity in crystals of glycerol kinase. *Acta Cryst.* **D54,** 479–480.
4. Stevenson, C. E. M., Mayer, S. M., Delabre L., and Lawson, D. M. (2001) Crystal annealing – nothing to lose. *J. Cryst. Growth* **232,** 629–637.
5. Weik, M., Kryger, G., Schreurs, A. M., et al. (2001) Solvent behaviour in flash-cooled protein crystals at cryogenic temperatures. *Acta Cryst.* **D57,** 566–573.
6. Harp, J. M., Hanson, B. L., Timm, D. E., and Bunick, G. J. (2000) Asymmetries in the nucleosome core particle at 2.5 Å resolution. *Acta Cryst.* **D56,** 1513–1534.
7. Timm, D. E., Mueller, H., Harp, J. M., and Bunick, G. J. (1999) Crystal structure and mechanism of a carbon-carbon bond hydrolase. *Structure* **7,** 1023–1033.
8. Hanson, B. L., Harp, J. M., Kirschbaum, K., et al. (2002) Experiments testing the abatement of radiation damage in D-xylose isomerase crystals with cryogenic helium. *J. Synchrotron Rad.* **9,** 375–381.
9. Boulanger, M. J., Bankovich, A. J., Kortemme, T., Baker, D. and Garcia, K. C. (2003) Convergent mechanisms for recognition of divergent cytokines by the shared signaling receptor gp130. *Mol. Cell* **12,** 577–589.
10. Samatey, F. A., Imada, K., Vonderviszt, F., Shirakihara, Y., and Namba, K. (2000) Crystallization of the F41 fragment of flagellin and data collection from extremely thin crystals. *J. Struct. Biol.* **132,** 106–111.
11. Kriminski, S., Kazmierczak, M. and Thorne, R. E. (2003) Heat transfer from protein crystals: implications for flash-cooling and X-ray beam heating. *Acta Cryst.* **D59,** 697–708.
12. Garman, E. F. and Doublié, S. (2003) Cryocooling of macromolecular crystals: optimization methods. *Meth. Enzymol*. **368,** 188–216.

4

First Analysis of Macromolecular Crystals

Biochemistry and X-Ray Diffraction

David Jeruzalmi

Summary

Methods are described for performing a first analysis of newly prepared macromolecular crystals. Biochemical analysis using denaturing gel electrophoresis on dissolved crystals informs on the content of the crystal. Detailed protocols are given for mounting crystals in capillaries or fiber loops in preparation for analysis of their X-ray diffractive properties.

Key Words: Macromolecular crystals; biochemical analysis of crystal contents; crystal mounting; X-ray diffraction; crystal symmetry; space groups.

1. Introduction

A proper first analysis of newly prepared macromolecular crystal will determine the future course of crystal structure determination by X-ray crystallography. The questions of interest to the investigator are: does the crystal contain a macromolecule or an undesired small molecule? What is the state of the macromolecule in the crystal? What is the symmetry of the crystal? What are the lattice constants of the crystal? What is in the asymmetric unit? To what scattering angle do the crystals diffract X-rays? What is the quality of the integrated intensities that may be recorded? What is the stability of the crystals in the X-ray beam? Can a formulation be found that enables preservation of the crystal to shock cooling? Successfully addressing these questions will determine the suitability of a crystalline specimen for structural analysis.

A first analysis of a newly prepared macromolecular crystal consists of:

1. Biochemical analysis of the crystal.
2. Mounting of the crystal for initial measurements.
3. Measurements of the X-ray diffractive properties of the crystals.

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

Table 1
Crystal Symmetry Elements

Type	Description
Rotation axis	Rotation of 360°/*n* about axis of rotation, where *n* = 1, 2, 3, 4, 6; for example a twofold axis involves a rotation of 180°. This element is common in crystals of biological macromolecules.
Mirror plane	Reflection through a plane. This element is not found in crystals of biological macromolecules.
Screw axis	Similar to a rotation axis but includes a fractional translation along the rotation axis. For example a 4_1 screw axis involves a 90° rotation followed by a translation of one-fourth along the axis of rotation. This element is common in crystals of biological macromolecules.
Glide plane	Reflection through a plane followed by a translation of half the unit cell in the direction parallel to the plane. This element is not found in crystals of biological macromolecules.

4. Software analysis of the recorded X-ray images to determine crystallographic and orientational parameters.
5. Calculation of the Matthews parameter ***(1)***, and examination of the native Patterson and self-rotation functions to set the stage for a full crystallographic analysis.

1.1. Lattices and Symmetry

A crystal is a periodic array of atoms or molecules, where each element of the smallest repeating unit makes the same kind of interaction with its neighbors. The smallest repeating unit of a crystal is termed the asymmetric unit. Asymmetric units can be arranged into a unit cell through application of 230 different combinations of symmetry elements (**Table 1**). However, because this chapter is dedicated to the analysis of crystals comprised of biomolecules, which are usually chiral entities, the number of combinations of symmetry elements (or space groups) that can be observed is reduced from 230 to 65 (**Table 2**). A unit cell of crystal is defined as a parallelepiped (**Fig. 1**) described by six values that define the lengths of its edges (a, b, and c) and the angles between the edges (α, β, and γ). Operationally, a lattice can be assembled by repeatedly depositing unit cells in the three directions defined by the angles α, β, and γ. Each unit cell in the lattice is separated by intervals that are equal to a, b, and c. The parameters a, b, c and α, β, and γ are designated crystallographic lattice constants.

Crystal symmetries with common elements can be grouped into seven possible classes or systems (**Table 3**). For example, the triclinic system possesses no

Table 2
The 65 Space Groups That are Allowed for Chiral Macromolecular Crystals

Crystal system	Diffraction symmetry	Space groups
Triclinic	1	P1
Monoclinic	2/m	P2, $P2_1$, C2
Orthorhombic	mmm	P222, $P222_1$, $P2_12_12$, $P2_12_12_1$, C222, $C222_1$ F222, I222, $I2_12_12_1$
Tetragonal	4/m	P4, $P4_1$, $P4_2$, $P4_3$, I4, $I4_1$
4/mmm		P422, $P4_122$, $P4_222$, $P4_322$, $P42_12$, $P4_12_12$, $P4_22_12$, $P4_32_12$, I422, $I4_122$
Trigonal	3	P3, $P3_1$, $P3_2$, R3
	3/m	P321, P312, $P3_121$, $P3_221$, $P3_112$, $P3_212$, R32
Hexagonal	6/m	P6, $P6_5$, $P6_4$, $P6_3$, $P6_2$, $P6_1$
	6/mmm	P622, $P6_122$, $P6_222$, $P6_322$, $P6_422$, $P6_522$
Cubic	m3	P23, $P2_13$, F23, I23, $I2_13$
	m3m	P432, $P4_132$, $P4_232$, $P4_332$, F432, $F4_132$, I432, $I4_132$

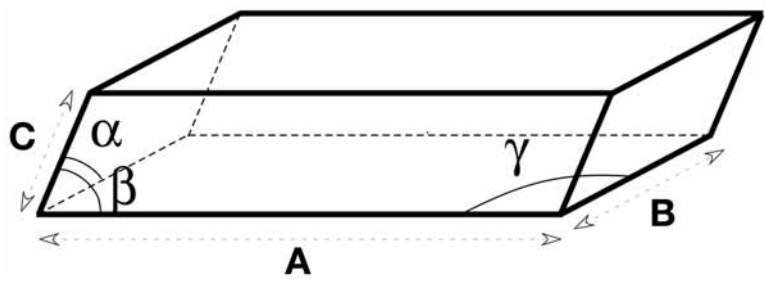


Fig. 1. Parameters that describe a crystallographic unit cell. The crystallographic unit cell is a parallelepiped, which is specified by six values. These are the three lengths (a, b, c) of the cell and the three angles (α, β, γ) between them. The symmetry exhibited by the unit cell can impose restrictions on the values of these parameters (**Table 3**).

symmetry axes and the unit cell edges and angles may adopt any value. A crystal in the monoclinic class is characterized by the presence of a twofold axis (by convention) along the *b*-axis. The β angle may adopt any value, but presence of the twofold axis restricts α and γ to a value of 90°; the cell edges may adopt any

Table 3
The Seven Crystal Systems

System	Bravais lattices	Symmetry	Restrictions on cell parameters
Triclinic	P	None	$a \neq b \neq c$, $\alpha \neq \beta \neq \gamma$
Monoclinic	P, C	Twofold axis (parallel to b)	$a \neq b \neq c$, $\alpha = \gamma = 90°$; $\beta \neq 90°$
Orthorhombic	P, C, I, F	Three perpendicular twofold axes	$a \neq b \neq c$, $\alpha = \beta = \gamma = 90°$
Tetragonal	P, I	Fourfold axis (parallel to c)	$a = b \neq c$, $\alpha = \beta = \gamma = 90°$
Trigonal/	P	Threefold axis (parallel to c)	$a = b \neq c$, $\alpha = \beta = 90°$; $\gamma = 120°$
Rhombohedral	R		$a = b = c$, $\alpha = \beta = \gamma \neq 90°$
Hexagonal	P	Sixfold axis (parallel to c)	$a = b \neq c$, $\alpha = \beta = 90°$; $\gamma = 120°$
Cubic	P, I, F	Four threefold axes along diagonal of cube	$a = b = c$, $\alpha = \beta = \gamma = 90°$

value. A more extreme example of symmetry is seen in the cubic system where the cell edges adopt the same value ($a = b = c$) and the angles between them are restricted to 90°. Implicit in the discussion of crystal systems is the assumption that a unit cell is drawn to contain one lattice point at each of its eight vertices. This type of lattice is referred to as a primitive lattice. It is often useful to describe a unit cell that contains additional lattice points that are not at the vertices. A cell with an additional lattice point in the center is designated as body centered (I). A cell that has an additional lattice point on one of its six faces is referred to as a C-centered cell (C). Presence of lattice points on all or each of the six faces of the unit cell is called a face-centered cell (F). Consideration of nonprimitive cells along with the seven crystal systems gives rise to 14 possible lattice arrangements (**Fig. 2**) that are named after Auguste Bravais, who was the first to classify them as such. **Table 2** summarizes the seven crystal systems and the restrictions that they impose on the crystallographic cell.

The precise arrangement of the lattice of unit cells endows the crystal with the power to diffract X-radiation. Diffraction is observed when waves are scattered by a periodic array that is arranged with the appropriate spacing to fulfill Bragg's law ($n\lambda = 2d\sin\theta$). Bragg's law is satisfied when a lattice with a characteristic spacing (d) scatters incident radiation of an integral (n) number of wavelengths (λ), which interfere with each other constructively and destructively to give rise

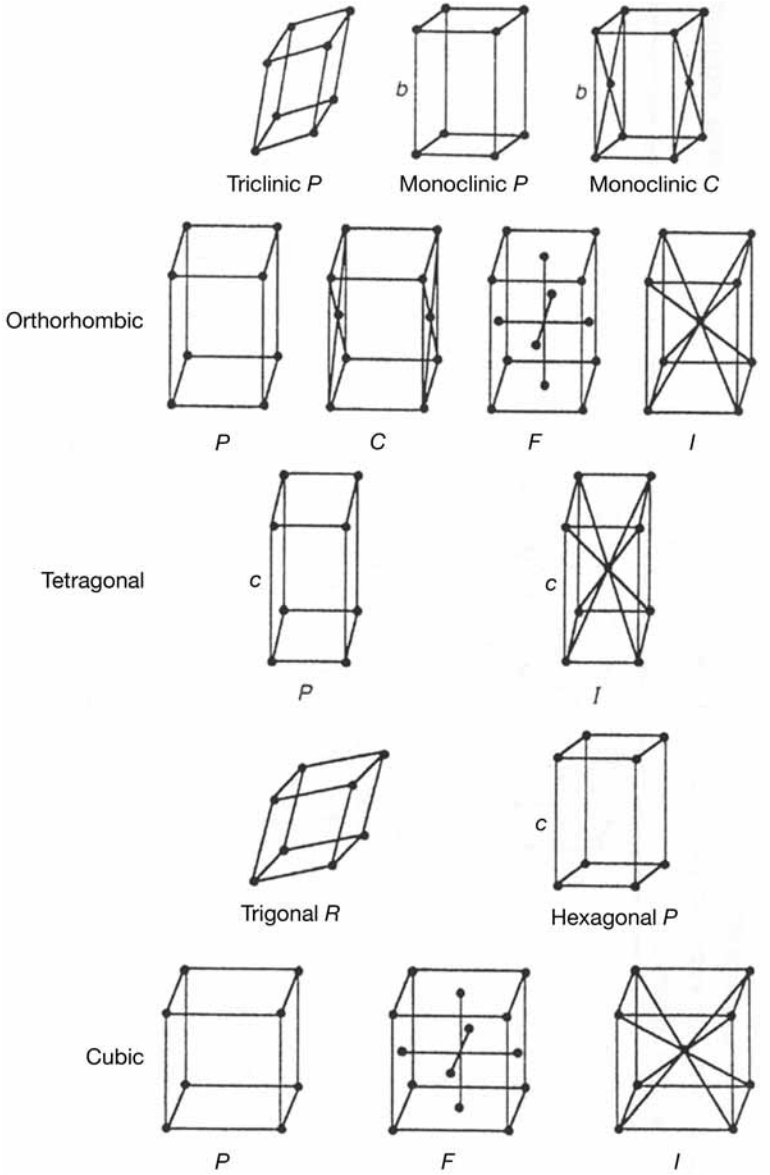


Fig. 2. The 14 Bravais Lattices.

to a diffracted wave. A diffracted wave is observed at a characteristic scattering angle (θ) relative to the source. Bragg's formalism likened diffraction to reflection of light by a plane mirror. In this view, diffraction arises from a set of imaginary planes that are parallel, equally spaced (d), and carries equivalent slices of

the atoms that form the structure. Low-angle diffraction arises from planes with large spacings and carry less information about the underlying structure (in practical terms, the equivalent of performing measurements with a ruler that is subdivided into centimeters) than higher angle diffraction, which comes about from Bragg planes with a smaller spacing (e.g., measurements performed with a ruler that is subdivided into millimeters). Equivalency of unit cells over the volume of the crystal, therefore, determines the scattering angle of the resultant diffraction and ultimately the information content that can be extracted. The higher the angle of the observed diffraction, the smaller the distance between the Bragg planes that give rise to it. Thus, a lower quality crystal, whose unit cells are arranged less perfectly over its volume, will only satisfy Bragg's law at relatively large planar spacings and, thus, only yield observable diffraction at a lower scattering angle. A high-quality crystalline sample will fulfill diffraction conditions at low and higher angles and provide higher resolution information about the underlying structure.

A first analysis of a newly prepared macromolecular crystal will focus on (1) discovery of the fundamental crystallographic constants (a, b, c, α, β and γ), (2) description of the symmetry that describes the diffraction (space group), and (3) estimation of the quality of the integrated diffraction that may be recorded.

2. Materials

2.1. Hardware

2.1.1. Biochemical Analysis of the Crystal

1. Sodium dodecyl sulfate (SDS) gel electrophoresis equipment or access to a liquid chromatographic system (e.g., high-performance liquid chromatography [HPLC], fast protein liquid chromatography [FPLC]); mass spectrometric services.
2. Fiber loop (Hampton Research, Aliso Viejo, CA).
3. Crystal-handling tools.
4. Dissecting microscope.
5. 250 μL of reservoir or crystal stabilization solution.
6. 4X SDS-gel loading buffer: 200 m*M* Tris-HCl (pH 6.8), 400 m*M* dithiothreitol, 8% SDS, 0.4% bromophenol blue, and 40% glycerol.

2.1.2. Crystal Mounting (Capillary)

1. X-ray diffraction-quality capillaries (cleaned by immersion in distilled water followed by ethanol). Some groups also siliconize capillaries.
2. Dissecting microscope.
3. Wax.
4. Source of heat (flame or soldering iron).
5. Rubber tubing (appropriate size for attachment to a capillary).

6. 1-mL Plastic syringe.
7. Crystal-handling tools (Hampton Research or other).
8. Scalpel.
9. Paper wick.
10. Reservoir or crystal stabilization solution.
11. High-vacuum grease.
12. Fine-tip forceps.
13. Diamond-tip pen.
14. Plasticine clay.
15. Goniometer head.
16. Brass-mounting pin.
17. Drawn-out Pasteur pipet (these can be made in the lab by heating a Pasteur pipet over a hot flame and pulling them to a fine bore).
18. Spindle stage (p/n 7058; Charles Supper Company, Natick, MA).

2.1.3. Crystal Mounting (Loop) and Shock Cooling

1. Fiber loops.
2. Loop-mounting pins.
3. Cryosolvents or oils.
4. Shallow low-form dewar (stainless steel, 1000 mL to hold the sample stage) flask.
5. Tall-form dewar flask.
6. Sample stage.
7. Propane.
8. Cryovials, canes, sleeves, and a storage dewar.
9. Liquid nitrogen.
10. Magnetic crystal wand.
11. Cryotongs.
12. Cryotube clamp.
13. “Dry shipping” Dewar.
14. Lint-free tissues.
15. Hand and face protection (cryogloves, finger cots, and goggles).

2.1.4. X-Ray Analysis

1. Aligned X-ray camera/X-ray generator/detector and supporting computer interface.
2. Gaseous nitrogen cold stream.
3. Goniometer head and adjustment key.

2.2. Software

1. DENZO/SCALEPACK (http://www.hkl-xray.com/) ***(2)***.
2. d*TREK (http://www.rigakumsc.com/protein/dtrek.html) ***(3)***.
3. MOSFLM (http://www.mrc-lmb.cam.ac.uk/harry/mosflm/) ***(4)***.
4. CCP4 suite (www.ccp4.ac.uk) ***(5)***.
5. DPS (http://staff.chess.cornell.edu/~szebenyi/DPS/) ***(6)***.
6. X-ray detector/camera operations software.

3. Methods

3.1. Biochemical Analysis of Macromolecular Crystals

Biochemical analysis of the contents of a newly prepared macromolecular crystal form is the first step in a series that will result in structure determination by X-ray crystallography. The investigator seeks to answer the following questions: does the crystal contain the macromolecule(s) of interest? Or does it consist of small molecules (e.g., salt and others)? Biochemical analysis is especially important when the structure being investigated consists of multiple macromolecular components, each of which could form a crystal as an isolated species. As the methods described next are destructive, the investigator should exercise judgment in sacrificing precious crystalline samples. Biochemical analysis (or seeding experiments) may be performed on crystals subsequent to X-ray analysis. This would be appropriate in a situation where crystallization trials have yielded few crystals.

Analysis (whether for biochemical or X-ray analysis) begins with optical inspection of a newly prepared crystal under a dissecting microscope. Crystals suitable for analysis should be single, display sharp edges, and have an approximate dimensions of 50–100 μ per edge or larger (the larger, the better). Analyses can be carried with smaller crystals, but an unambiguous interpretation might tax the methods described next.

3.1.1. Protein Analysis Using SDS-Polyacrylamide Gel Electrophoresis

1. Select a macromolecular crystal (approximate dimensions 50–100 μ per edge) for analysis.
2. Harvest a macromolecular crystal from its mother liquor using a fiber loop of appropriate size.
3. Immerse the harvested crystal in a large volume (e.g., 50 μL) of reservoir solution (or some other stabilizing solution, *see* **Subheading 3.2.2.1.**). The crystal should be harvested nearly dry in order to minimize contamination by the still-dissolved macromolecular components that might confuse the analysis (*see* **Note 1**).
4. Remove still-dissolved macromolecular components by serial transfer of the crystal (using the fiber loop) successively into four identical 50-μL aliquots of mother liquor.
5. Add 17 μL 4X SDS-sample buffer to the four samples and the sample with the dissolved crystal and perform electrophoretic analysis after boiling (in some cases, crystalline samples may be refractory to dissolution and might require more drastic treatment, e.g., 6 *M* urea). The results of the gel electrophoretic analysis are visualized by staining with Coomassie brilliant blue or silver stain, according to standard protocols *(7)*.

A successful analysis should show a declining (down to undetectable) presence of the macromolecule of interest in the lanes that correspond to the crys-

tal washes, followed by the presence of each component in the sample corresponding to the dissolved crystal. In the case of analysis of crystals of a multiprotein complex, the ratio of intensities of the individual components should mirror those of the sample precrystallization; deviations might signal substoichiometric complexes. Crystals without the desired constellation of macromolecular components should be easily diagnosed by inspection (*see* **Note 2**). Crystals containing nucleic acids can be similarly analyzed using high-percentage polyacrylamide or agarose gels.

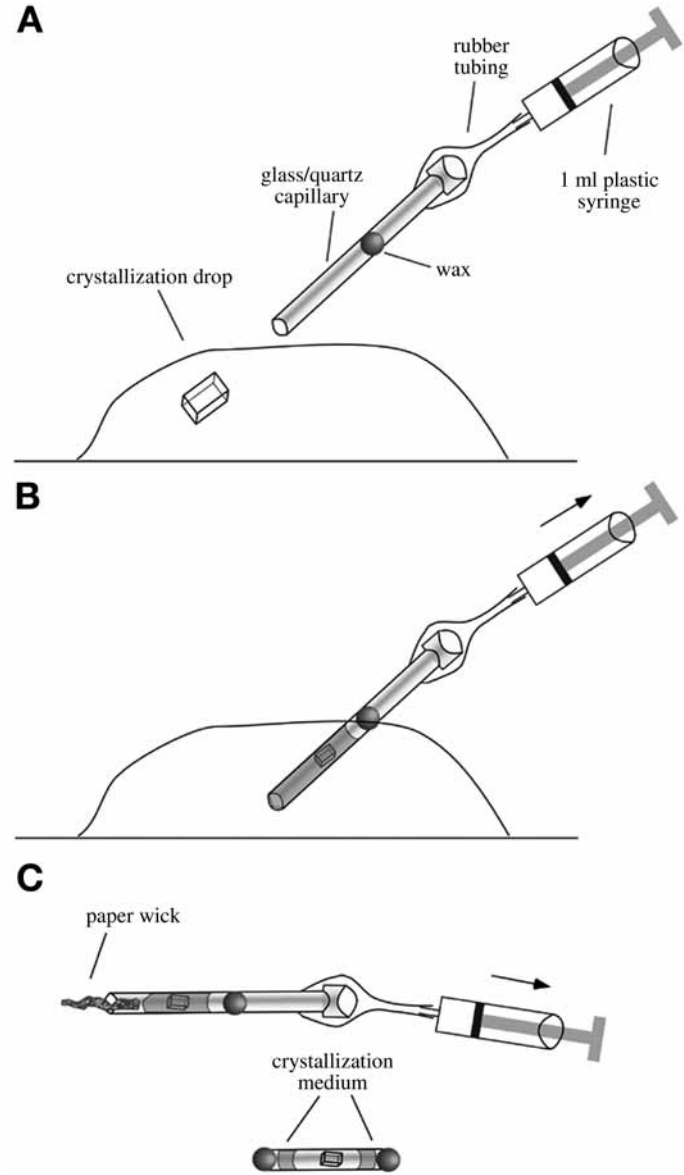
3.2. Mounting Macromolecular Crystals for X-Ray Diffraction

There are two methods to mount crystals for X-ray analysis, either in thin-walled capillaries (glass or quartz) or in fiber loops. Capillary-mounted crystals are currently only used for preliminary analyses or for recording intensity data from those samples not amenable to shock cooling. Measurements of X-ray diffraction at cryogenic temperatures are routinely performed using shock-cooled crystals that have been mounted in thin films supported by fiber loops. Both types of mounts will be described next.

3.2.1. Capillary Mount: Mounting From Natural Mother Liquor

Mounting a crystal directly from its natural mother liquor is often the preferred method to make a first observation of the X-ray diffractive properties of a newly prepared sample. Crystals taken from this environment are in their natural milieu and have not been adversely affected by contact with an inexactly formulated stabilizing solution or by the mechanical handling that accompanies crystal transfers. The X-ray diffractive properties observed from a crystal mounted from the natural mother liquor can be considered as a baseline against which to measure the efficacy of subsequent manipulations (i.e., synthetic mother liquors, cryostabilization, and others).

1. Select a glass capillary of the appropriate size for the crystal to be analyzed. A small capillary is most appropriate for a small crystal to minimize the effects of X-ray absorption during the analysis; it is, however, also more fragile. Plate-like crystals might fare better in larger diameter capillaries. Some prefer the higher mechanical strength of quartz capillaries for mounting crystals.
2. Decorate the capillary with a drop of wax as shown in **Fig. 3**.
3. Attach the capillary to a 1-mL syringe via a short segment of rubber tubing.
4. Place the crystallization drop onto the stage of a low-power dissecting microscope.
5. Dislodge the crystal, if necessary, using the appropriate handling tools.
6. Gently draw the crystal into the capillary by pulling the plunger out of the syringe. The plunger should be drawn out until the crystal has been positioned approx 1 cm from the central drop of wax. The crystal should be positioned in the capillary so that when it is mounted on a goniometer head, it will be positioned in the X-ray



beam. The exact location for the crystal in the capillary will depend on local geometry and should be verified prior to beginning the mounting procedure.

7. Using a paper wick of appropriate size, remove the mother liquor from both sides of crystal. Care should be exercised not to mechanically stress the crystal in any way. A small amount of mother liquor should be left on the syringe side of the crystal to prevent desiccation.

Fig. 3. Mounting a macromolecular crystal into a capillary from the drop. **(A)** A capillary (glass or quartz) is selected and attached to a 1-mL plastic syringe by way of a small section of rubber tubing. A bead of wax is placed on the capillary. This will be where the capillary will be broken after the mounting procedure is complete. **(B)** Under a low-power dissecting microscope, the crystal is gently drawn out of the crystallization drop and into the capillary. **(C)** A paper wick is used to gently remove the crystallization medium. **(D)** Crystallization medium should be left on either side of the crystal (or both) in order to prevent desiccation of the sample. The capillary is broken at the location of the wax bead. The open ends of the capillary are sealed with wax, vacuum grease, or epoxy glue.

8. Seal the open end of the capillary with a drop of wax, high-vacuum grease, or epoxy glue. If sealing the capillary with a bead of wax (the preferred method), the capillary should be wrapped with a moist paper towel. The open portion of the capillary should be left exposed. This precaution protects the crystal from heating owing to the application of a wax plug.
9. Use a pair of "fine-tip forceps" to gently break the capillary on the syringe side of the wax bead.
10. Seal the open end of the capillary (as a result of detachment from the syringe) with a bead of wax, high-vacuum grease, or epoxy glue.
11. Attach the capillary to the goniometer head by gently burying one end in a mound of plasticine.

3.2.2. Capillary Mount: Mounting From Synthetic Mother Liquors

3.2.2.1. Formulation of the Synthetic Mother Liquor

Formulation of synthetic mother liquor is a critical step in the X-ray analysis of macromolecular crystals. Efficient stabilizing buffers play important roles in the mounting process as well as in subsequent manipulations (e.g., heavy atom derivatization and others). Although handling of crystals in their natural mother liquor is often preferable, this can prove difficult because the solution might be close to a phase boundary and, thus, not fully stabilize the crystal to dissolution. Additionally, natural mother liquor is generally in short supply (1 mL or less from the well solution for a hanging-drop experiment). For these reasons, devising synthetic mother liquor is often a necessity. Most often, synthetic mother liquor is comprised of the components of the natural version, but with an elevated amount of precipitant. Candidate conditions can be iteratively improved by evaluating their effect on the physical appearance and X-ray diffractive properties of the sample. In difficult cases, determining an appropriate stabilizing buffer may require screening of numerous chemical conditions (salts, various polyethylene glycols, solvents, compatible solutes, pH shifts, and others) prior to success (*see* **Note 3**). Synthetic mother liquors are deemed to be stabilizing when they preserve intact (or improve) the physical (sharp edges,

extinction under polarized light) and X-ray-diffractive properties of a crystal. In all cases, synthetic mother liquor should be equilibrated to the temperature of the crystal that will be harvested into it.

3.2.2.2. Harvesting the Macromolecular Crystal

1. Select a crystal for harvesting by inspection under a dissecting microscope (this discussion will focus on crystals prepared using vapor-diffusion methods [hanging or sitting drops] as these are the most common).
2. Place the crystallization experiment on the stage of a low-power dissecting microscope.
3. In a favorable case, the crystal is single and loosely adhered to the crystallization support (glass or plastic). Alternatively, crystals may grow as a clump, from the "skin" that forms at the air–mother liquor interface, embedded in precipitated material or from foreign particulate material. In this case, use a fiber loop (or commercially available tools, glass fiber, and others) to gently dislodge a single crystal from the support. If necessary, gently separate the crystals from "skins" or precipitated material.
4. Gather the crystal with a fiber loop of similar size as the crystal and lift it out of the crystallization drop.
5. Quickly immerse the loop containing the crystal (embedded in a thin film of mother liquor) in a dish containing synthetic mother liquor equilibrated at the same temperature as the harvested specimen. Simple passage of the loop through the solution is sufficient to float the crystal off the loop.

3.2.2.3. Mounting Into a Capillary

1. Attach a glass capillary to a brass-mounting pin by first filling the pin with hot wax and then embedding the glass capillary in the wax. Place the capillary/mounting-pin assembly onto a goniometer head. Mount the goniometer head onto a spindle stage. Arrange the spindle stage in the vertical position (**Fig. 4**).
2. Use a drawn-out Pasteur pipet to fill the capillary with synthetic mother liquor. Filling should be from the bottom of the capillary starting from a point approx 2 cm from the brass-mounting pin and should proceed to the top of the capillary. The lower level of the liquid relative to the location of the brass mounting pin should be selected with care (e.g., appropriate for the specific requirements of the camera to be used) because this will the approximate final location of the mounted crystal.
3. Under a dissecting microscope, pick up the crystal from the dish with a fiber loop.
4. Deposit the crystal into the top of the capillary by brief immersion of the loop in the solution. Allow the crystal to drop (several minutes) to the meniscus at the bottom of the column of mother liquor.
5. Reposition the spindle stage to the horizontal position.
6. Under a dissecting microscope, use a drawn out Pasteur pipet to remove most of the synthetic mother liquor. A zone of liquid in the immediate vicinity of the crystal should be left untouched.
7. Use a thin paper wick to remove the remaining liquid (**Fig. 4**). The crystal should be nearly dry (*not dry*).

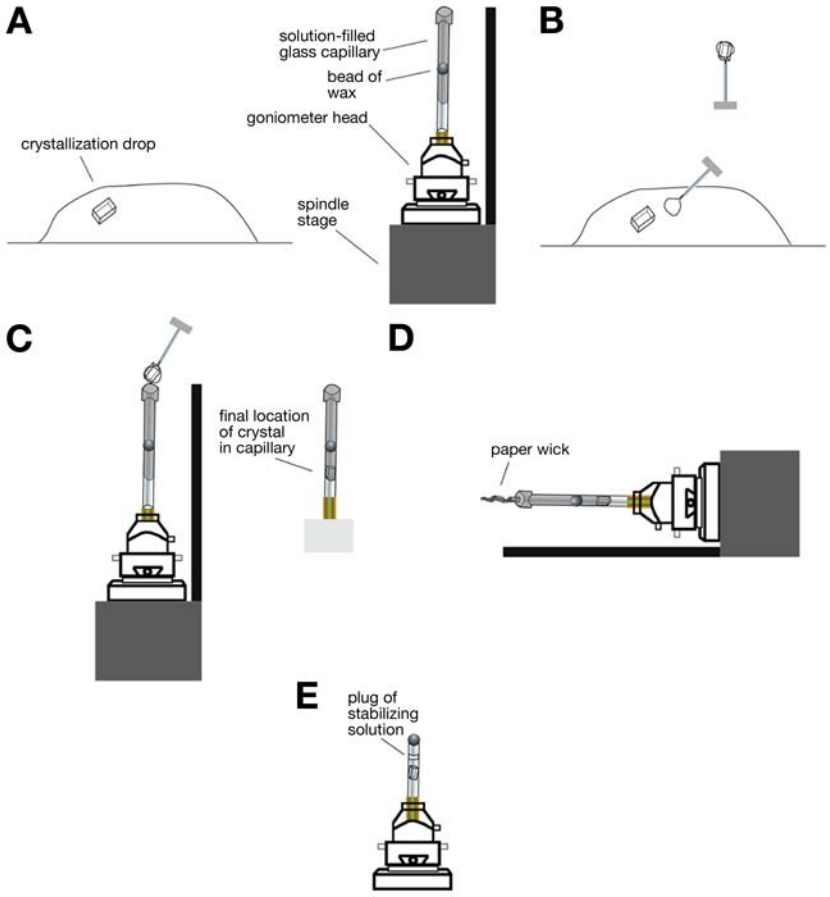


Fig. 4. Mounting a macromolecular crystal into a capillary from a stabilizing solution. **(A)** A glass capillary is attached to a brass-mounting pin and the pin is placed into a goniometer head. The goniometer head is placed on a spindle stage. The spindle stage is deployed in the vertical configuration. Using a drawn-out Pasteur pipet, the glass capillary is filled with crystal-stabilizing solution. Filling of the capillary is carried out starting at the bottom (which sets the final position of the crystal) and ending at the top of the capillary. The final position of the crystal should be selected appropriately to be an approximate match for the location of the X-ray beam. **(B)** A fiber loop is used to capture the crystal in a thin film of stabilizing solution. **(C)** The captured crystal is quickly immersed in the stabilizing solution held in the capillary. This should release the crystal, which will sink to the bottom of the column of liquid. **(D)** The spindle stage is deployed in the horizontal configuration. A paper wick is used to remove the crystal stabilizing solution from around the crystal. A small amount of liquid is left near the crystal in order to forestall desiccation during X-ray measurements. **(E)** The capillary that houses the mounted crystal is sealed with wax, vacuum grease, or epoxy glue to complete the mounting procedure.

8. Apply a plug of synthetic mother liquor into the capillary above the crystal. This volume of liquid will prevent desiccation of the sample, once the capillary has been sealed.
9. Seal the open end of the capillary with a bead of wax, a dab of grease, or epoxy glue.

3.2.3. Fiber Loop Mounts

In order to better exploit the advantages (attenuation of radiation damage) of recording X-ray diffraction data at cryogenic temperatures, Teng *(8)* proposed that crystals be mounted in thin films of solvent supported by loops. Mounting crystals for cryogenic applications requires more extensive preparation of the crystalline sample than with capillary mounts. The past decade has witnessed an explosion in the development of hardware and methodologies that has greatly transformed this method of mounting into routine practice *(9,10)*. Detailed cryocrystallography protocols are given in Chapter 1.

3.3. Assembly of the Mounted Crystal Onto the X-Ray Diffraction Camera

3.3.1. Capillary Mounts

1. Place the capillary-mounted crystal onto the goniometer head. The capillary can be affixed to the goniometer head using plasticine or through a brass-mounting pin if the crystal was mounted in a capillary with a pin.

3.3.2. Fiber Loop Mounts (Propane as Cryogen)

1. Adjust the position of the cold stream to provide more working space. The edge of the cold stream should be positioned approx 1–1.5 cm from the presumptive position of the crystal.
2. Select a vial containing a shock-cooled crystal (**Fig. 5**). Place the cryovial onto a flat surface and allow the most external layer of solid propane to thaw (ca. 10 s). This will release the loop/mounting pin from the cryovial. At this point, a substantial amount of propane will remain in solid form around the crystal. The slow-thawing properties of propane provide a favorable time window to transfer the crystal to nitrogen stream without exposing the crystal to ambient temperatures. Liquid propane should always be handled with appropriate hand and eye protection. Prolonged skin contact with liquid propane at any temperature can cause severe burns owing to evaporative freezing *(11)*.
3. While the propane immediately around the crystal is still solid, use a pair of curved forceps to lift the pin (containing the loop-mounted crystal) out of the cryovial. Place the mounting pin on the magnetic mount that is held on the goniometer head. The solid propane that remains around the crystal will slowly thaw under the cold stream, efficiently transferring the crystal from one cryogen (propane) to another (gaseous nitrogen). If the propane around the crystal is allowed to thaw outside of the cold stream, the sample may sustain damage.

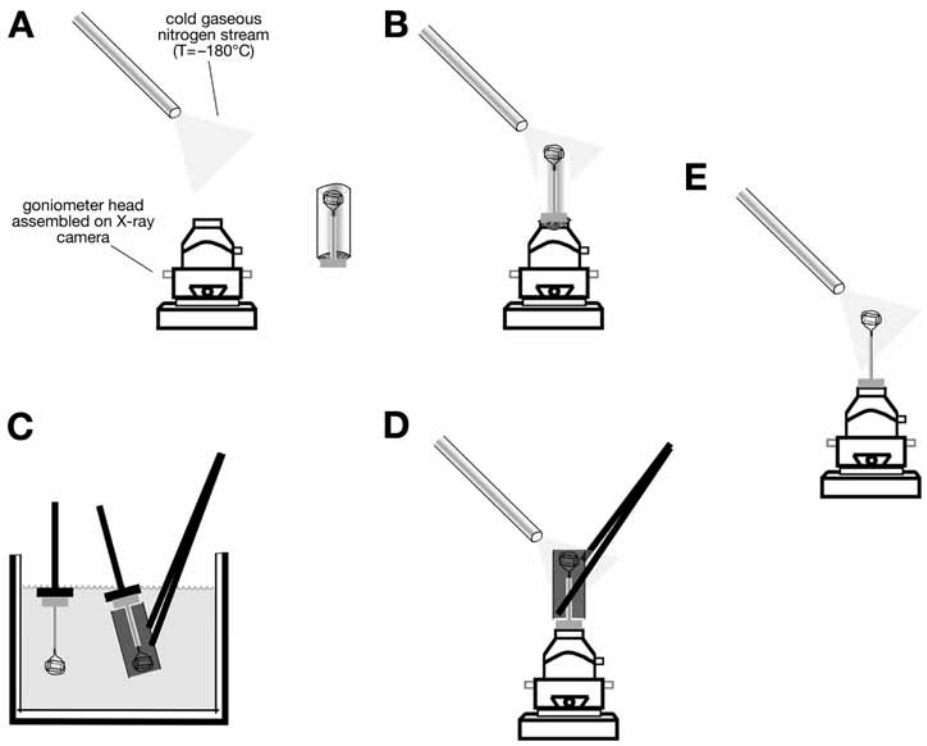


Fig. 5. Procedure for transfer of shock-cooled macromolecular crystals to the X-ray camera. **(A)** The goniometer head is assembled on the X-ray camera and the cold gaseous nitrogen stream is centered on the eucentric point of the camera. **(B)** For samples prepared with propane, a cryovial containing a crystal shock cooled in propane is selected and allowed to partially thaw (5–10 s). Once the external layer of propane has thawed, a pair of forceps is used to lift the loop-mounted crystal, still encased in solid propane, out of the cryovial and placed on the goniometer head. The solid propane is allowed to thaw completely to liquid, which falls away from the crystal. Excess liquid should be collected and disposed of properly. **(C)** For samples prepared with liquid nitrogen, a cryovial is selected and placed in a bath of liquid nitrogen (held in a shallow dewar flask). The pin containing the loop-mounted crystal is removed from the cryovial and captured with the magnetic crystal wand as shown above. The mounting pin (held by the magnetic crystal wand) is maneuvered into the head of the cryotong. With the cryotong in its locked position, lift the crystal out of the liquid nitrogen bath and position the mounting pin on the magnetic mount held on the goniometer head. Quickly unlock the cryotongs to expose the crystal to the nitrogen stream as in **E**.

4. As the propane continues to thaw, the cold stream should be brought closer to the sample. Optimally, the tip of the nozzle should be brought as close as possible to the sample without casting a shadow on the X-ray detector.
5. If necessary, wick away any remaining liquid propane from around the crystals with a lint-free tissue. Thawed liquid propane should be collected and allowed to safely evaporate in a chemical fume hood.

3.3.3. Fiber Loop Mounts (Nitrogen as Cryogen)

1. Select a vial containing a shock-cooled crystal and place it into a bath of liquid nitrogen housed in a shallow dewar flask. Hand and eye protection should always be used when handling liquid nitrogen.
2. The pin containing the loop-mounted crystal is removed from the cryovial and captured with the magnetic crystal wand.
3. Maneuver the mounting pin (held by the magnetic crystal wand) into the head of the cryotong. Encased within the head of the cryotong, the loop-mounted crystal will remain at cryogenic temperatures for up to 20 s.
4. With the cryotongs in their locked position, lift the crystal out of the liquid nitrogen bath and position the mounting pin on the magnetic mount held on the goniometer head. The time that the head of the cryotongs spends outside of a cryogen (liquid or gaseous nitrogen) should be kept to a minimum.
5. Quickly unlock the cryotongs to expose the crystal to the nitrogen stream. The crystal will thus have been efficiently transferred from liquid nitrogen to gaseous nitrogen for X-ray diffraction undamaged.

3.3.4. Recovery of Shock-Cooled Crystals

1. Recovery of crystals from the cold stream involves performing the procedure described for the mounting of crystals to the X-ray camera but with the steps reversed (**Subheading 3.3.3.**). Cool the cryotongs, cryovial, and vial clamp immersion in liquid nitrogen.
2. Open the cold cryotongs and rapidly capture the mounting pin/loop. This step should be carried out rapidly so that the crystal leaves the cold stream and enters the protected (cold) space within the cryotongs in a minimum amount of time.
3. Quickly lift the cryotongs (with enclosed crystal) off the goniometer head. Rapidly plunge the head of the cryotongs into a bath of liquid nitrogen.
4. Within the liquid nitrogen, carefully transfer the mounting pin to the magnet of the crystal wand.
5. Maneuver the mounting pin (attached to crystal wand) into a cryovial (held with the cooled vial clamp). Release the mounting pin into the cryovial.
6. Secure the recovered crystal (in the cryovial) to a cryocane and place the cryocane into a liquid nitrogen storage dewar.

3.4. Optical Alignment of a Crystal on the X-Ray Camera

1. Alignment of the axis of rotation of the goniometer with the X-ray beam is established through an optical alignment procedure. (This procedure assumes an

aligned X-ray camera with a method of visually [e.g., crosshairs] identifying the point of coincidence of the axis of rotation with the X-ray beam). Initiate the procedure by setting the rotational and translational adjustments of the goniometer head to their zero points.
2. Set the goniometer to an angle of 0°.
3. Look at the crystal through an optical microscope (or related) and use the translational adjustments of the goniometer head to bring the crystal into the crosshairs.
4. Set the goniometer at an angle of 180°. Repeat **step 3**.
5. Repeat **steps 2–4** until the crystal remains in the crosshairs after each rotation and requires no further adjustment.
6. Repeat this procedure but with goniometer angle settings of 90° and 270°.
7. Check the translational adjustments at goniometer settings of 0° and 180°. It may be necessary to iterate between translational adjustments at 0°/180° and 90°/270°. A crystal that is properly aligned at the eucentric point of the X-ray camera will rotate in place as the goniometer sweeps through 360°.
8. Set the locks on the goniometer head and the goniometer.

3.5. X-Ray Photography of the Reciprocal Lattice

X-ray photography of the reciprocal lattice of a newly prepared macromolecular crystal has two primary objectives. The investigator seeks to identify the crystallographic cell constants and determine the maximal extent of measurable data. In carrying out these analyses, information about the efficacy of the cryocooling procedure will be obtained that can provide valuable feedback to iterative efforts.

Measurement of the intensities of X-ray diffracted by crystals has evolved greatly over the past two decades with respect to the hardware used to record images and the software used to analyze the resultant images. As this is meant to be a practical guide, the focus of the sections below will deal with measuring images by the rotation/oscillation method with imaging-plate or charge-coupled device detectors and associated analysis software. A full discussion of precession photography, which played a prominent role in macromolecular crystallography, is beyond the scope of this chapter. The interested reader who wishes treatment of its theory and practice is referred to other texts ***(12)***.

3.5.1. Obtaining the First Images

The first images that are recorded will be used to assess the quality of the cryocooling, the suitability of the sample for subsequent analysis, and to adjust the camera parameters for the optimal measurement of data.

1. Set the camera parameters (oscillation sector size, crystal-to-detector distance, exposure time) to reasonable values. For example, the first images should record 1° sectors at an exposure time of approx 5–20 s when using bending magnet-derived synchrotron radiation or 10–20 min when working with a rotating anode source.

The crystal-to-detector distance should be set such that diffracted spots up to 2–3 Å will be captured (*see* **Note 4**).

2. Record two images at settings of 0° and 90°. Examine the image to determine if the sector size been selected appropriately for the size of the unit cell. Are the diffracted spots well resolved? If not, the distance between crystal and detector should be increased. Is the time of exposure to the X-ray beam optimal? If it is too short, few recorded intensities will have the required statistical significance to be used by software. If the exposure time is too long, inaccurate measurements will result from saturating the detector. The operating principle is to obtain diffraction images containing 100 or more well-measured (intensity [I] divided by sigma [σ] > 5) reflections.
3. If necessary, adjust the starting camera parameters based on the appearance of the first image(s) and repeat.
4. Supply the recorded images to an oscillation data processing package (d*TREK, DENZO, DPS, or MOSFLM) for obtaining estimates of the orientation of the crystal with respect to the laboratory axes, the crystallographic unit cell, and the space group. Follow the instructions described in the operating manual of the appropriate software. A more detailed treatment of software analysis of X-ray oscillation camera images appears in Chapter 5.

3.5.2. Measuring a Complete Dataset

A full discussion of the operation of software to scale intensities across images is found in the operating manual for each software package and is discussed in more detail in Chapter 5.

3.5.3. Postprocessing Analysis

The first analysis of a newly prepared macromolecular crystal is completed by the (1) calculation of the Matthews coefficient ***(1)*** to estimate the contents of the asymmetric unit, (2) inspection of the native Patterson and self-rotation functions to identify elements of noncrystallographic symmetry (if any), and, finally, (3) calculations to detect the presence of twinning in the diffraction data. These topics are covered in greater detail in Chapter 6.

4. Notes

1. In some cases, crystals grow embedded in precipitated protein or attached to the proteinaceous “skin,” which forms at the air–mother liquor interface. Best results will be obtained when this skin is carefully teased away using an appropriate tool (e.g., Micro-Knife from Hampton Research).
2. Alternatives to SDS-PAGE analysis of dissolved crystals include analysis by a variety of chromatographic chemistries that include size exclusion, ion exchange, and reverse phase or mass spectrometric methods. These would be especially applicable when looking for the presence of species not amenable to gel analysis. The exact experimental condition will vary with the macromolecule under investigation.

3. It is sometimes necessary to add protein to the synthetic mother liquor in order to preserve the diffraction properties of the crystal.
4. When one is not sure whether a particular crystal is macromolecular or salt, it is advisable to first collect a 5° oscillation for 30 s (on a rotating anode), with the detector set at a distance such that 2 Å data or better can be recorded. A diffraction pattern with dark spots at high resolution and no reflections at low resolution will indicate that the crystal does not contain a macromolecule.

References

1. Matthews, B. W. (1968) Solvent content of protein crystals. *J. Mol. Biol.* **33,** 491–497.
2. Otwinowski, Z. and Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode. In: *Macromolecular Crystallography, 1.6 edit., Vol. 276,* (Sweet, R. M. and Carter, C. W., eds.), Academic Press, New York, NY, pp. 307–326.
3. Pflugrath, J. W. (1999) The finer things in X-ray diffraction data collection. *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 1718–1725.
4. Leslie, A. G. W. (1992) Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography* 26.
5. Winn, M. D. (2003) An overview of the CCP4 project in protein crystallography: an example of a collaborative project. *J. Synchrotron. Radiat.* **10,** 23–25.
6. Rossmann, M. G. and van Beek, C. G. (1999) Data processing. *Acta Crystallographica* **D55,** 1631–1653.
7. Sambrook, J. and Russell, D. W. (2001) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
8. Teng, T. -Y. (1990) Mounting of crystals for macromolecular crystallography in a free-standing thin film. *J. Appl. Cryst.* **23,** 387–391.
9. Rodgers, D. (1997) Practical cryocrystallography. In: *Macromolecular Crystallography, Vol. 276,* (Carter, C. W. and Sweet, R., eds.), Academic Press, New York, NY, pp. 183–202.
10. Garman, E. F. and Schneider, T. R. (1997) Macromolecular crystallography. *J. Appl. Cryst.* **30,** 211–237.
11. Hicks, L. M., Hunt, J. L., and Baxter, C. R. (1979) Liquid propane cold injury: a clinicopathologic and experimental study. *The Journal of Trauma* **19,** 701–703.
12. Blundell, T. L. and Johnson, L. N. (1976) *Protein Crystallography*, Academic Press, New York, NY.

5

X-Ray Data Collection From Macromolecular Crystals

Elspeth Garman and Robert M. Sweet

Summary

Instruments, methods, and software for modern macromolecular crystallography is becoming so effective that molecular biologists often can solve structures from their crystals by working "without a license." In this chapter, the authors attempt to demystify some of the apparatus and techniques by providing a roadmap. Current methods for collecting X-ray diffraction data from macromolecular crystals are described. The principles of operation of the required X-ray sources, optics, goniometers, and detectors are outlined, and a typical data collection protocol is presented. Optimization of data quality is a pivotal stage in the whole crystallographic process, so much attention is given to the detailed setting up of the experiment. This is followed by a summary of the basic ideas behind the diffraction image-processing packages and their application to data reduction. Despite the increasingly "black box" nature of these computer programs, understanding how they extract the intensities, errors, and indices from the data can make subsequent structure solution and refinement much easier.

Key Words: Data collection; optics; rotating anode generator; synchrotron; X-ray detector; data processing; integration; autoindexing; scaling; multiplicity; completeness; *R*-value.

1. Introduction

Once a suitable macromolecular crystal has been grown, the next stage is to test it to see if it diffracts well enough to be used for three-dimensional (3D) structure solution. This test involves mounting the crystal on an X-ray source, which is usually either a rotating anode generator in the home laboratory or an electron storage ring fed by a synchrotron at a remote site. Testing and characterization of crystals bigger than approx 80 μm in their largest dimension can usually be carried out at home, but smaller crystals require the extra brilliance of the synchrotron source. The use of synchrotron radiation (SR) has grown enormously over the last 15 yr because of increased availability and improved technology; of the structures deposited in the Protein Data Bank during 1991,

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

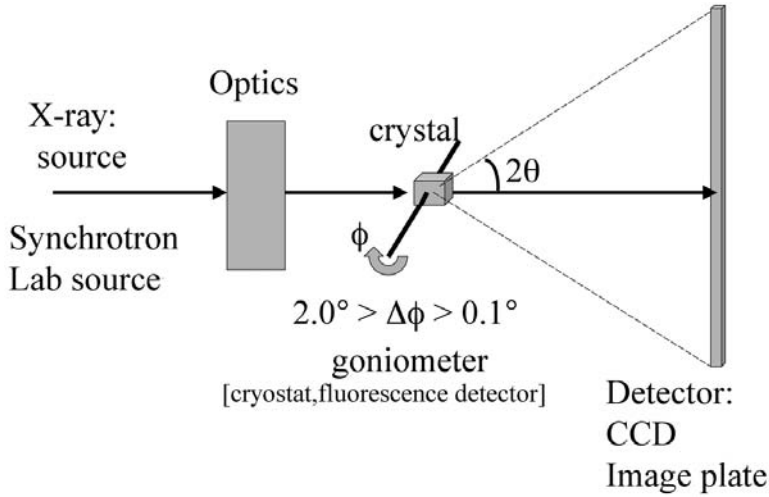


Fig. 1. A bird's eye view of an X-ray experiment. The X-ray beam is produced by the source on the left and is conditioned by the optics before hitting the crystal mounted on the goniometer. Scattered X-rays (diffracted and background) are detected on the CCD or imaging-plate detectors.

which required *de novo* phase determination, 18% were solved using diffraction data collected at synchrotrons, whereas in 2003 this had risen to close to 80%. The synchrotron has thus become an essential tool for a structural biologist.

An X-ray source for protein crystallography is fitted with the following array of equipment as shown diagrammatically in **Fig. 1**:

1. Optical elements to produce an intense beam of defined wavelength (energy).
2. A collimator system or set of slits to define the beam envelope.
3. A "goniometer" that is a motorized stage allowing accurate rotation of the crystal about one or more axes. This enables the crystal to be rotated by a small angle (typically 0.1°–1.0°) in the beam during each X-ray exposure, so that diffraction is sampled in 3D space, the so called "rotation method" ***(1)***. The goniometer has a removable "goniometer head" for convenient alignment of the crystal at the beam position.
4. A cryostat using an open flow of cooled gaseous nitrogen gas to keep the crystal at around 100K during the experiment.
5. A beamstop downstream from the crystal and in line with the collimator or slits to catch the main beam.
6. A computer from which the experiment is driven and controlled.

Once the crystal is mounted on this hardware, described in depth in the next section, decisions must be taken on how to collect the data. There is a bewildering number of parameters to be input by the experimenter, and these are

detailed in **Subheading 3**. The measurement of the diffraction data is a vital step in the experiment, because although the X-ray data collection is fast compared with the other parts of a structural project, one has to work with the data for a long time afterwards. Thus, taking extra time to optimize the data quality is very worthwhile because good data make structure solution more straightforward and can avoid frustration and failure. This optimization requires attention to many seemingly trivial aspects of the experiment: the devil is indeed in the detail. We give some pointers for making the most of your sample in **Subheading 4**.

The software for analysis of diffraction images is now extremely powerful and enables almost "blind" processing of standard cases. The principles underlying these programs are outlined in **Subheading 5.**, as well as some clues on which output statistics to monitor during processing.

2. Equipment

The apparatus used to perform crystallographic experiments is evolving quickly. The most rapidly changing component is probably the source of X-rays. Modern home-lab X-ray generators rival the power of older synchrotron sources, and modern synchrotron sources surpass the sun in brightness. The optical systems used on the modern generators are a small but crucial component, employing cunning aspects of diffraction physics and materials science to make these generators so powerful. The X-ray detectors, which supplant X-ray film as virtual video cameras for X-rays, provide sensitivity, spatial resolution, and flexibility to allow measurement of diffraction intensities to unprecedented accuracy. Finally, cryogenic equipment to protect specimens from rapid radiation damage in the X-ray beam closes the loop in the modern X-ray crystallographic laboratory. We will describe and give some examples of each of these components.

2.1. Sources

Conventional X-ray sources, the basic scheme for which is shown in **Fig. 2**, depend on a beam of electrons, drawn from a hot filament by a high-voltage electric field. The electrons strike an anode, usually made out of a pure elemental metal (copper or molybdenum), driving an electron from an atomic orbital. The emitted radiation depends on the fluorescence spectrum of the emitting element. The high-power generators used for macromolecular crystallography typically employ a water-cooled, rotating anode surface to carry away the electrical power used to generate the X-rays, a scheme for which is shown in **Fig. 3**. This power is typically several kilowatts, and the area of the source is not the roughly 1 mm^2 shown in **Fig. 3**, but a fraction of that, so the power density is huge ***(2)*** (*see* **Note 1**). Another important source of X-rays for

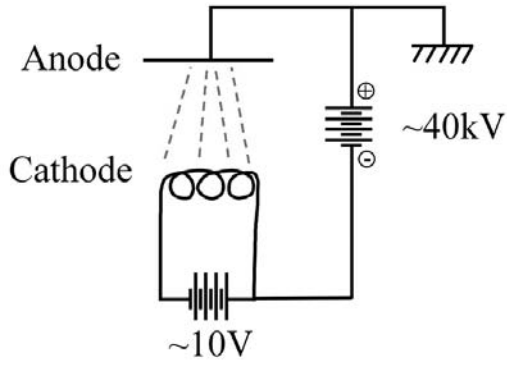


Fig. 2. A basic scheme for the simplest possible X-ray generator.

diffraction studies these days is the "synchrotron X-ray source." Enter these words into your favorite web browser and you will easily find a source nearby if you are in Japan, Europe, or the United States. A cartoon view of the National Synchrotron Light Source (NSLS) on Long Island in New York is shown in **Fig. 4**. The NSLS, like all other second-generation light sources, produces most of its usable SR from the dipole, or "bending" magnets. These magnets cause an electron beam to follow a closed, polygonal path as they traverse a vacuum pipe in the shape of a "ring." There are two accelerator rings producing radiation at the NSLS. The lower-energy ring (750 MeV) produces light of wavelengths from the infrared through to low energy ("soft") X-rays. The higher energy ring (2.6 GeV) produces higher energy ("hard") X-rays for diffraction studies. The X-rays pass to the experimental stations through vacuum-filled pipes, known as beam lines. The experimental stations are inside metal enclosures known commonly as "hutches." Both of these concepts are shown in the cartoon of **Fig. 4**.

An oversimplified view of how SR is produced is shown in **Fig. 5**. The relativistic electron is "flicked" to the side when it passes between two magnets. This induced electric field, traveling essentially at the speed of light, becomes a photon traveling in the direction that the electron is traveling at the time of the photon emission. To produce a high-energy X-ray photon, the energy of the electron must be high enough (a few giga-electron volts, GeV), and the curving of the path must be tight enough (a few meters bending radius).

There are three reasons why SR can be especially useful for crystallographic measurements. First, the total intensity of the X-ray beam can be several orders of magnitude greater than from any conventional source. Second, because the source of the X-rays can be a fairly small packet of electrons (a

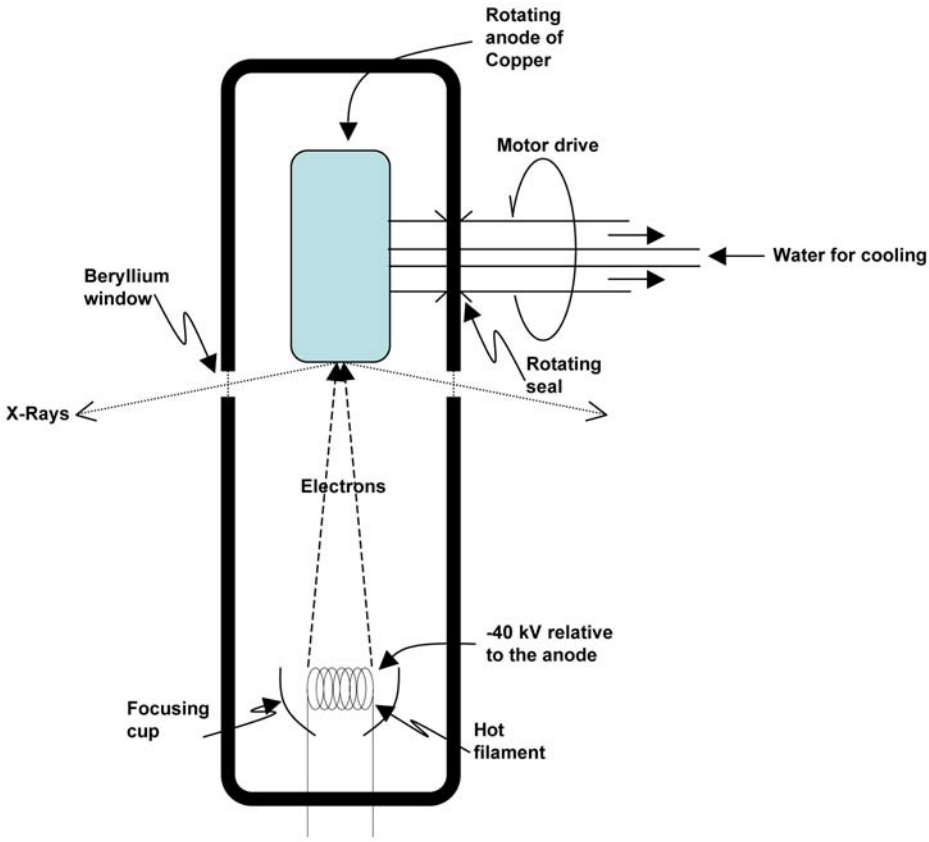


Fig. 3. Rotating-anode X-ray generator: the anode rotates at 3–6000 rpm to allow the surface of the anode, often of copper, to withstand the 3–5 kW of power. (The actual current of electrons is in the range of 100 mA.) The focusing cup, at a slightly negative voltage relative to the filament to repel the electrons, produces a focus that is 0.3×3.0 mm or so. This is reviewed at an angle, foreshortening the shape to an approximate square; 5.7° gives a 10:1 decrease. The electrons travel through a good quality vacuum of about 10^{-7} Torr.

few tenths of a millimeter), and because the emission is somewhat directional, the beam can be well collimated to produce nearly parallel rays. The availability of high-beam intensity means that the diffraction experiments can be performed more quickly than with conventional X-ray sources. The nearly parallel beam means that very rich diffraction patterns, say from viruses or multicomponent complexes like ribosomes, can be resolved well *(3,4)*. A third advantage is that there is usually a range of X-ray photon energies or wavelengths

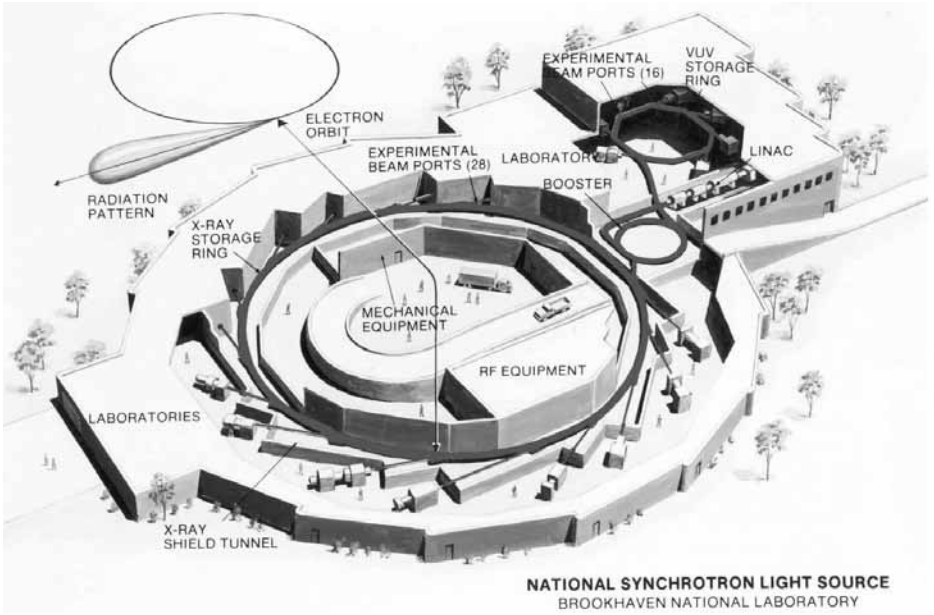


Fig. 4. The National Synchrotron Light Source (NSLS) is at Brookhaven National Laboratory on Long Island, NY. Electrons circulate at an energy of 2.8 GeV (10^9 electron volts) in a polygonal ring with a diameter of about 54 m.

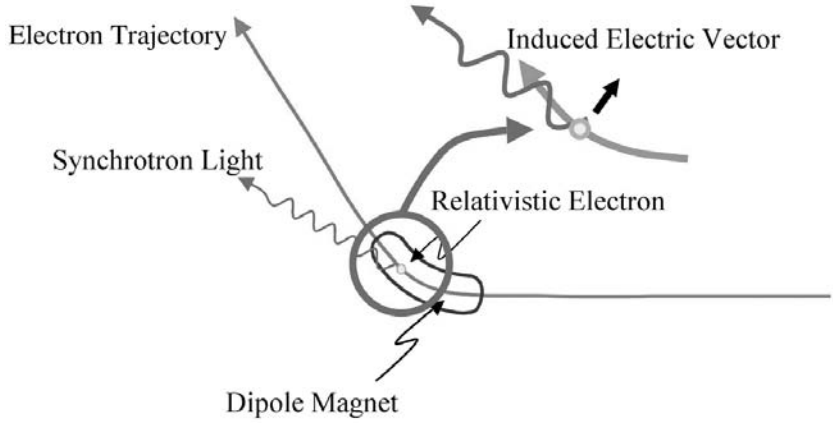


Fig. 5. Simplistic view of the source of synchrotron radiation. The electron is "flicked" to the side when it passes between the magnets. This induces an electric vector, which stimulates the photon emission.

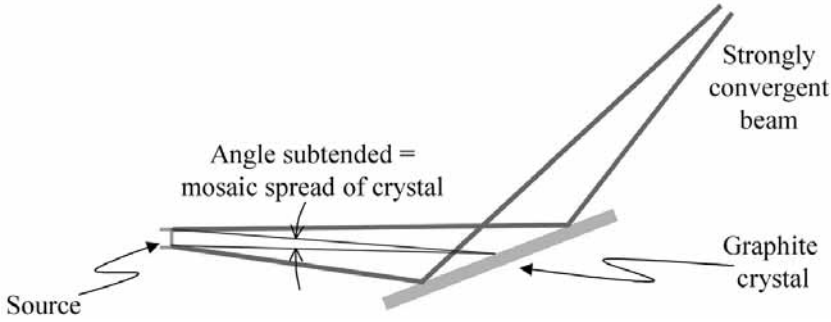


Fig. 6. X-ray monochromators depend on Bragg diffraction. For conventional sources, one used to use a graphite monochromator: the mosaic spread matches the crossfire of the beam for greatest reflectivity.

produced. Often X-rays of selected energies will scatter anomalously from heavy atoms in the macromolecule, and these can be used to solve "the phase problem."

Although many useful structures can be solved by use of conventional X-ray sources, the use of SR has revolutionized several aspects of macromolecular structure determination resulting from the properties of the beam. First, work simply can be done much faster. Second, structures of dramatically large complexes, as previously mentioned, have been solved, and these studies simply could not be done with conventional sources. Finally, the usefulness of anomalous scattering effects for structure solving has made some problems trivial. Although it is not routine, it is not unusual at a SR source for one to be looking at an interpretable electron density map only a few hours after an uncharacterized crystal has been put into the X-ray beam.

2.2. X-Ray Optics

The X-ray beam is often conditioned or altered in some way before it is used for a diffraction experiment. In the case of a conventional source, the X-rays are from a single emission line of the anode element, so they are nearly monochromatic. However, the "color" or X-ray wavelength bandwidth can be narrowed and extraneous radiation can be eliminated through use of a monochromator. In the case of a polychromatic SR source, the wavelength might be chosen directly. Typically, Bragg diffraction is used for this purpose, and monochromator crystals are of different sorts. For years, but now decreasingly, one used a man-made "crystal" of pyrolytic graphite. The sort of diffraction one gets from this is shown in **Fig. 6**: the beam was typically not very parallel, owing partly to the not-too-

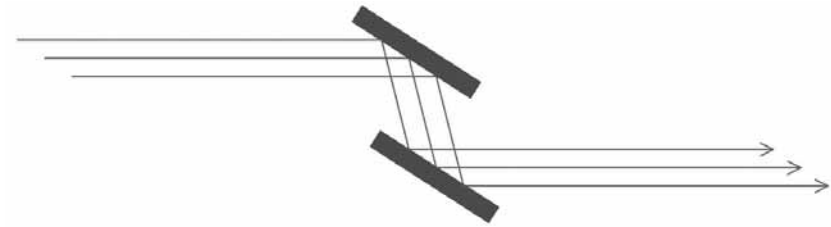


Fig. 7. A silicon crystal monochromator. Two crystals are used, one to monochromatize and one to deflect the beam back up into the horizontal plane. Bragg's law determines the wavelength.

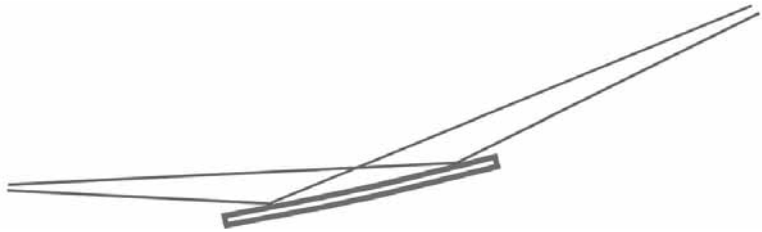


Fig. 8. General scheme for Franks mirrors. The glancing angle is typically 0.5°. On a conventional source the distances from source to mirror and mirror to specimen are 100–200 mm.

small size of the source and the imperfection in the graphite crystal. The very parallel synchrotron beam can be monochromated by a perfect crystal, and silicon cut to use the (1,1,1) planes is the usual choice. Often we use two crystals: one to monochromatize and one to deflect the beam back up into the horizontal plane. Bragg's law determines the wavelength. Such a scheme is shown in **Fig. 7**.

One might condition the beam by focusing the beam diverging from the source to get a more intense beam. We cannot make a lens for X-rays, but a method that suffices is very low-angle reflection from mirrors. One actually can use a pair of curved, very smooth mirrors to focus the beam from a conventional source in first the vertical and then the horizontal direction. Properly called Franks or Kirkpatrick-Baez mirrors, they are commonly known as "Yale" mirrors because of the popular mounting apparatus devised at Yale. The general scheme for a Franks mirror is shown in **Fig. 8**. The glancing angle is typically 0.5°. On a conventional source the distances from source to mirror and mirror to specimen are 100–200 mm.

At a SR source, the beam is only slightly divergent, but because the distances are often tens of meters, one would like to focus it, and thereby concentrate the beam. One should be able to focus with a mirror shaped like an ellipsoid of revolution (*see* **Fig. 9**). Technically we cannot actually achieve that, so we make a cylinder, then bend it to the shape of a toroid.

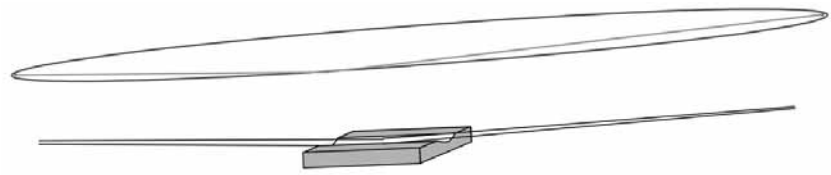


Fig. 9. The geometry of the ellipse suggests a way in which we can focus a diverging source of light. We approximate this by bending a cylinder to form a small patch of toroid.

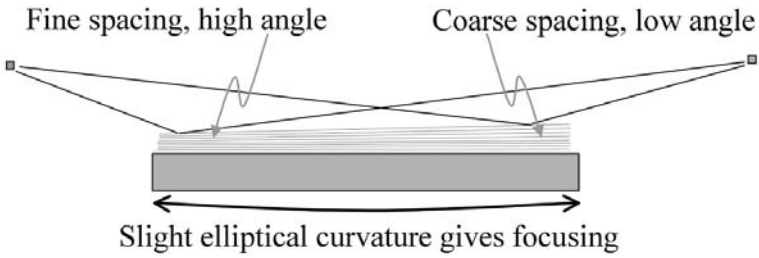


Fig. 10. Focusing of the X-rays is accomplished with the combination of a graded multilayer coated on a curved surface.

A commercial firm has devised a clever optical arrangement that mixes these two concepts. The Osmic company creates a synthetic crystal by deposition of alternate layers of dense and lighter material: a multilayer crystal. This material is deposited on a base or substrate that is curved to a gentle optical shape. Two of these are then joined together at right angles along their long edge: the complete L-shaped Osmic confocal focusing multilayer mirror is capable of focusing in two dimensions *(5)*. **Figure 10** shows the scheme that describes the functioning of this graded multilayer, which is coated on a curved backing. The brightness and overall quality of the beam from this device has made "home" sources extremely useful for many crystallographic experiments.

2.3. Diffractometers

The apparatus used to measure diffraction data goes by various names; here we call it a diffractometer. It typically comprises some sort of "collimation" system consisting of an aperture or slit pair to give a final limit to the size of the X-ray beam, a mechanism to hold the specimen crystal precisely in the beam and to move it in some controlled way to sample diffraction space (crystal orienter), and a detector to record the diffracted X-rays. The sort of diffractometer used with conventional X-ray sources is typically a commercial self-contained unit. Those used at synchrotron sources usually contain commercial off-the-

shelf components, but often the assembly somehow manifests the personality of the builder. The collimation system must be chosen carefully, and we will discuss this further in **Subheading 4**. The crystal orienter is often a single axis, set to rotate perpendicular to the X-ray beam, to enable the crystal to be oscillated through a small angle in the beam during data collection *(1)*. Occasionally, especially at a SR source, one might find a three-axis system wherein two oblique (or orthogonal) axes are used to put the crystal in some particular orientation before having the whole assembly rotated around the third axis.

The most important component of the diffractometer is the X-ray detector. The original X-ray detector for crystallography was X-ray film (a two-dimensional [2D] detector). For some time in the mid-1960s through the 1980s a movable X-ray counter rather like a Geiger counter (usually crystal scintillation or gas-filled proportional counter) was placed behind a small aperture (a zero-dimensional detector). The crystal and detector were moved systematically to bring individual reflections into a diffracting position one at a time so one could measure the diffraction intensity. Finally in the 1980s electronic 2D detectors of various sorts were devised, and these are what we mostly see today. In the mid-1980s a commercial firm produced a fairly popular video-based detector (the "FAST"), which depended on phosphors turning X-rays to visible light for the video system to detect *(6,7)*. About the same time two other firms popularized xenon-filled multiwire proportional counters (*[8–10]*; reviewed in **ref.** *11*). The video detector was used successfully at both synchrotron sources and on conventional X-ray generators. The multiwire detectors were used on conventional sources only. Both of these early electronic 2D detectors were responsible for many important structures. Soon after, a system based on "storage phosphors" was developed (*[12]*; reviewed in **ref.** *13*). These phosphors (often europium-doped barium halide) would absorb an X-ray photon and then would go into a meta-stable excited state that would be stable for a long time. The stored energy was released with the scanning of the phosphor plate with a red-light laser, and the blue light that was emitted could be measured. These devices were hugely successful because they were much more sensitive than X-ray film, the sensitive surface and dynamic range were bigger than that of the video system, and they had finer resolution elements (pixels) than the multiwire systems. They could be automated to work reliably for years and had a high dynamic range. Essentially every SR source used these for a time, and they still can be found in both conventional and SR laboratories.

The most abundant 2D electronic detector now seen at SR sources is based on a solid-state electronic imager known as a charge-coupled device or CCD (reviewed in **ref.** *14*). This is closely akin to the imagers used in modern digital cameras. The most common application of these is shown schematically in **Fig. 11**. A light-emitting phosphor is bonded to the large end of a tapered fiber optic device, and a CCD chip is bonded to the small end. The large (several tens to a

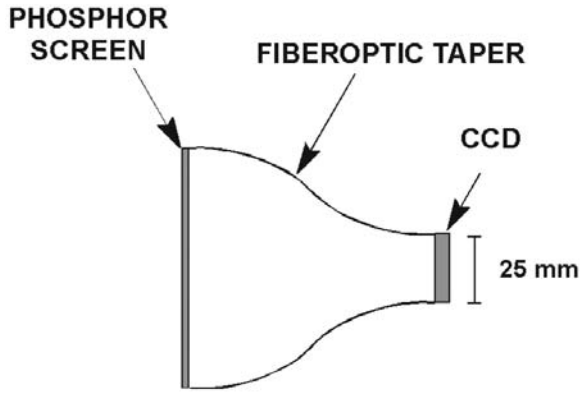


Fig. 11. The general scheme of modern CCD-based X-ray detectors.

Fig. 12. The Brandeis B4 four-module CCD-based detector. The front-left surface would be covered with a phosphor. One can see the originally cylindrical surfaces of the tapered fiber optic before it was cut to a square. Readout electronics are in the metal cylinders. (Reproduced from **ref.** *15*. with permission from IUCr.)

few hundred millimeters) diffraction pattern is mapped with adequate efficiency onto the small (a few tens of millimeters) semiconductor chip. Great effort has been made to make these chips nearly perfect in their imaging (no dead pixels), to keep the read-out noise low, and to provide rapid readout. In order to make

large-surface area detectors, several of these modules are connected together to act as a single, large detector. An example of the insides of such a detector is shown in **Fig. 12** *(15)*. The front surface of this detector is 20 × 20 cm.

2.4. Cryostats

An important advance in macromolecular crystallography during the 1990s was development of techniques for cryocooling of the crystalline specimens *(16–18)*. Macromolecular crystals are quite sensitive to X-ray damage, owing partly to the necessity to maintain covalent bonds between light atoms, and partly to the fact that the crystals are essentially clumps of matter sticking together in places, with all the empty space filled with solvent. Absorption of an X-ray photon by any part of a macromolecular crystal will be accompanied by production of a large number of electrons. These electrons can diffuse freely through the water-filled interstices of the crystal, wreaking havoc on the bonds in the molecules themselves. Many workers helped to develop techniques gradually during the decade to stabilize crystals, to cool them quickly enough that water-ice crystals cannot form, and to suspend them in a stream of cold gas in a way suitable for a diffraction experiment, as described in Chapter 1.

The usefulness of crystal-cooling methods is only as great as the reliability of the apparatus that will keep the crystal cold during the diffraction experiment. Cryocooling of crystals is so important an aspect of macromolecular crystallography that the cryocoolers are ubiquitous in X-ray labs around the world. There are several devices made by commercial vendors. They vary somewhat in their approach to the handling of cryogenic liquids and gases (liquid nitrogen is the usual cryogen), but the goal is to produce a stable stream of gas that is parsimonious in use of liquid nitrogen, does not cost too much to purchase, and will not ice up with water ice from the air during a long experiment.

3. Methods

Figure 13 shows a generic diffraction experiment with the steps labeled as in the sections following. The aim of the experiment can vary: it might be either to test if the crystal diffracts at all, to prescreen it for future data collection, or to collect the data immediately. These days the crystal is most commonly cryocooled, and some effort and preparation are required to optimize the flash-cooling of the sample: procedures for this are described in Chapter 1 (*see* **Note 2**).

3.1. Checks Before Beginning

There are several checks that are worth reviewing before the crystal is placed on the goniometer head.

1. X-ray flux maximization. If it is possible for you to control the table alignment, try to maximize the X-ray flux that on some systems is displayed on an LCD as

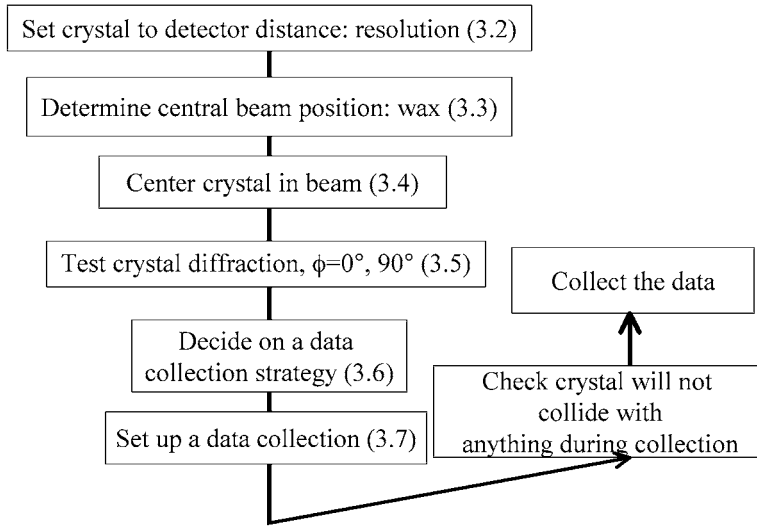


Fig. 13. A flow diagram of the steps involved in setting up an X-ray data collection.

the output of an ion gage in the slits. The alignment adjustment is now automatic on some systems, once requested by the experimenter, but in others manual controls that move the vertical and horizontal position of the detector table are provided.

2. Cryostat alignment. The center of the cold gaseous nitrogen must be centered on the crystal position so that the crystal is maximally protected from the surrounding warm air. You can check it with a so called "cryoalignment nozzle" ***(19)***, which fits onto the end of the cryostat and has a pin for you to align at the crystal position (an empty properly aligned cryoloop will help here) using the *x*, *y*, and angle adjustments usually available on the cryostat stand. If there are no fine adjustments on the stand, the alignment is a two-person job, as when you undo the locking screws on the stand, the cryostat can suddenly move a lot and hit the beamstop/collimator/goniometer head etc.
3. Liquid nitrogen supply. If your cryostat uses liquid nitrogen from a supply Dewar rather than manufacturing its own liquid nitrogen, check that there is enough to last through the data collection (usual consumption is between 0.6 and 1.5 L of liquid per hour depending on which cryostat you are using).
4. Data storage capacity. Sometimes computer disk space runs out part of the way through a data collection and the acquisition software stalls as there is nowhere to write the images. Find out the size of each image, and the spare capacity of your disk, to ensure that this problem will not occur.
5. On some detectors, although these days increasingly few, the spatial distortion has to be calibrated prior to data collection, and the distortion correction lookup table stored for use by the processing software later.

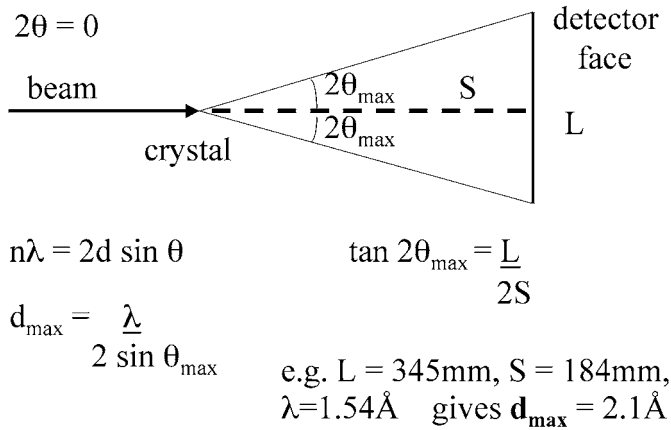


Fig. 14. Geometry of the diffraction experiment for calculation of the maximum resolution of the data (d_{dmax}).

3.2. The Detector-to-Crystal Distance: Resolution

The closer the detector is to the crystal, the higher the resolution of the data that can be collected. The relationship is illustrated in **Fig. 14** and is calculated using basic trigonometry and Bragg's law (*see* **Fig. 14** legend). Modern data collection software usually tells you the resolution at a specific detector distance, but it is often useful to be able to calculate it yourself.

The resolution you will need for your experiment depends critically on what is the problem being addressed: for heavy atom derivative searching (multiple isomorphous replacement) 3.5Å is usually enough, whereas for a final publishable dataset, you will want to collect as high a resolution as possible (although *see* **Subheading 3.8.**).

3.3. The Beam Position Relative to the Rotation Axis/Axes

The goniometer rotation axis/axes must intersect with the beam path and with the line of sight of the crystal viewing/alignment camera. The monitor attached to the camera will show a set of cross wires, and it can be dangerous to assume that if the crystal is placed there, the beam will hit it. In-house you can check the alignment by mounting a metal pin on the goniometer head, removing the backstop (only do this test under the supervision of the person responsible for the X-ray equipment) and then placing a fluorescent paddle or much better, a CCD camera, in front of the detector. With the lights off, you can then see the shadow of the pin on the screen or on the CCD monitor, and check that the rotation axis and the beam are coincident. You can then check the true center of the camera by rotating the pin and aligning it using the sledges on the

goniometer head until it does not move upon being rotated 180°. Mark this position on your camera monitor screen with washable felt-tip pen, so you know the right place for your crystal to sit.

Both parts of the previously mentioned alignment test are essential steps at synchrotron sources as much time can be lost if the beam is not hitting the crystal. Sometimes they are already part of the daily checks carried out by the beamline scientists, and the current beam position will be drawn on the camera monitor screen, but you should always ask when it was last checked and by whom. To find the position for yourself on a beamline that does not have a specified procedure, you will need either a small fluorescent paddle or a 1×0.5 cm strip of X-ray sensitive paper (it is usually pink or green) stuck to a piece of stiff card to hold it flat. The paddle or pink paper is mounted on a goniometer head, most conveniently via a cryocap. If the paddle/paper is carefully aligned on the rotation axis and then exposed for 30 s to the X-ray beam, the beam will be visible glowing on the paddle in the crystal-viewing camera, or as a black silhouette on the paper. Rotate the paper so that it is normal to the line of sight of the viewing camera, and that gives the beam position.

Once the internal alignment has been established, you can find the central beam position relative to detector face by melting a blob of wax onto a cryoloop and exposing it to the X-ray beam for a short time. The center of the rings on the detector, which can be found using available software (e.g., MOSFLM), is the main beam position. Sometimes one can find that a preparation of fine silicon-crystal powder has been placed on a crystal cryopin for the same purpose. The rings are at known resolution, so software exists to compute the X-ray wavelength or the crystal to detector distance (but not both as they are correlated [*see* **Subheading 3.2.**]). If you are unfortunate/careless enough to have ice rings on your crystal diffraction pattern, they also can be used to find the direct beam position. At many synchrotron sources an attenuated image of the direct beam itself is simply recorded on the detector to get the same answer directly. Either way, it is *absolutely vital* to know the coordinates of the beam position because this is the origin of the reciprocal lattice. An incorrect value can result in misindexing of your images at the processing stage (*see* **Subheading 5.2.**).

3.4. Center Crystal in the Beam

On some systems this step now simply involves clicking a mouse button when the cursor is over the crystal on a crystal-viewing screen, rotating the crystal by 90°, and clicking again. However, more usually, the experimenter has to use a "goniometer key" to adjust the two perpendicular translation slides manually on the goniometer head until the crystal does not move when it is rotated on the goniometer. If it moves it is *not* on the rotation axis and, thus, may move out of the beam during the rotation range covered by the data collection.

The centering of the crystal must be checked at the maximum zoom mode of the crystal viewing camera: most cameras do not retain the same center when their zoom is changed.

3.5. Test the Crystal Diffraction

The first two images to collect are $\Delta\phi = 1.0°$ oscillations at $\phi = 0°$ and at $\phi = 90°$. This is because a crystal can often appear to be diffracting well, but when rotated through 90°, it turns out to suffer from disorder. This most usually manifests in the image as an overabundance of spots on one of the images. Such crystals are unsuitable for data collection and it is a waste of time to carry on with them.

You can judge the diffraction quality from these two initial images. There are a number of possibilities which include: the crystal is in fact salt (collect an image with a large $\Delta\phi$, e.g., 5°), the crystal is obviously split because there are two superimposed diffraction patterns, the crystal is internally twinned (this may not be apparent until the end of the image processing or even until there are difficulties during refinement of the structure), or the crystal may be disordered. Disorder along one or more axes of the crystal lattice can give high mosaicity or statistical disorder, in which the diffraction spots appear to be sharp in parts of the image and smeared in others in a regular pattern.

Alternatively the diffraction may be weak or there may be no diffraction at all (if the latter, do a room temperature test before abandoning your crystallization conditions). At this stage you may need to reassess crystal-to-detector distance, because you will improve the signal-to-noise ratio of your data by pulling the detector back to place the highest diffracting spots at the edge of it, rather than collecting data with the outer third of the detector bereft of spots.

3.6. Decide on a Data Collection Strategy

There are now several excellent programs available for planning a data collection, and there is absolutely no excuse for failing to use them. They will allow you to collect as complete data as possible in the time available to you. If you autoindex your first or first few diffraction images to find the probable unit cell, Laue group, and orientation of the lattice with respect to the beam, you can feed the strategy programs with the information they need to compute the minimum range of rotation required for a completeness specified by the user.

Software packages for this task are: the "Strategy" option in MOSFLM *(20)*, Predict *(21)*, XPLAN for XDS *(22–24)*, Strategy *(25)*, COSMO (Bruker Nonius), or the internal strategy option in HKL2000 *(26)*. The crystallographic theory coded in these packages can be found in **ref.** *27*.

The importance of obtaining complete data cannot be over emphasized. Incomplete data, especially if it is the low-resolution data that are systematically missing, can result in failed structure solution or uninterpretable electron density maps.

For low-symmetry space groups where it is physically difficult to access some parts of reciprocal space, you can use the arc adjustments on the goniometer head to offset the crystal (while checking that the crystal is still centered in the beam). If this strategy does not provide enough offset, an additional flexible piece of metal wire can be attached between the end of the cryopin and the loop, which can be bent during the experiment ***(18)***.

3.7. Set Up a Data Collection

The control software for the diffraction experiment requires you to specify a number of experimental parameters. The choices you make will affect the ultimate data quality, so it is important to understand the compromises you may have to make when deciding each value.

1. Crystal-to-detector distance: resolution. The crystal-to-detector distance, S, must be large enough so that the spots of the diffraction are separated on the image, but also small enough to collect the desired resolution at the edge. It is best to have at least one clear pixel (with an intensity near that of the rest of the background) between neighboring spots, as otherwise the processing software will not be able to separate them and integrate them accurately. The further away the detector is, the better the signal-to-noise ratio will be. This is because the background decreases as $1/S^2$ from the crystal, whereas the diffraction intensity/pixel decreases as $1/D^2$, where D is the distance between the detector and the source of the X-rays. D is a longer distance than S, and at a synchrotron D is very large, so as the detector is pulled back, $1/S^2$ decreases faster than $1/D^2$.
2. The oscillation angle for each image, $\Delta\phi$. If you choose $\Delta\phi$ to be too large, the data will suffer from "overlaps," where spots with consecutive values of *h*, *k*, or *l* are not separated in the ϕ direction. They impinge on the detector in exactly the same place so they are indistinguishable from one another, and they will be rejected by the processing software. There is absolutely no way to rescue them, and the data will be incomplete.

   The maximum $\Delta\phi$ that can be used if overlaps are to be avoided is approximately:

   $$\Delta\phi = (d_{max}/\text{maximum primitive unit cell dimension}) - \text{mosaicity}$$

   where the answer is in radians (multiply by $180/\pi$ to get degrees). The parameter d_{max} is the minimum Bragg spacing one hopes to achieve. It may be possible to use a larger $\Delta\phi$ than given by the previously shown formula, depending on the orientation of the maximum primitive unit cell dimension with respect to the beam.

   If you choose $\Delta\phi$ to be small (e.g., 0.1°), all the reflections are likely to be partially recorded (i.e., the total intensity is split over two or more images in ϕ) rather

than fully recorded (i.e., all the diffraction intensity of a reflection recorded on one image because the $\Delta\phi$ is larger than the mosaic spread). If the detector has a fast readout time compared to the exposure time, so-called "fine slicing" (small $\Delta\phi$) can be advantageous because the fine sampling of the reflection in phi allows its profile to be fitted in the phi direction as well as in *x* and *y* (i.e., 3D profile fitting). The relative merits of fine and coarse slicing can be found in **ref.** *28*, but the principal advantage is that the smaller the rotation for each image, the smaller the X-ray background. This must be balanced against a slightly increased error on each image, more experimental overhead in time spent reading out the detector, and increased data to be stored and reduced.

3. Exposure time or 'dose.' The experimenter will either have to specify the time for each image, or the X-ray 'dose' for each. The latter has the advantage that if the beam flux varies with time, the exposure per image will remain constant. The exposure should be sufficient to obtain an I/σ (I) (intensity/the error in the intensity) at the edge of the detector of at least two, but not so great that the low-resolution spots are overloaded, i.e., each pixel has reached its full dynamic range. This number is around 60,000 analog units per pixel (2^{16}) for most commonly used detectors. Overloaded reflections should be rejected explicitly at the processing stage because the pixels are essentially saturated, and any further X-rays incident on them are not recorded (*see* **Note 3**). This makes the response of the detector nonlinear for the intense reflections that have the lowest statistical error and thus purport to be the most reliable. When deciding the exposure time, another consideration is the radiation sensitivity of the crystals. Radiation damage is increasingly being recognized as the source of nonisomorphism in multiwavelength anomalous dispersion (MAD) experiments and of poor scaling, as well as resulting in erroneous biological information owing to specific structural damage (reviewed in **ref.** *29*). It is usually better to settle for a complete lower resolution (shorter exposures) dataset than to collect a higher resolution (longer exposures) one that is incomplete from radiation decay. In any case, it is always useful to have a lot of *good* data. Do not hesitate to make long exposures to get the high-resolution data you want, but then be prepared to use multiple crystals if the crystals die fast, and certainly make sure there is a sweep of data taken rapidly enough that *none* of the reflections are overloaded.

   A final consideration is the motivation for the data collection. If one needs a "routine" dataset to search for a bound effector, for a molecular replacement structure solution, or for a search for an isomorphous heavy atom, the previously mentioned procedure is appropriate. If the motivation is to find subtle anomalous signals for a structure solution where the heavy atom is in low abundance, the $I/\sigma(I)$ at a reasonably high-resolution range (say 2.5 Å) should be at least five, and all low-resolution data should be measured. If the motivation is to get the very best dataset possible for high-resolution refinement of a structure, then it will pay to assure an exposure long enough to achieve $I/\sigma(I)$ at the edge of two or more, but then to rerecord the data at a much shorter exposure if the low-angle reflections are saturated (overloaded). This may require use of more than one crystal to avoid the radiation damage that may destroy the high-angle reflections you seek.

4. The total number of images. The strategy program will have told you how many degrees of data are necessary for the completeness you wanted. This value divided

by the $\Delta\phi$ value will give the number of images required. For more multiplicity, extra images above the minimum are beneficial, but only if the crystal has not suffered too much radiation damage.

5. The starting ϕ value. This will be known from the output of whatever strategy program you have used. Most modern goniometers track any manual rotations you make, and update the software value accordingly, but it is worth checking once that the physical ϕ and the software ϕ are in agreement, as otherwise you may not be at the correct start ϕ for your strategy to succeed.
6. Number of oscillations ("passes") per image. This is the number of times during one image that the ϕ axis sweeps out the specified $\Delta\phi$. At a synchrotron this is usually one per image, but on an in-house image plate it is advisable to do one oscillation every 5 min to reduce systematic errors caused by the decay of the fluorescent centers of the phosphor. Thus for a 20 min per image exposure time, typical of an in-house data collection, four oscillations per image would be appropriate.
7. Incident X-ray wavelength. This question arises only at the synchrotron, where the wavelength is often tuneable. For a MAD experiment, the wavelength will be defined by an excitation scan carried out over the absorption edge of the element that is the anomalous scatterer in the crystal. It is common to collect data at around 1 Å at the synchrotron, other things being equal. It is a nice compromise of minimizing absorption errors, minimizing crystal damage, separating spots on the detector, and providing strong anomalous signals for some heavy atoms (*see* **Note 4**). Once all the previously mentioned parameters have been set, the data collection is ready to start. During the data collection, images should be monitored to check that the crystal is diffracting as expected and that ice is not forming on the crystal (particularly for long in-house collections).
8. Modern high-precision methods: single anomalous diffraction (SAD), sulfur-phasing, and so on. The use of anomalous dispersion signals from the heavy atoms in macromolecules provides great leverage in structure solving. The MAD method is still the gold standard for high-quality phases that are independent of any model of the protein. They require that data be measured quite accurately, and of course require that the data be measured on a tuneable (synchrotron) source so the particular multiple wavelengths that are needed can be selected.

Within the last few years a method of phasing has become popular that depends on only the anomalous scattering from heavy atoms in a crystal—the SAD method. The method depends on getting monomodal phase information similar to single isomorphous replacement from Friedel mates, and then using solvent flattening and similar methods to converge to the correct phases. The first example of a macromolecular structure being solved from anomalous difference alone was that of the structure of crambin, solved by Hendrickson and Teeter ***(30)***. They used the anomalous signal from sulfur atoms in the structure to locate and refine these atoms, then used the normal scattering from the S atoms in the protein alone to resolve the phase ambiguity. The method is possible because of the accuracy of data that can be measured these days (high statistical precision plus low systematic errors) and the powerful methods of phase improvement (solvent flattening,

maximum entropy, and so on). The first of these depends very strongly on the high-performance position-sensitive X-ray detectors that are in use in most laboratories.

To some extent the SAD method is employed when one can optimize the anomalous signal by careful choice of wavelength at a synchrotron source. The first example we can find of a correct structure being solved by the method in its present form (single-wavelength anomalous followed by solvent flattening) is that by Ramakrishnan et al. of the translation factor IF3 ***(31)***. In addition, with very good quality detectors, X-ray optical systems, and beamstop one can measure accurate enough data on a Cu rotating anode source with 1.54 Å radiation to solve structures by the SAD method with the anomalous signal of the natural S atoms in the protein (*see* Chapter 10).

What is the art in measuring data of sufficient accuracy to use the MAD or SAD methods of phasing? An important consideration is to attempt to measure the Friedel mates (F[*h, k, l*] and F[*-h, -k, -l*]) in a way that minimizes any systematic error *in the measured difference*. The point is, each measurement will have systematic error with contributions from absorption of X-rays by the crystal, oblique incidence of the X-ray beam on the detector, damage to the crystal by the X-ray beam, and so on. One wants to attempt to have these errors the same for each member of the pair. To ensure this, there are several steps that can be taken. (1) Use cryocrystallography, because minimizing crystal damage minimizes errors that it causes. (2) Attempt to measure members of a Friedel pair at the same time and with similar geometry. Occasionally one can use a three-axis diffraction camera (goniostat) to prealign a crystal so that Friedel mates (actually Bijvoet discovered this) appear mirror imaged on the same diffraction image. This is possible when there is a mirror plane in the Laue point group of the crystal, which arises when there is a two-, four-, or sixfold axis in the crystal itself. (3) Measure members of the pair in a similar geometry and close together in time. With any crystal of any symmetry and orientation one can perform the operation we call the "Friedel Flip" to accomplish this. First, measure a sweep of data through some rotational sweep with the axis of rotation perpendicular to the X-ray beam (the usual situation). Then increment the axis by 180° from the original starting point and take the same rotational sweep again. One will find that the second sweep contains the Friedel mates (*-h*, *-k*, *-l*) of the original sweep, related to the first by a mirror plane perpendicular to the axis of rotation. (The reader may want to prove to themselves that the mirror plane is really the correct relationship.) If the sweeps of data are chosen to be short enough that the difference in radiation damage between them is small, this can provide quite accurate anomalous differences.

3.8. High-Resolution Data Collection

Sometimes the crystal diffracts to much higher resolution than you had expected, and it is tempting to chase the weak reflections at the edge of the detec-

Table 1
Three Typical Data Passes Required to Obtain Very High-Resolution Data[a]

Resolution (Å)	Δϕ	S (mm)
0.83–2.0	0.5°	110
1.35–8.0	0.8°	250
2.66–25.0	1.5°	550

[a]This table shows the resolution ranges, oscillation angle per image, and crystal-to-detector distance, S.

Table 2
The Crystal-to-Detector Distance[a]

Incident wavelength (Å)	S (mm)
1.0	100
0.9	128
0.8	160

[a]The crystal-to-detector distance, S, for 1.0 Å data at the edge of a Mar345 image plate detector as a function of incident X-ray wavelength.

tor. For a high-resolution data collection you will almost certainly have to collect two or more passes at different resolutions to avoid overloading the detector (*see* **Subheading 3.7.**). Time must be allowed for these extra passes, and the crystal may suffer significant radiation damage before they are complete.

Table 1 shows values of crystal-to-detector distance and resulting resolution, and the resulting usable resolution limits for each pass for an atomic resolution data collection. The wavelength of the X-rays was only 0.79 Å: we chose the lowest practicable value for that beamline in order for the diffraction pattern to occupy a smaller cone on the detector, thereby enabling us to collect the highest resolution possible at the shortest crystal-to-detector distance. **Table 2** illustrates this point, showing the distances necessary to collect 1.0 Å data for a mar345 image plate detector (Marresearch) at different incident wavelengths. For such an experiment, you must therefore select your beamline carefully to achieve the target resolution.

4. Improving Data Quality

One of the most important issues in producing good data is the quality of the beam-collimation system. The importance of this is reflected in the fact that it is

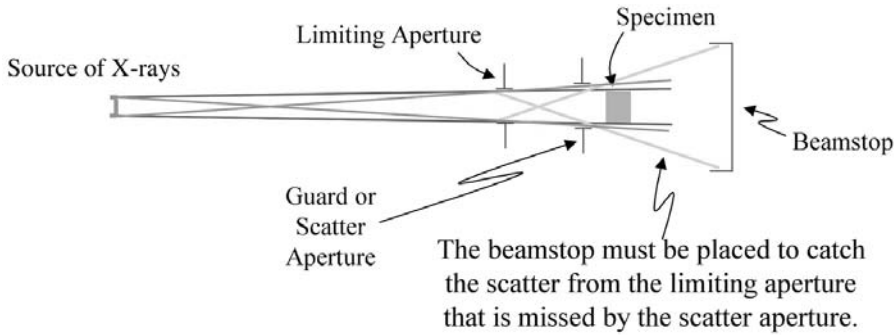


Fig. 15. A typical "collimation" system. The size of the beam is determined by the limiting aperture. The guard aperture catches scatter from the limiting aperture. The beamstop should capture all of this scatter as well as the main beam.

an aspect of the experiment over which the typical experimenter has very little control. What one finds is that the apparatus is often fairly inflexible as regards collimation, except possibly for adjustment of aperture sizes. This is to prevent someone with too little experience putting things *out* of adjustment. Nonetheless, it will pay to examine this mysterious subject in some detail.

Figure 15 shows a typical "collimation" system for either a conventional or a synchrotron source. In a classic sense, collimation means "to make the rays parallel." We generally use the term to mean limiting the beam and avoiding extra scattered rays. One wants to think of the X-rays as coming from all points on a source of finite size. Notice in the figure that the rays that come from one end of the source and go out the opposite side of the aperture produce a weak beam that gets wide faster than the main beam. The final dimensions of the beam are determined by a limiting aperture that in general should be close to the specimen. The greatest difficulty with this aperture is that X-rays scatter from it in all directions.

If one were to use a beamstop sufficient to capture only the direct X-ray beam, one would see huge amounts of scatter from this aperture, probably in rings from the metal, around the shadow of the beamstop. Therefore, one follows it by a "guard" aperture, chosen to be just big enough not to touch any of the main beam. Its size, and the distance the guard aperture lies from the limiting aperture, defines the path of all of the X-rays that must be captured by the beamstop. A small beamstop must be close; a larger one may be farther away.

Given these principles, one can think about how to employ a collimation system to make the experiment work better: (1) one wants to keep background low. First, the final aperture should be as close to the specimen as practical to keep the air path short. Second, in general, one wants the beamstop to be as close to the specimen as possible for the same reason. (2) One wants to measure all of

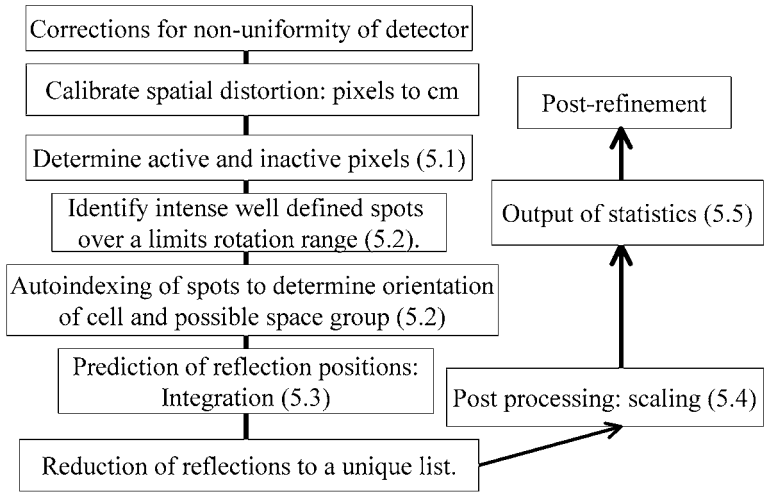


Fig. 16. A flow diagram of the steps involved in data reduction.

the low-resolution (low-angle) data. First, make the beamstop as small as possible. Second, notice that to get low-angle reflections it always will pay to put the beamstop as far as possible from the specimen, making the beamstop shadow on the detector smaller as one goes back. Of course this exposes one to having excessive background.

In reality, collimation systems are built so one can use a fairly small beamstop, fairly close, giving low-angle data while keeping background low. A thoughtful experimenter will push the limits to achieve this. In the case where one wants very-low-angle data, there are two approaches. One can construct a collimator where the two apertures are far apart. Then a smaller beamstop can be used close to the crystal. Alternatively, one can make a tunnel with thin plastic film at each end, filled with helium gas, with the beamstop inside the tunnel but far from the specimen. Helium gas scatters X-rays very little, and this will keep background from the scattered beam fairly low (for additional ways to improve diffraction, *see* **Note 5**).

5. Data Processing

After diffraction data have been measured, one must "reduce" the data to extract observed structure factor amplitudes from the diffraction images. The steps involved are summarized in **Fig. 16**. Explicitly, for each diffraction spot in the images, one must assign it its general indices *(hkl),* must measure the integrated intensity of that spot, must estimate the error of that measurement, and must make all known geometric corrections to that intensity, such as the

Lorentz and polarization factors. In general, the data-reduction process can be split into discrete steps:

1. "Index" the diffraction pattern, i.e., to determine the indices *h*, *k*, *l* of each reflection, as well as the parameters that define the size and shape of the crystal's repeating unit, and its orientation on the diffractometer.
2. Integrate the intensity of each reflection (over the total extent of diffraction onto the detector), typically using predictions of spot positions that come from the crystal parameters determined during indexing. At this point the intensity of the reflection, and the consistency of measurement, allow estimation of the error in the integrated intensity.
3. Put all the reflections from all of the images on the same relative numeric scale and then merge and average them to produce a unique reflection list with associated intensities and realistic error estimation.

Data processing packages that are currently available are as follows:

Mosflm	http://www.ccp4.ac.uk/html/mosflm.html.
HKL2000 (Denzo and Scalepack)	http://www.hkl-xray.com/.
D*Trek	http://www.msc.com/protein/dtrek.html.
XDS	http://www.mpimf-heidelberg.mpg.de/~kabsch/xds/.
SAINT	http://www.bruker-nonius.com/products/scd/saint.php.

5.1. Preliminary Work on the Diffraction Images

The modern electronic detectors previously described are, by themselves, not perfectly uniform in either their recording of the position of the diffraction spot or of its intensity. Almost always these nonuniformities are easily measured, are quite stable, and, therefore, are corrected automatically as the data are measured. If you do not know whether this is automatic, it will not hurt to ask. With some of the data-reduction software available to you, it will pay for you to determine which are the usable portions of the detector and which are not. You may need to put masks around the beam-stop shadow and possibly any cryostat shadows, and so on. In addition, one will need to know certain parameters of the experiment such as the wavelength of the radiation and the crystal-to-detector distance. Also one needs to know the main X-ray beam position to better than half the spot separation, otherwise the algorithm will assign the wrong *hkl* indices to the reflections.

5.2. "Indexing" the Diffraction Pattern

This step is really more than just determining the indices of the reflections, and it requires several steps. First, a program will be run to locate a few strong spots on each of one or more images. The user may need to adjust one or more parameters in the spot-finding process to get the right number: aim for about 100 per image. Then an "autoindexing" routine attempts to identify

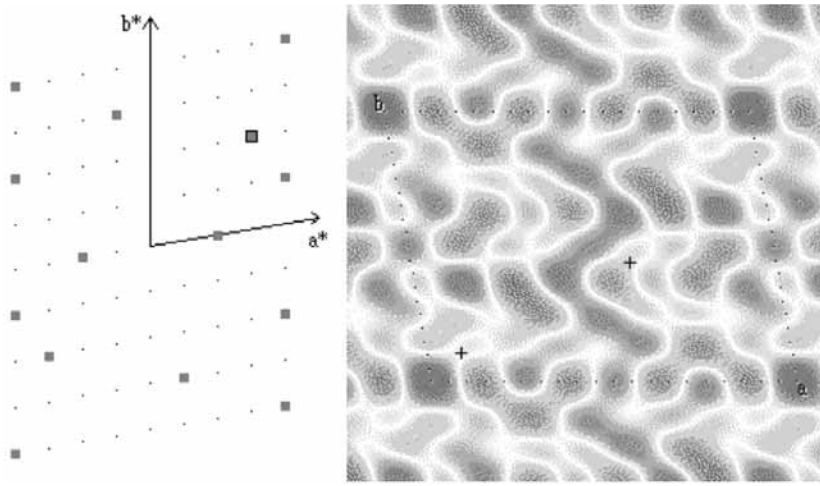


Fig. 17. The Fourier indexing method. On the left, diffraction spots; on the right, the cosine Fourier transform of the spots. The unit cell of the crystal appears in this map. (K. Cowtan, Structure-Factor Applet http://www.ysbl.york.ac.uk/~cowtan/sfapplet/sfintro.html.)

the indices of every reflection. The programs that perform this operation all work slightly differently, but fundamentally one can start with a group of reflections that are accurately located in the diffraction image (left panel in **Fig. 17**). Then one calculates a cosine Fourier transform of their positions. The resulting "density" map (right panel in **Fig. 17**) gives a clear indication of the crystal lattice orientation and dimension. If there are any regularities in the unit cell shape or symmetry, like perhaps a hexagonal lattice, the software will suggest this, and the user may choose a solution as judged from a "penalty" or "score." Once this has been done, and the positions of the spots predicted from the new indexing have been adjusted to match the actual positions (refinement), one is ready to integrate the diffraction pattern.

5.3. Integration of the Intensities

During integration the objective is to measure the total intensity of each reflection. **Figure 18** is a schematic view, in one dimension, of the scheme for integration. Typically there are some X-rays recorded on the area-sensitive detector that are not part of the diffraction spot; we call these "background." Usually we measure the intensity over an area that goes beyond the edges of the actual reflection itself. The integrated intensity we want is the area of the curve under the peak that is above the background. It is straightforward to extend the idea described in **Fig. 18** to a 2D box around the spot.

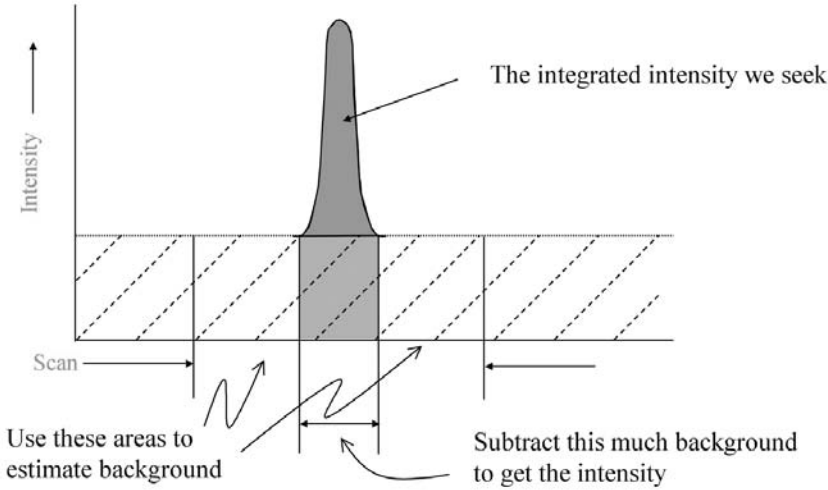


Fig. 18. Integration of diffraction peaks.

Typically, we use the refined parameters for the unit cell size and the crystal orientation to predict where to find each reflection on the detector. Then the program will compare the predicted to the observed position. If the fit is not good it will refine again, and then go on.

One can show that for weak reflections, it pays to fit the weak ones with a profile for the reflection that is learned from the strong ones. Intensity data are digitized from the detector along equal increments in the horizontal and vertical directions. These increments are typically the pixels of the CCD imager. Thinking in one dimension as in **Figs. 18** and **19**, we could name the intensity value at each pixel S_i and the background at each point B_i. The profile is "learned" by averaging values of the background-corrected intensity, S_i–B_i, over many strong reflections. This would give the average reflection profile, given by p_i. We normalize the learned profile so that $\Sigma p_i = 1$. Then using the variance of each measurement as V_i, (usually S_i), we have that the best estimate of intensity, I, is ***(32)***:

$$I = \frac{\sum \frac{p_i(S_i - B_i)}{V_i}}{\sum \frac{p_i^2}{V_i}}$$

Most data-reduction programs will integrate the peaks as illustrated in **Fig. 18**, and fit profiles, as in **Fig. 19**, then compare the two. Reflections will be rejected where the values are widely different.

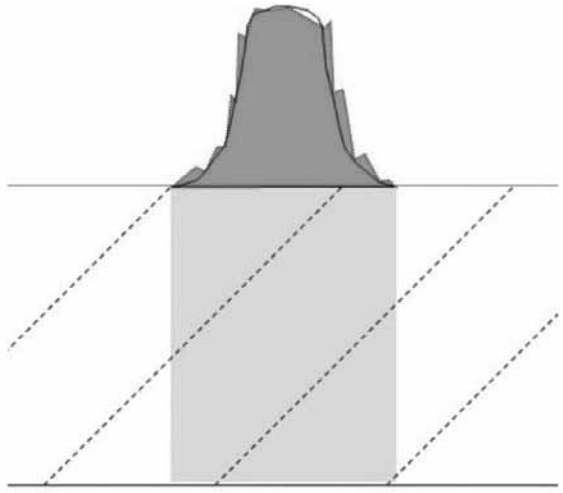


Fig. 19. One can show that for weak reflections, it pays to fit weak reflections with a profile that is learned from the strong reflections.

5.4. Scaling

These days, diffraction data are typically recorded on multiple diffraction images taken one after the other. These images may or may not be all on the same scale; we presume they are not and calculate scale factors. These scale factors are adjusted to make the residual discrepancy among the various measurements of the same reflection or ones related by the symmetry of the crystal as small as possible. The discrepancy indicator we use is typically an "*R*-value." This *R*-value will be of the form:

$$R = \frac{\sum \left| \langle I \rangle - I_j \right|}{\sum \langle I \rangle}$$

where <*I*> is the average intensity for each *h,k,l,* and I_j is the intensity after the scale factor is applied to a reflection measured on the *j*th image for the same *h,k,l.* The sum is over all observations, and the values of *R* are typically 0.05 for reasonably well-measured data.

We typically try to measure as many symmetry-related observations as possible. The ratio of observations to final scaled measurements we term the "redundancy" or "multiplicity." To some extent measuring data to high multiplicity mops up various systematic errors, which we should try to minimize anyway. These are absorption, X-ray beam fluctuations, mechanical errors from the instrument (detector sensitivity, crystal-axis errors, and so on). As part of

the scaling process, we take into account two geometric factors. The first is the "Lorentz" correction, which accounts for the variability in the time that each reflection spends in the diffracting condition. The second is the polarization correction, which accounts for the fact that any reflection of light imparts some degree of polarization on the reflected light.

5.5. General Processing Points

Just as it will pay the experimenter to think a little about the X-ray beam and its collimation, it will pay to spend a little time looking at what the data-reduction programs do to the data. Each program output is different, and generally the strongest indication of problems will not appear until the scaling program is run. Nonetheless, study the output of the integration program for signs of success or failure on each image. The most important indicators will be principally the residuals that result from the refinement of the spot positions to the calculated model. After the scaling step, one can scan the table of scale factors and *R*-value residuals for each image for aberrant values. The scale factors should mostly be similar for a single crystal, and should vary only a little from image to image. The *R*-value should be close to this range of 0.05, and all should be similar. If there are a few images with bad values, discard them and run the scaling again. If there are many and you do not know why, seek advice. Look especially for signs of radiation damage, where the *R*-value or the scale factor rise significantly toward the end of the data collection. If the data were collected with high redundancy (multiplicity), do not hesitate to throw out bad data.

Some of the programs allow examination of the "error model." As the individual reflections are integrated, the program will use some statistical tests to estimate the error in each, and a "standard deviation," often designated with the Greek character σ (sigma), is determined. At the end of scaling of the data the test to see if the σ's are estimated correctly is to evaluate the "goodness-of-fit," or χ^2 (chi-square) value. This is an expression similar to the *R*-value mentioned previously, but it has the σ's in it:

$$\chi^2 = \frac{\left(\langle I \rangle - I_j\right)^2}{\sigma^2}$$

The errors in the I_j values, should be approximately $(<I> - I_j)$ if the <I> values are nearly correct. So if the σ^2 values are close to these errors, the χ^2 must be close to 1.0. Sometimes values different from 1.0 tell you that there are real differences (such as, for example, the presence of anomalous scattering). Sometimes something else is wrong, like the value of the conversion from intensity on the detector (the gain) to X-ray photons is wrong (*see* **Note 6**).

6. Conclusions

Currently, researchers interested in developing high-throughput "pipelines" for structure solution are devoting considerable energy to the automation of the data collection steps previously described. Robots that can transfer a crystal from a liquid nitrogen Dewar onto the goniometer head are now being tested at most synchrotrons. The crystal (or at the moment the cryoloop supporting the crystal) will then be aligned under the control of the automated software for initial testing and then the best crystal, selected on the basis of various criteria set in the diffraction image processing software, will be used for data collection. These automatic pipelines will soon have the capability of steering a structure solution to the point of producing an electron density map.

Despite the rapid progress of these developments, there will always be a vital role for those experimenters who understand the detail hidden in the "black boxes" and who can use this knowledge to tackle problematic and nonstandard cases that are not amenable to predetermined pipelines.

Finally, when one is collecting data, attention to detail and a basic understanding of the underlying principles are always worthwhile: they can make the difference between successful structure solution and failure.

7. Notes

1. There is another type of patented X-ray source known as a "MicroSource" that employs a tiny (0.02-mm diameter) electron-beam focus on a water-cooled anode to produce a surprisingly bright source ***(2)***. The whole source fits on a desktop, and instead of putting several kilowatts of power into the anode, it puts, typically, 25 W. It is a low maintenance, moderate brightness source that is suitable for some experiments.
2. If a crystal does not diffract when cryocooled, it is well worth carrying out a room temperature test to find out whether the crystal has no intrinsic diffracting power or if the order was destroyed by the cryocooling protocol.
3. If a dataset is incomplete because of overloads, the reflections can be rescued by profile fitting their tails to an estimated peak height. This procedure is clearly second best, and it is much better not to collect overloads in the first place.
4. Make sure you record the wavelength in case it is not automatically entered into the image header.
5. Other ways to improve the quality of diffraction images are to (1) choose a limiting aperture to match the crystal size, (2) center the beamstop properly to avoid asymmetric scattering into the background, and probably larger than necessary background, (3) minimize the quantity of vitreous cryobuffer around the crystal: one way to accomplish this is to match the size of the loop to the crystal.
6. Experienced experimenters will tell you that for the same detector and collimation system, the χ^2 value will always be near to 1.0 unless there is something critically wrong.

References

1. Wonacott, A. J. (1977) Geometry of the rotation method. In: *The Rotation Method in Crystallography,* (Arndt, U. W. and Wonacott, A. J., eds.), North-Holland Publ., Amsterdam, The Netherlands, pp. 77–103.
2. Arndt, U. W., Long, J. V. P., and Duncomb, P. (1998) A microfocus X-ray tube used with focusing collimators. *J. Appl. Cryst.* **31,** 936–944.
3. Gouet, P., Diprose, J. M., Grimes, J. M., et al. (1999) The highly ordered double-stranded RNA genome of bluetongue virus revealed by crystallography. *Cell* **97,** 481–490.
4. Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A.(2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289,** 905.
5. Yang, C., Courville, A., and Ferrara, J. (1999) Optics systems for the home laboratory: *caveat emptor*. *Acta Cryst.* **D55,** 1681–1689.
6. Arndt, U. W. (1982) X-ray television area detectors. *Nucl. Instrum. Methods* **201,** 13–20.
7. Arndt, U. W. and Thomas, D. J. (1982) High-speed single crystal television X-ray diffractometer (Hardware). *Nucl. Instrum. Methods* **201,** 21–25.
8. Hamlin, R., Cork, C., Howard, A., et al. (1981) Characteristics of a flat multiwire area detector for protein crystallography. *J. Appl. Cryst*. **14,** 85–93.
9. Durbin, R. M., Burns, R., Moulai, J., et al. (1986) Protein, DNA, and virus crystallography with a focused imaging proportional counter. *Science* **232,** 1127–1132.
10. Howard, A. J., Gilliland, G. L., Finzel, B. C., Poulos, T. L., Ohlendorf, D. H., and Salemme, F. R. (1987) The use of an imaging proportional counter in macromolecular crystallography. *J. Appl. Cryst*. **20,** 383–387.
11. Kahn, R. and Fourme, R. (1997) Gas proportional detectors. In: *Methods in Enzymology, Vol. 276,* (Carter, C. W., Jr. and Sweet, R. M., eds.), Academic Press, San Diego, CA, pp. 268–285.
12. Miyahara, J., Takahashi, K., Amemiya, Y., Kamiya, N., and Satow, Y. (1986) A new type of X-ray area detector utilising laser stimulated luminescence. *Nucl. Instrum. Methods* **A246,** 572–578.
13. Amemiya, Y. (1997) X-ray storage-phosphor imaging-plate detectors: high sensitivity X-ray detector. In: *Methods in Enzymology, Vol. 276,* (Carter, C. W., Jr., and Sweet, R. M., eds.), Academic Press, San Diego, CA, pp. 233–243.
14. Westbrook, E .M. and Naday, I. (1997) Charge-coupled device based area detectors. In: *Methods in Enzymology, Vol. 276* (Carter, C. W., Jr., and Sweet, R. M., eds.), Academic Press, San Diego, CA, pp. 244–267.
15. Phillips, W. C., Stanton, M., Stewart, A., Qian, H., Ingersoll, C., and Sweet, R. M. (2000) Multiple CCD detector for macromolecular crystallography. *J. Appl. Cryst.* **33,** 243–251.
16. Garman, E. F. and Schneider, T. R. (1997) Macromolecular cryocrystallography. *J. Appl. Cryst.* **30,** 211–237.
17. Rodgers, D. W. (1997) Practical cryocrystallography. In: *Methods in Enzymology, Vol. 276,* (Carter, C. W., Jr. and Sweet, R. M., eds.), Academic Press, San Diego, CA, pp. 183–202.

18. Garman, E. (1999) Cool data: quality and quantity. *Acta Cryst*. **D55,** 1641–1653.
19. Mitchell, E. M. and Garman, E. F. (1994) Flash freezing of protein crystals: investigation of mosaic spread diffraction limit with cryoprotectant concentration. *J. Appl. Cryst* **27**, 1070–1074.
20. Leslie, A. G. W. (1999) Integration of macromolecular diffraction data. *Acta Cryst.* **D55,** 1696–1702.
21. Noble, M. E. M. (1996) Sofware package PREDICT, Available at http://biop.ox.ac.uk/www/distrib/predict.html. Last accessed June 4, 2006.
22. Kabsch, W. (1988) Evaluation of single-crystal X-ray diffraction data from a Position-Sensitive Detector. *J. Appl. Cryst.* **21,** 916–924.
23. Kabsch, W. (1988) Automatic indexing of rotation diffraction patterns. *J. Appl. Cryst.* **21,** 67–71.
24. Kabsch, W. (1993) Automatic processing of rotation diffraction data from crystals of initially unknown cell constants. *J. Appl. Cryst.* **26,** 795–800.
25. Ravelli, R. G. B., Sweet, R. M., Skinner, J. M., Duisenberg, A. J. M., and Kroon, J. (1997) STRATEGY: a program to optimise the starting spindle angle and scan range for X-ray data collection. *J. Appl. Cryst.* **30,** 551–554.
26. Otwinowski, Z. and Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode. In: *Methods in Enzymology, Vol. 276,* (Carter, C. W., Jr. and Sweet, R. M., eds.), Academic Press, San Diego, CA, pp. 307–326.
27. Dauter, Z. (1997) Data collection strategy. In: *Methods in Enzymology, Vol. 276,* (Carter, C. W., Jr. and Sweet, R. M., eds.), Academic Press, San Diego, CA, pp. 326–343.
28. Pflugrath, J. W. (1999) The finer things in X-ray diffraction data collection. *Acta Cryst.* **D55,** 1718–1725.
29. Garman, E. (2003) 'Cool' crystals: cryocrystallography and radiation damage. *Curr. Opin. Struct. Biol.* **13,** 545–551.
30. Hendrickson, W. and Teeter, M. (1981) Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulfur. *Nature* **290,** 107–113.
31. Biou, V., Shu, F., and Ramikrishnan, V. (1995) X-ray crystallography shows that translational initiation factor IF3 consists of two compact alpha/beta domains linked by an alpha-helix. *EMBO J*. **14,** 4056–4064.
32. Rossmann, M.G. (1979) Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting. *J. Appl. Cryst.* **12,** 225–238.

6

Characterizing a Crystal From an Initial Native Dataset

Michael R. Sawaya

Summary

Methods are presented for characterizing a crystal given an initial X-ray diffraction dataset. These methods can facilitate the structure determination process and illuminate the oligomeric state and symmetry of your molecule before the crystal structure is determined. Specifically, these methods include (1) calculation of Matthews coefficient to estimate the number of molecules in the asymmetric unit; (2) calculation and interpretation of a self-rotation function to evaluate the point group symmetry of the crystallized oligomer, if contained in the crystal; (3) calculation and interpretation of a native Patterson map to evaluate the presence of noncrystallographic translational symmetry; and (4) calculation of statistics to evaluate the possibility of merohedral twinning in a crystal.

Key Words: Self-rotation function; native Patterson map; merohedral twinning; noncrystallographic symmetry; Matthew's coefficient.

1. Introduction

In the race to obtain phases for an initial set of X-ray diffraction intensities, it is tempting to limit crystal characterization to the bare minimum: unit cell, space group, and resolution limits. However, it often pays to take time to acquaint yourself fully with the properties of your native crystal. Detection of noncrystallographic symmetry (NCS) elements and merohedral twinning (if present) provide additional information to guide and speed the structure determination process. Both characterizations are simple to perform and require just a native dataset.

2. Significance of NCS

NCS is a commonly occurring feature in macromolecular crystals. Approximately 27% of the unique crystal forms in the PDB contain proper n-fold symmetry in the asymmetric unit *(1)*. Many others contain improper or translation symmetry. Knowing the number of copies and symmetrical arrangement of

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

these copies can be an enormous advantage in structure determination, either by facilitating molecular replacement (NCS locked rotation and translation functions) or map interpretation (improvement of electron density maps through NCS symmetry averaging). Furthermore, clues about the molecule's natural oligomeric state can be discovered from analysis of NCS symmetry because crystals often capture the symmetry of naturally occurring complexes in their NCS elements. This section describes the use of the Matthews coefficient, the native Patterson map, and the self-rotation function to characterize the contents and arrangement of macromolecules in a crystal once having obtained an X-ray diffraction dataset.

2.1. Estimating the Number of Molecules in the Asymmetric Unit With V_M

The number of molecules in the asymmetric unit (n) can be readily estimated using a method developed by Matthews *(2)*. His survey of 116 protein crystals revealed that the v/m ratio of the asymmetric unit (V_M) typically ranges between 2.0 to 3.0 $Å^3$/Da (corresponding to a solvent content of 40–60% by volume). He hypothesized that this range of V_M is characteristic of the majority of protein crystals and, thus, could be used as a guide to estimate n for any protein crystal given its unit cell volume and protein mass. Specifically, if one calculates V_M as a function of n, one can assume that the most probable n corresponds to the value that produces a V_M in the range 2.0 to 3.0 $Å^3$/Da. V_M is called the Matthews coefficient and is directly related to the solvent content in its most general form:

$$V_M = \frac{a*b*c*(1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma+2\cos\alpha*\cos\beta*\cos\gamma)^{1/2}}{M_r*Z*n} = \frac{\text{volume of unit cell}}{\text{mass of unit cell}}$$

M_r corresponds to the molecular weight of the protein molecule specified in Daltons and Z is the number of asymmetric units per unit cell (*see* **Note 1**). The fractional volume occupied by the solvent can then be derived from V_M by the equation: $V_{solv} = 1 - (1.23/V_M)$.

2.1.1. Calculating V_M

1. Calculate V_M as a function of n beginning with $n = 1$. This may be done with a calculator using the equation above. Alternatively, V_m can be calculated with the matthews_coef program from the CCP4 suite of programs (**Fig. 1**) *(3)*. The CCP4 suite is the cornerstone of crystallographic software. It may be obtained from http://www.ccp4.ac.uk/. Another convenient method is to simply enter your unit cell parameters, molecular weight, and space group in the V_M server at http://www.rigaku.co.jp/xrl/group1/tips/prg/vm.html or http://www-structure.llnl.gov/mattprob/ *(4)*.
2. Repeat the V_M calculation for all n that produce a V_M in the range from 1.6 to 3.5 $Å^3$/Da. It is helpful to enter these values in a table with two columns, n and V_M.
3. Evaluate results.

A
```
matthews_coef << eof
CELL 79.1 79.1 37.9 90. 90. 90.
SYMM P43212
MOLWEIGHT 14296
AUTO
eof
```

B

Nmol/asym	Matthews Coeff	%solvent
1	2.1	40.2

Fig. 1. (A) Input script and **(B)** output table from CCP4's matthews_coef program. Example taken from hen egg white lysozyme crystals.

2.1.2. Evaluation of V_M Results

In the best case scenario, only one value of *n* will produce a V_M in the range from 2 to 3 Å^3/Da. Generally, this value of *n* would be quite reliable. However, the exact value of *n* can become ambiguous as *n* gets larger. In this scenario you may find multiple values of *n* will produce a V_M in the 2–3 Å^3/Da range. There are several things you can do to help narrow down the choice of *n*.

1. Consider that crystals that diffract to high resolution (better than 2.0 Å) typically have a lower V_M (denser packing) than crystals that diffract poorly. Thus, a higher *n* is more probable for crystals that diffract to higher resolution and lower *n* is more probable for crystals that diffract to lower resolution ***(4)***.
2. Check for consistency with the self-rotation function. For example, if V_M suggests that *n* = 4, 5, or 6 are all possible, a peak on the κ =72° section, would suggest *n* = 5 is the correct choice (360°/5=72°) (*see* **Subheading 2.3.**).
3. Perform a crystal density measurement. A direct measurement of the crystal density might help narrow the range of possible *n* ***(5–7)***.
4. Determine the oligomeric state of the protein in solution—it is likely that this oligomeric state will be trapped in the crystal. Indications from native gel electrophoresis, light scattering, or sedimentation velocity experiments can help pinpoint the oligomeric state. In favorable cases, the number of subunits may even be counted from an electron micrograph.
5. If you find that *n* = 1/2 is the only value of *n* that produces a reasonable V_M, twinning is likely indicated (*see* **Subheading 3.2.**).

2.2. The Native Patterson Map Reveals the Presence of Pure Translational Symmetry

The native Patterson map reveals the presence of pure translational symmetry elements in a crystal. Recall that the native Patterson map contains peaks corresponding to interatomic vectors *(8)*. The Patterson function is given by the following equation.

$$P(u,v,w)=1/v \sum_{h}\sum_{k}\sum_{l}|F_{hkl}|^2 |e^{-2\pi i(hu+kv+lw)}$$

If one were to draw all possible vectors between atoms in one unit cell, then translate those vectors so that their tails lie at the origin, the heads of the vectors would correspond to the position of the peaks in the Patterson map. In the general case, these interatomic vectors would have a variety of lengths and directions producing scattered distribution of low peaks in the Patterson map. But, when two protein molecules are related by a translation, the interatomic vectors between equivalent atoms in the two molecules would all have the same length and direction, summing up to an exceedingly intense peak in the map. The position of the peak describes the vector between the pair of molecules that produced it (*see* illustration at http://www-structure.llnl.gov/xray/Patterson/Native_Patterson.htm).

2.2.1. Patterson Map Calculation

The native Patterson map may be calculated with FFT then viewed with mapslicer and analyzed with peakmax, all programs from the CCP4 suite (**Fig. 2**). XtalView offers another route; one can produce a map using xfft then view the map with xcontur.

2.2.2. Interpretation of the Patterson Map

The presence of pure translational symmetry elements in a crystal is indicated by the appearance of exceedingly intense peaks in the native Patterson map. Look for off-origin peaks greater than 5σ. Usually, peaks resulting from translational symmetry are greater than 10σ above the mean (**Fig. 2B**). **Caution:** native Patterson maps in nonprimitive cells (C, I, F) will contain strong peaks at coordinates corresponding to their centering operations; these peaks should not be considered as noncrystallographic translations.

The presence of translational symmetry allows for some simplifications in molecular replacement. Two molecules related by translational symmetry will share the same cross-rotation function solution (because they have the same orientation). They will be related by a vector having the same length and orientation as the vector having its tail at the origin and head at the coordinate of the Patterson peak. Generally, translational NCS is easier to recognize in electron

A

```
#PATTERSON CALCULATION
#
fft HKLIN mydata.mtz MAPOUT patterson.map<<END >patterson.log
PATTERSON
RESOLUTION 20 3.0
FFTSPACEGROUP 2   ! Original space group + no trans + center of sym
TITLE 20-3.0A Native Patterson Map
LABIN F1=F SIG1=sigF
END

#PEAKSEARCH
#
peakmax MAPIN patterson.map  << END >patpeak.log
THRESHOLD RMS 2.5
OUTPUT PEAKS
END
```

B

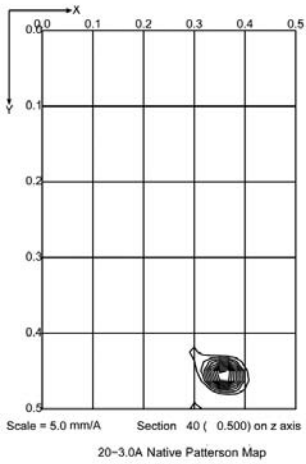


Fig. 2. (A) Input script and **(B)** output native Patterson map from CCP4's fft program. The peakmax script generates a list of peaks higher than a specified threshold. The map is viewed with CCP4's mapslicer program.

density maps because the protein–solvent boundaries will have the same shape in the related molecules.

2.3. The Self-Rotation Function Specifies Oligomeric Symmetry and Orientation

The self-rotation function captures the orientational relationships between molecules in the unit cell. Specifically, it reveals (1) the presence of all rotational and screw symmetry axes in the unit cell, (2) the orientations of these

axes, and (3) the degree of rotation about these axes required to bring a pair of symmetry-related copies into coincidence. This information is implied from an analysis of the rotational symmetry of the Patterson map; each rotation or screw axes in the crystal produces a corresponding rotational symmetry axes in the Patterson map (*see* **Note 2**). The self-rotation function reveals the rotational symmetry of the Patterson map by comparing two identical, but misoriented, copies of the native Patterson map. The two copies of the map can be brought into coincidence for certain misorientation angles corresponding to rotational symmetry of the crystal. To search for these angles, one copy of the map is held stationary while the other is rotated about its origin throughout the unique part of rotation space. Rotation space is specified as intervals of three angles, usually (ϕ,ψ,κ) in the spherical polar coordinate system, where ϕ and ψ specify the orientation of the rotation/screw symmetry axis, and κ specifies the value of the rotation required to bring the symmetry-related copies into coincidence. The Patterson maps are overlaid and integrated within a specified radius of the origin (radius of integration) and the value of the integration is written to a map.

$$R(\phi,\psi,\kappa)=\int_U P(\mathbf{u}) \times P_r(\mathbf{u}_r)d\mathbf{u}$$

$P(u)$ is the Patterson function at point u. $P_r(u_r)$ is the rotated version. U is the volume of the sphere of the Patterson contained by the radius of integration. Rotation angles that bring the Patterson maps into coincidence (high value of integration) correspond to peaks in the rotation function map. This map is conveniently plotted in sections of κ. For example, peaks on the κ=180° section indicate the presence of all the twofold rotation and screw axes, peaks on the κ=120° section indicate the presence of all threefold, and so on. Longitude and latitude lines mark divisions of ϕ and ψ, respectively. If you think of each κ section as a spherical projection (analogous to a two-dimensional [2D] projection of a hemisphere of the earth), peaks mark the locations where rotational symmetry axes enter and exit the sphere. All axes intersect at the center of the sphere. The orientations of the axes in the sphere correspond to the orientation of the n-fold symmetry axes in the crystal. The position of the n-fold symmetry axes in the crystal is *not* specified.

2.3.1. Self-Rotation Function Calculation

Several modern programs calculate the self-rotation function. Among these are AMoRe ***(9)***, MOLREP ***(10),*** polarrfn ***(3)*** (all part of CCP4 suite), CNS ***(11)***, and GLRF ***(12)***. GLRF is featured in the example in **Fig. 3** for the following reasons: it is easy to use; it can be run directly on Scalepack output (or any ascii format); coordinate system conventions can be explicitly chosen; the option of origin removal is provided; the option of fast or slow rotation functions is provided; the output is easy to interpret; relevant parameters are conveniently print-

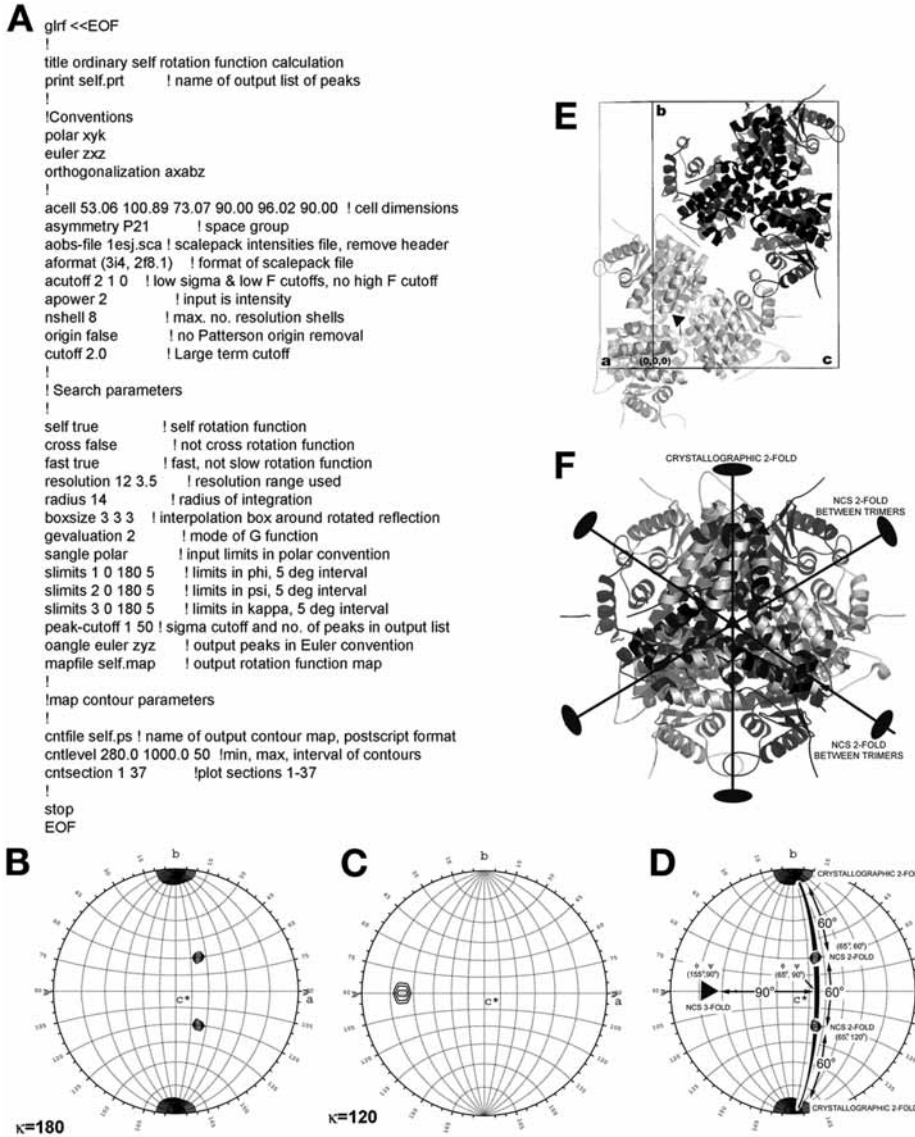
ed on each plot; and lastly, it is part of the "Replace" suite of programs and so provides a direct route to the locked rotation and translation functions that aids molecular replacement. GLRF and may be obtained by sending a request by e-mail to Liang Tong, tong@como.bio.columbia.edu.

There are two variables of the rotation function that strongly affect its appearance, (1) the radius of integration and (2) the resolution range. The radius of integration, as previously mentioned, excludes from the self-rotation function all points further than distance *r* from the origin of the Patterson map. Generally, self-rotation function peaks become sharper and more significant with a longer *r* because a larger volume of the Patterson map is used in the integration. Indeed, if *r* is too short, then the self-rotation function peaks will appear low and broad. But, if *r* is too long, the self-rotation functions peaks will diminish in height because Patterson cross vectors (vectors between different molecules) will be included in the integration, obscuring the rotational symmetry evident in the pattern of Patterson self vectors (vectors within the same molecule). When chosen judiciously, *r* serves as the boundary, excluding cross vectors and retaining mainly self vectors in the integration. A rule of thumb is to begin with a radius equal to the expected diameter of the molecule. However, one must always sample different radii of integration to find the radius that produces the clearest features in the self-rotation function.

The resolution range used in the self-rotation function refers to the resolution range used to calculate the Patterson map. Both high- and low-resolution limits must be optimized to give the clearest features possible. A high-resolution limit anywhere between 3 and 5 Å is optimal. High-resolution terms tend to sharpen the peaks, but if significant conformational differences exist between molecules then these terms will produce noise. In the later case, one may need to reduce the high-resolution limit. Low-resolution terms (lower than 15 Å) are often excluded because they are insensitive to rotation. These terms dominate the rotation function, reducing the level of detail in the self-rotation function. For the lower resolution limit, test a range between 12 and 8 Å.

2.3.2. Interpretation of the Self-Rotation Function

The immediate goal of interpreting the rotation function is to build a working model of homo-oligomeric assembly. Once a hypothesis about the NCS of the assembly is obtained, it can be used to help solve a structure. This potential was realized by Rossmann and Blow in the infancy of the crystallographic enterprise ***(13,14)*** though it was not until much later that techniques were developed to take advantage of it ***(15)***. Here is outlined a procedure for interpreting the self-rotation function map. Because biological assemblies often exist in certain point group symmetries ***(16–18)*** attention is focused on finding a set of peaks in the self-rotation function consistent with one of these point groups.



1. A plausible model for the point group symmetry of the biological assembly must be compatible with the number of molecules in the unit cell and/or space group symmetry. Many point groups can be eliminated as implausible even before calculating the self-rotation function.
   a. The number of molecules in the point group ($N_{point\ group}$) of the biological assembly cannot exceed the number of molecules in the unit cell ($N_{unit\ cell}$). Calculate $N_{unit\ cell}$ by multiplying the number of molecules in the asymmetric

Fig. 3. The self-rotation function calculated, displayed, and interpreted for Protein Data Bank (PDB) entry 1esj, thiazole kinase ***(36)***. **(A)** The self-rotation function script input to the program GLRF. Input structure factors were downloaded from the PDB. **(B)** $\kappa = 180°$ section illustrating the orientation of three twofold axes. **(C)** $\kappa = 120°$ section illustrating the orientation of a single threefold axis. **(D)** Superimposition of the two sections illustrating that the three twofold axes lie in a plane perpendicular to the threefold axis. This distinctive pattern of peaks suggests either point group C_3 or D_3. However, point group D_3 can confidently be ruled out given the clear indication from Matthews coefficient (2.1 $Å^3$/Da) that there are only three molecules in the asymmetric unit of this cell ($N_{point\ group} > n_{asu}$, 6>3) and $P2_1$ contains no pure rotational twofold axis that could contribute to D_3 (*see* **Subheading 2.3.2.**, **item 1b**). The appearance of NCS twofold result from the relationship between the two trimers of the unit cell. **(E)** A depiction of the unit cell packing for 1esj, viewed down the NCS threefold of the trimer. **(F)** An exercise to show how the NCS twofold are generated by the combination of a threefold NCS axis perpendicular to the crystallographic 2_1 screw. The appearance of the NCS twofold becomes obvious after translating both trimers to the origin (to mimic the loss of the translational information inherent in the self-rotation function). When the twofold axes are drawn, you can see that each twofold is 60° from its neighboring twofold and that they all lie in a plane that is perpendicular to the NCS threefold. You may perform this exercise on any crystal to rationalize the presence of unanticipated NCS peaks owing to the operation of crystallographic symmetry operators on the NCS operators of the biological assembly (*see* **Subheading 2.3.2.**, **item 5c**).

unit (n_{asu}) (**Subheading 2.1.**) by the number of symmetry operators in the crystal's space group (Z) (i.e., $N_{unit\ cell} = n_{asu} * Z$). Point groups with $N_{point\ group} > N_{unit\ cell}$ may be eliminated as implausible models. For example, if $N_{unit\ cell} = 4$, then point groups with $N_{point\ group} > 4$ are implausible (*see* column $N_{point\ group}$ in **Table 1**). The only possible choices remaining are C_2, C_3, D_2, and C_4. These choices may be further narrowed down when $N_{point\ group} > n_{asu}$ (*see* **item 1b**).

b. $N_{point\ group} > n_{asu}$ is plausible only when one (or more) symmetry axis of the point group coincides with crystallographic rotation operator(s). If the space group lacks a pure rotational symmetry operator that can be used as a component of the point group assembly, then the point group can be eliminated as an implausible model. Continuing with the example above, a biological assembly with point group C_2 ($N_{point\ group} = 2$) is plausible in space group $P2_12_12$ under conditions where $n_{asu} = 1$ (i.e., $N_{point\ group} > n_{asu}$) only because $P2_12_12$ contains a crystallographic twofold axis that can theoretically be used as a component of the point group symmetry. But, given the same n_{asu}, a biological assembly with point group C_2 is impossible in space group $P2_12_12_1$ because it has no pure rotation axis to contribute to the point group assembly.

c. $N_{point\ group} = n_{asu}$ imposes no requirement for pure rotational symmetry operators. Continuing with the example above, point group C_2 ($N_{point\ group}$=2) is plausible

Table 1
Self-Rotation Function Peak Patterns for 12 Point Groups[a]

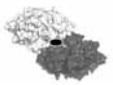
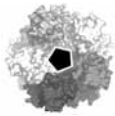
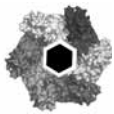
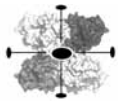
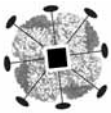
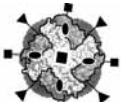
			Number of axes indicated on κ section				
	Point group	$N_{point\ group}$	180°	120°	90°	72°	60°
 1d1s	2 (C_2)	2	1				
 1esj	3 (C_3)	3		1			
 6nn9	4 (C_4)	4	1		1 (same ϕ, ψ as peak on κ = 180°)		
 1vps	5 (C_5)	5				1	
 1d2n	6 (C_6)	6	1	1 (same ϕ, ψ as peak on κ = 180°)			1 (same ϕ, ψ as peak on κ=180°)
 4pgm	222 (D_2)	4	3 (mutually ⊥)				
 1kqn	32 (D_3)	6	6 (coplanar, 60° apart)	1 (⊥ to κ = 180° peaks)			
 1qlt	422 (D_4)	8	8 (coplanar, 45° apart)		1 (⊥ to κ = 180° peaks)		

Table continues

Table 1 (continued)

	Point group	$N_{point\ group}$	Number of axes indicated on κ section				
			180°	120°	90°	72°	60°
 1di0	52 (D_5)	10	10 (coplanar, 36° apart)			1 (⊥ to κ= 180° peaks)	
 1hto	622 (D_6)	12	12 (coplanar, 30° apart)				1 (⊥ to κ= 180° peaks)
 1dps	23 (T)	12	3 (mutually ⊥)	4 (109.5° apart)			
 1hrs	432 (O)	24	9 (in 3 mutually ⊥ planes)	4 (109.5° apart)	3 (mutually ⊥)		

[a]Coordinates were obtained from the Protein Quaternary Structure database at http://pqs.ebi.ac.uk. PDB ID codes in the left column refer to the following proteins: 1d1s, alcohol dehydrogenase *(35)*; 3eca, thiazole kinase *(36)*; 6nn9, neuraminidase *(37)*; 1vps, virus coat protein *(38)*; 1d2n, hexamerization domain of *N*-ethylmaleimide-sensitive fusion protein *(39)*; 4pgm, phosphoglycerate mutase *(40)*; 1kqn, NMN/NAMN adenylyltransferase *(41)*; 1qlt, vanillyl-alcohol oxidase *(42)*; 1di0, lumazine synthase *(43)*; 1hto, glutamine synthetase *(44)*; 1dps, Dps protein *(45)*; 1h, ferritin *(46)*.

in space group $P2_1$ (n_{asu} = 2) despite the fact that space group $P2_1$ contains no pure rotational symmetry elements.

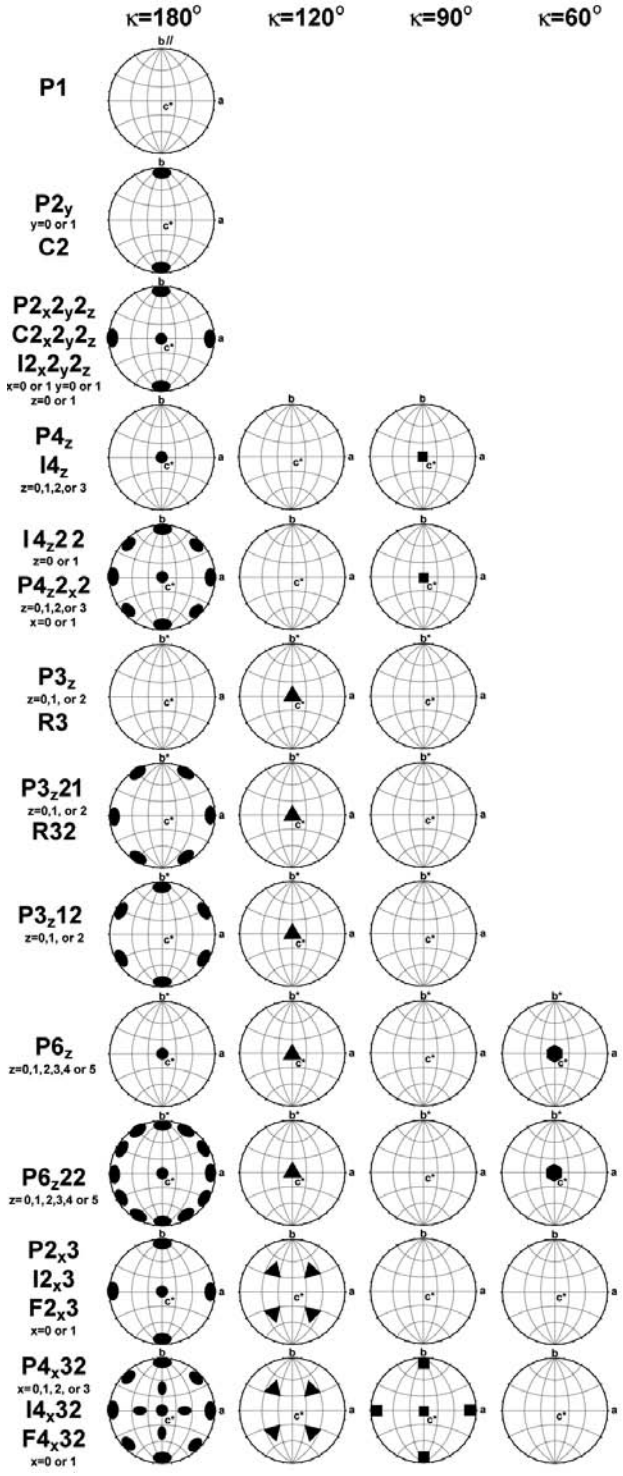
d. $N_{point\ group} < n_{asu}$ also imposes no requirement for pure rotational symmetry operators. Under these conditions, one could expect multiple biological assemblies in the asymmetric unit. The number of biological assemblies need not be an integer providing that the crystallographic rotation operators play a role in generating one of the biological assemblies.

2. Label all crystallographic rotational/screw symmetry axes. To determine whether NCS rotational symmetry is present, one should first account for all the peaks arising from crystallographic symmetry. Both NCS and crystallographic rotational/screw symmetry produce peaks in the rotation function map; the only distinction is that peaks arising from crystallographic symmetry are located at positions dictated by the space group conventions, whereas peaks arising from NCS can appear any-

where in the rotation function map. If you are not familiar with the rotational symmetry of your space group, refer to **Fig. 4**, which shows the 11 Laue symmetries and the expected location of crystallographic rotation/screw axes. All space groups within the same Laue group have a common set of peaks in the self-rotation function owing to crystallographic symmetry.

3. Among the crystallographic axes previously labeled, distinguish between pure rotation and screw with an additional annotation. If you know the space group of your crystal then you are able to distinguish which crystallographic axes are pure rotation and which are screw axes. Such information cannot be derived from the self-rotation function itself, but only by careful examination of the systematic absences in your dataset (*see* **Note 2**). This additional information will become helpful in eliminating incorrect choices of point group symmetry of the biological assembly because screw axes cannot participate in point group symmetry (**step 1b–d**).
4. Label the (ϕ,ψ) coordinates of the NCS rotational/screw symmetry axes. What is not crystallographic is noncrystallographic. If there are no peaks other than those arising from crystallographic symmetry, then either no NCS rotation/screw axis exists or there is a NCS rotation/screw parallel to the crystallographic rotation or screw axes. If the latter is true, then you should see a peak on the native Patterson map (**Subheading 2.2.**).
5. Recognize point group symmetry. Examine the self-rotation function for a set of NCS peaks that match the criteria for any of the 12 point groups listed in **Table 1**. Recognition of the point group symmetry of your biological assembly could be simple or complex depending on the clarity of the self-rotation function and the number of symmetry operators involved. To aid you in the recognition process, consider the following:
   a. Cyclic point groups (C_n), require that the peak must be located at a specific κ section, though there is no restriction on the ϕ, ψ value (**Table 1**). Cyclic point groups 2 (C_2) and 3 (C_3) are common, 4 (C_4) and 6 (C_6) are rare.
   b. Dihedral, tetrahedral, and octahedral point groups (D_n, T, O) all contain more than one symmetry axis. These axes maintain relative orientations specified in **Table 1**. For example, in order for a set of NCS two- and threefold peaks to meet the requirements of point group 32 (D_3), the twofold must be in a plane, and this plane must be perpendicular to the threefold axis. To test this relationship graphically, go to the $\kappa = 180°$ section and draw a "great circle" connecting the peaks corresponding to the three twofold axes (*see* **Note 3**). If the twofold do not lie in a great circle (you can judge this by eye) then they do not meet the first requirement of dihedral (D_n) symmetry. Next, the plane containing these twofold must be shown to be normal to the threefold axes. Draw a small triangle symbol on the $\kappa = 180°$ section marking the (ϕ,ψ) position of the NCS three-

Fig. 4. Self-rotation function peaks attributed to crystallographic symmetry for the biological space groups.


κ=180°
κ=120°
κ=90°
κ=60°
P1
P2y
y=0 or 1
C2
P2x2y2z
C2x2y2z
I2x2y2z
x=0 or 1 y=0 or 1
z=0 or 1
P4z
I4z
z=0,1,2,or 3
I4z22
z=0 or 1
P4z2x2
z=0,1,2,or 3
x=0 or 1
P3z
z=0,1, or 2
R3
P3z21
z=0,1, or 2
R32
P3z12
z=0,1, or 2
P6z
z=0,1,2,3,4 or 5
P6z22
z=0,1,2,3,4 or 5
P2x3
I2x3
F2x3
x=0 or 1
P4x32
x=0,1, 2, or 3
I4x32
F4x32
x=0 or 1

fold axis. Draw a latitude or longitude curve that connects the threefold with the curve. The length of this connecting curve should correspond to 90° (i.e., if you subtract the [ϕ,ψ] coordinates of the two endpoints of the curve, ϕ should be 90° and ψ should be 0°, or vice versa) (**Fig 3B–D**).

c. Recognize that multiple sets of NCS axes are generated by crystallographic operators. The self-rotation function depicts all rotational symmetry axes in the unit cell, not just the asymmetric unit. Crystallographic symmetry operators act on the biological assembly in the asymmetric unit producing multiple orientations of the biological assembly. Each orientation of the biological assembly will have its own (set of) NCS peak(s), which can lend a crowded and confusing appearance to the self-rotation function, especially for the higher symmetry biological assemblies in high-symmetry space groups. Crystallographic symmetry, in addition, produces NCS axes between biological assemblies in the unit cell. In general, a biological assembly having m-fold symmetry placed in a crystal having an n-fold crystallographic rotation axis would produce n m-fold and m n-fold symmetry axes (although there might appear to be fewer peaks in the self-rotation function in the case where these operations produces axes that are parallel). These additional peaks between biological assemblies could be confused as being part of the biological assembly. Hence, it is easier to interpret NCS symmetry if the molecule crystallizes in a lower symmetry space group. If you can identify one set of NCS peaks as belonging to a certain point group characteristic of the biological assembly, you can often rationalize how the crystallographic rotation operators would produce the remaining peaks in the self-rotation function (**Fig. 3E,F**). It sometimes helps the deduction process to color the sets of peaks according to the biological assembly, which produced them. You can test your hypothesis for the symmetry of the biological assembly, by building a model with the proposed point group symmetry, calculating structure factors, and then run a self-rotation function on the calculated data. You should be able to reproduce the features of the experimental self-rotation function map. The molecule used in the model is irrelevant, as long as it does not contain any rotational symmetry itself.

6. Start by searching for the highest allowable point group according to **step 1**. But, be aware that observation of a high n-fold peak does not necessarily mean that the biological assembly contains n-fold symmetry (recall **step 1d**). Instead, a high n-fold peak could be the result of multiple independent biological assemblies of lower point group symmetry having a special relative orientation. For example, two independent C_3 trimers in the asymmetric unit aligned along their threefold axes would produce a peak at $\kappa = 40°$ if one trimer were rotated with respect to the other trimer by 40° about the NCS threefold axis. The indication of ninefold symmetry by the $\kappa = 40°$ peak would be misleading in this case.
7. Consider improper rotational symmetry. Finding a set of peaks that meet the criteria for a certain point group is necessary, but not sufficient, to conclude its presence. There are two types of ambiguities that cannot be resolved from the self-rotation function alone. (1) Proper rotation cannot be distinguished from screw axis symmetry. For example, a peak at $\kappa = 60$ could mean a sixfold ring or a sixfold

helix (*see* **Note 2**). (2) The positions of the NCS axes cannot be determined. For example, 432 point group symmetry requires the intersection of all symmetry axes at a point, in reality these axes may not intersect at all, but still maintain the orientation characteristic of 432—the cubic symmetry would be distorted.

2.3.3. Using the Information From the Self-Rotation Function to Solve a Structure

Besides building a working model of homo-oligomeric assembly, the self-rotation function lends information that can be used to help solve a structure by the following paths:

1. Molecular replacement: the locked rotation function (GLRF, AMoRe) improves chances of finding a rotation function solution by including NCS symmetry information. Similarly, the locked translation function of GLRF can take advantage of point group symmetry, but not improper symmetry. MOLREP performs a special translation function that works with both proper and improper symmetry ***(19)***.
2. Restraints for heavy atom location (*see* RSPS from the CCP4 suite).
3. Assistance in finding NCS in electron density maps. The usefulness of this information depends on the exactness of the NCS operator, which can be judged by the height and sharpness of the self-rotation function peaks.

3. Significance of Merohedral Twinning

Merohedral twinning is a common crystal growth defect in which a crystal is composed of two or more crystal domains related by a twofold "twin" operator such that their lattices coincide in three-dimensional space (*see* **Note 4**). Because the lattices coincide, the diffraction patterns from the separate domains superimpose as if from a normal, single lattice. But, because the domains have nonequivalent orientations, the observed intensities no longer correspond to a single crystal structure but instead to the superposition of twin-related structures. Merohedral twinning is especially insidious because it is difficult to diagnose unless specifically tested. If left undetected it can lead to much frustration and lengthy delays in structure determination (e.g., uninterpretable Patterson maps, ghost peaks in Fourier maps, high R_{free} after prolonged refinement). Although procedures have been established to overcome twinning with relative ease, early detection is the only key to avoid frustration when twinning is encountered. Merohedral twinning can be detected quickly and easily by application of twin tests to the native dataset. Reviews on merohedral twinning can be found in **refs.** ***20*** and ***21***.

3.1. Conditions for Merohedral Twinning

Not all space groups are prone to twinning, and so not all crystals require the application of twinning tests. The possibility of twinning requires certain relationships among cell parameters. In general, merohedral twinning can occur

whenever the rotational symmetry of the lattice exceeds that of the underlying crystal point group *(20)*. Take for example a P4 crystal; the tetragonal lattice contains, in addition to the crystallographic fourfold, a perpendicular twofold symmetry that exceeds the underlying P4 symmetry. This additional symmetry in the lattice allows the possibility for two nonequivalent orientations of the lattice (related by this twofold) to coincide in three dimensions (**Fig. 5**). All space groups falling under point group symmetry 3, 32, 4, 6, or 23 meet this criterion and so, therefore, are suspect of twinning.

In addition to these space groups, one must also suspect space groups with point group symmetry 622 and 432. Although these space groups cannot be twinned in theory (because their lattice symmetry does not exceed the underlying point group symmetry), twinning in 32 and 23 point groups can give rise to apparent 622 and 432 symmetry when the twin domains are of equal proportion (i.e., twinning fraction = 50% or perfectly twinned). Equal proportions of twin domains combined with the twofold symmetry of the twinning operator introduce an additional twofold symmetry in the diffraction pattern. This additional twofold masks the ”true symmetry” of the crystal, giving rise to a higher “apparent symmetry.” Less severe twinning (twin fractions less than 50%) is referred to as “partial twinning,” and does not mask the true symmetry of the crystal.

In summary, whenever you encounter a crystal that appears to belong to any of the space groups listed in the first column of **Table 2**, it is suspect for merohedral twinning and the twinning test should be applied automatically. One further note of caution: crystals in lower symmetry space groups can also be twinned in unfortunate circumstances (pseudomerohedral twinning). For example, crystals that appear to be orthorhombic can be twinned if two of the cell dimensions are equal. Similarly, crystals that appear to be monoclinic can be twinned if two of the cell dimensions are equal *(22,23)*, if β is close to 90° *(24–26)*, or if a = 2*c*cos(180-β) (or equivalently c=2*a*cos(180-β)) *(27)*.

3.2. Detection of Merohedral Twinning by Examination of Intensity Distribution Statistics

Merohedral twinning can be diagnosed from a myriad of different symptoms. Some have been previously discussed, e.g., inability to interpret difference Patterson maps, observance of ghost peaks in Fourier maps, maintenance of a high R_{free} after prolonged refinement, and molecular replacement solutions that overlap. But, these symptoms develop too late in the structure determination process and might be symptomatic of problems other than twinning. Earlier observations can aid identification of many (although not all) twinned crystals *(28)*. For example, the observance of concave faces or boundaries between halves of the crystal is a property of many twinned crystals *(20)*. An impossibly low V_M can also serve as an indication of perfect twinning. In this instance

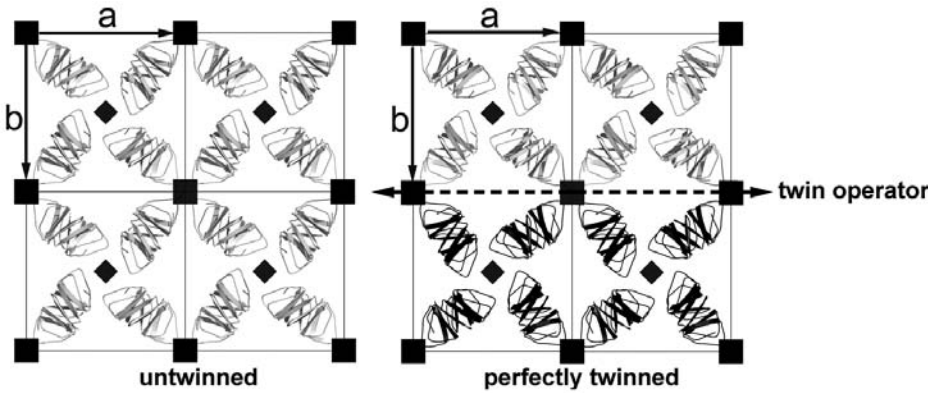


Fig. 5. A representation of merohedral twinning. Two crystals are depicted in space group P4, viewed down the principal crystallographic fourfold axis. The crystal on the left is not twinned. The crystal on the right is merohedrally twinned with a twin fraction of 50%. Half the cells belong to one crystal domain, the other half belong to a separate crystal domain. The domain on the top is related to the domain on the bottom by a twin operator. This operator is not part of the normal symmetry operators of space group P4.

the twin operator has misled you to assume a higher symmetry space group than is actually present (apparent vs true symmetry). However, if there are two or more molecules in the asymmetric unit, the V_M will appear normal even in the case of perfect twinning.

The most reliable methods to detect twinning involve the recognition of the aberrant intensity distributions that are characteristic of twinned crystals. A dataset from a twinned crystal contains fewer extremely weak and extremely strong reflection intensities than are predicted by Wilson statistics. This situation arises from the fact that the intensity of each reflection is the linear combination of contributions from two twin domains ***(20)***. Because there is only a poor likelihood that both contributions will be very large or that both contributions will be very small, very few extreme values result ***(28)***. Several tests have been devised to recognize this feature of twinned crystals ***(29)***, but one particular test has been singled out as being the most robust: a plot of the second moment of intensity vs resolution ***(20)***. This test can be implemented with the truncate program of the CCP4 suite, the CNS suite, or through the Merohedral Crystal Twinning Server. All three means of implementation are described next.

3.2.1 Use of CCP4's Truncate Program

1. Import your data into the CCP4's mtz format. The mtz file should contain at least H, K, L, intensity, and the standard deviation of the intensity measurement. The

Table 2
Twinning Relationships Sorted by Apparent Space Group[a]

Apparent space group	Twin operator	Partial or perfect	True space group
(75) P4	k,h,-l	partial	P4
(76) $P4_1$	k,h,-l	partial	$P4_1$
(77) $P4_2$	k,h,-l	partial	$P4_2$
(78) $P4_3$	k,h,-l	partial	$P4_3$
(79) I4	k,h,-l	partial	I4
(80) $I4_1$	k,h,-l	partial	$I4_1$
(89) P422	k,h,-l	perfect	P4
(91) $P4_122$	k,h,-l	perfect	$P4_1$
(94) $P4_222$	k,h,-l	perfect	$P4_2$
(95) $P4_322$	k,h,-l	perfect	$P4_3$
(97) I422	k,h,-l	perfect	I4
(98) $I4_122$	k,h,-l	perfect	$I4_1$
(143) P3	-h,-k,l	partial	P3
	k,h,-l	partial	P3
	-k,-h,-l	partial	P3
(144) $P3_1$	-h,-k,l	partial	$P3_1$
	k,h,-l	partial	$P3_1$
	-k,-h,-l	partial	$P3_1$
(145) $P3_2$	-h,-k,l	partial	$P3_2$
	k,h,-l	partial	$P3_2$
	-k,-h,-l	partial	$P3_2$
(146) R3	k,h,-l	partial	R3
(149) P312	k,h,-l	perfect	P3
	-h,-k,l	partial	P312
(150) P321	-k,-h,-l	perfect	P3
	-h,-k,l	partial	P321
(151) $P3_112$	k,h,-l	perfect	$P3_1$
	k,h,-l	perfect	$P3_2$
	-h,-k,l	partial	$P3_112$
(152) $P3_121$	-k,-h,-l	perfect	$P3_1$
	-k,-h,-l	perfect	$P3_2$
	-h,-k,l	partial	$P3_121$
(153) $P3_212$	k,h,-l	perfect	$P3_1$
	k,h,-l	perfect	$P3_2$
	-h,-k,l	partial	$P3_212$
(154) $P3_221$	-k,-h,-l	perfect	$P3_1$
	-k,-h,-l	perfect	$P3_2$
	-h,-k,l	partial	$P3_221$
(155) R32	k,h,-l	perfect	R3
(168) P6	-h,-k,l	perfect	P3
	k,h,-l	partial	P6

Table continues

Table 2 (continued)

Apparent space group	Twin operator	Partial or perfect	True space group
(169) $P6_1$	k,h,-l	partial	$P6_1$
(170) $P6_5$	k,h,-l	partial	$P6_5$
(171) $P6_2$	k,h,-l	partial	$P6_2$
	-h,-k,l	perfect	$P3_1$
	-h,-k,l	perfect	$P3_2$
(172) $P6_4$	k,h,-l	partial	$P6_4$
	-h,-k,l	perfect	$P3_1$
	-h,-k,l	perfect	$P3_2$
(173) $P6_3$	k,h,-l	partial	$P6_3$
(177) P622	-h,-k,l	perfect	P312
	-h,-k,l	perfect	P321
	k,h,-l	perfect	P6
(178) $P6_122$	k,h,-l	perfect	$P6_1$
(179) $P6_522$	k,h,-l	perfect	$P6_5$
(180) $P6_222$	-h,-k,l	perfect	$P3_n12$
	-h,-k,l	perfect	$P3_n21$
	k,h,-l	perfect	$P6_2$
(181) $P6_422$	-h,-k,l	perfect	$P3_n12$
	-h,-k,l	perfect	$P3_n21$
	k,h,-l	perfect	$P6_4$
(182) $P6_322$	k,h,-l	perfect	$P6_3$
(195) P23	k,h,-l	partial	P23
(196) F23	k,h,-l	partial	F23
(197) I23	k,h,-l	partial	I23
(198) $P2_13$	k,h,-l	partial	$P2_13$
(199) $I2_13$	k,h,-l	partial	$I2_13$
(207) P432	k,h,-l	perfect	P23
	k,h,-l	perfect	$P2_13$
(209) F432	k,h,-l	perfect	F23
	k,h,-l	perfect	$F2_13$
(211) I432	k,h,-l	perfect	I23
	k,h,-l	perfect	$I2_13$

[a]Based on information from **ref. *21***.

corresponding column labels should be H, K, L, IMEAN, and SIGIMEAN. If your data is in scalepack format, this conversion can be performed using the scalepack2mtz program from CCP4. The same conversion can be performed with the CCP4i graphical user interface (GUI) by selecting "import scaled data" from the "Data Reduction" menu (**Fig. 6**). Use of the GUI has the added advantage of performing **step 2** in the same script. If your data is in an ascii format other than

```
scalepack2mtz HKLIN intensities.sca HKLOUT intensities.mtz
SYMM myspacegroup
ANOMALOUS no
END

f2mtz HKLIN intensities.hkl HKLOUT intensities.mtz
TITLE convert CNS format to MTZ format
SYMMETRY myspacegroup
CELL 72.95 72.95 42.86 90.0 90.0 90.0
FORMAT '(6X,3F5.0,6X,F10.3,17X,F10.3,)'
SKIPLINE 5
LABOUT H K L IMEAN SIGIMEAN
CTYPEOUT H H H F Q
PNAME myproject
DNAME mydataset
end

truncate HKLIN intensities.mtz HKLOUT structurefactors.mtz >truncate.log
TITLE convert Is to Fs and test for twinning
LABOUT F=FP sigF=sigFP
NRESIDUES 335
END
```

Fig. 6. Input script file for calculating twinning statistics via CCP4's truncate program.

scalepack, the f2mtz program from CCP4 may be used to converted to mtz format. *See* the f2mtz manual for further details.

2. Run the truncate script. If you used the CCP4i GUI in the previous step, then this step is unnecessary.
3. Plot the contents of the truncate log file. This task may be performed by typing "loggraph truncate.log." From the "Tables in File" menu, select "Acentric Moments of (I**k)/(I)**k." Then from the "Graphs in Selected Table" menu select "2nd Moment of I". The second moment of intensity (vertical axis) will be plotted as a function of resolution (horizontal axis).

3.2.2. Use of CNS Twinning Detection Script

1. Import your data into CNS format. If your data is in scalepack or d*trek format, use to_cns from the CNS suite. If your data is in CCP4's mtz format, use mtz_to_cns.
2. Edit the detect_twinning.inp script in the CNS GUI. Enter cell dimensions, resolution range, and reflection file name. Execute the script (cns_solve <detect_twinning.inp).
3. Examine the twinning statistics table in the log file (detect_twinning.list). The sixth column reports the second moment of intensity, $\langle |I|^2 \rangle / (\langle |I| \rangle)^2$, as a function of resolution.

3.2.3. Use of the Merohedral Crystal Twinning Server

1. Import your data into CNS (X-PLOR) format. The twinning server requires H,K,L, F_{obs}, and sigma (F_{obs}) in CNS format. If your data is in scalepack format I recommend that you use xprepfin from the XtalView ***(30)*** suite to convert to CNS format. Although the to_cns utility provided in the CNS package is a more obvious choice to convert scalepack to CNS format, it splits HKL and FOBS on separate lines causing the twin server to report an error. If xprepfin is inconvenient you can modify the following awk script to convert any ascii file to CNS format: cat intensities.sca |awk '{printf("INDE= %4i%4i%4i FOBS= %8.2f SIGMA= %8.2f \n",$1,$2,$3,sqrt($4),sqrt($5)/2);}'> strucfactors.fobs.
2. Perform the perfect twinning test. In your web browser, go to the twinning server located at http://www.doe-mbi.ucla.edu/Services/Twinning. Select "perfect merohedral test." Enter your full-resolution range, your unit cell parameters, and select the apparent space group symmetry. Select "postscript" output. Finally, upload your CNS formatted native structure factors.
3. Examine the graph.

3.3. Interpretation of the Twinning Test

Perfect twinning is indicated when values of the second moment of intensity is 1.5 across the resolution range. Values between 1.5 and 2.0 indicate partial twinning. Values close to 2.0 indicate no twinning is present. The test is generally reliable except in special cases where there is translational NCS ***(31,32)*** or if there is severe anisotropy. Both these phenomena can foil twinning detection by altering the intensity distribution in a manner opposite to that caused by twinning (sometimes pushing the second moment of intensity to values greater than 2.0). To avoid this potential pitfall it is best to check the native Patterson map for translational NCS before running the twinning test. When anisotropy or translational NCS is indicated, one may turn to a more recently developed twinning statistic ***(33)***.

If your crystal tests positive for twinning, you will be faced with a decision whether to continue using this data in structure determination or to search for another untwinned crystal (preferred). Your decision may be influenced by the severity of twinning (twin fraction) and the means of structure determination to be employed. Molecular replacement has proven to be highly successful with crystals of any twin fraction and is relatively straightforward. There are some procedural modifications, but conventional molecular replacement software may be used ***(20)***. On the other hand, the use of isomorphous replacement or anomalous scattering can be significantly complicated by the effects of twinning ***(20)***. Higher twin fractions (greater than 40%) complicate interpretation of difference Patterson maps and electron density maps. Nevertheless, there are a growing number of examples of its successful use in the literature ***(34)***. For the less adventurous, the most rapid path to structure determination may be to search for an untwinned crystal. An

untwinned crystal may be obtained by adjusting the crystallization conditions, (e.g., adjusting pH) or by screening for another crystal form entirely.

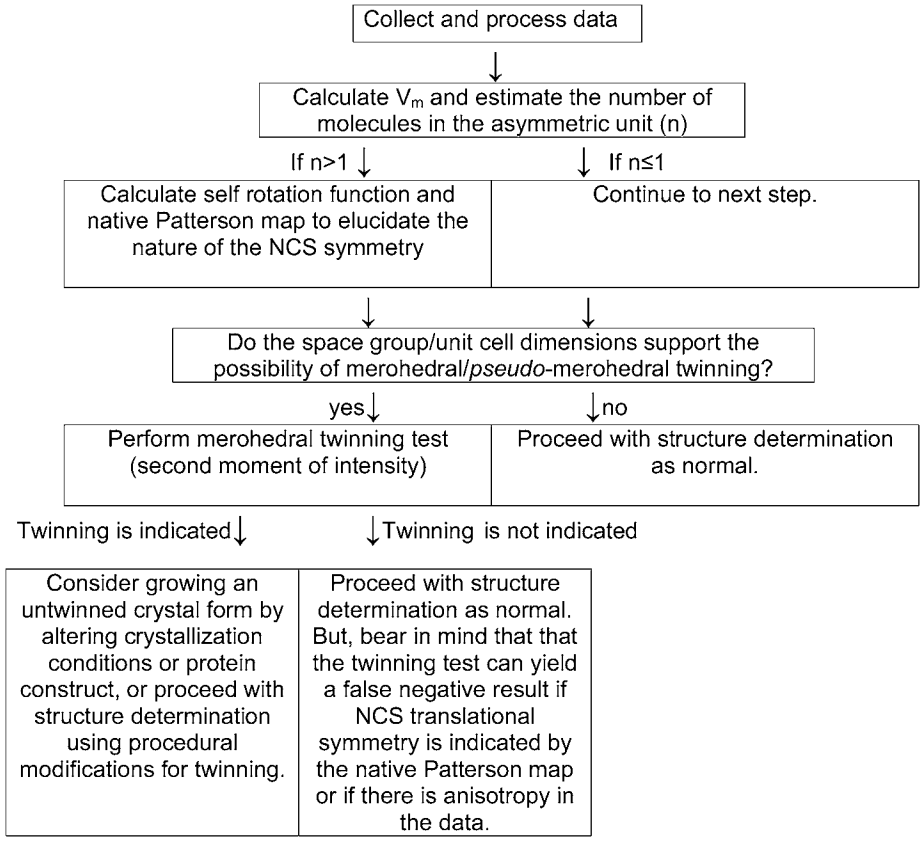
If you decide to continue structure determination with a twinned crystal, the next step is to determine the degree of twinning (twin fraction), which twinning operator caused the twinning (if more than one is possible), and what is the symmetry of the true space group. **Table 2** lists the combinations of these parameters that are possible for your apparent space group symmetry. To discern among the possibilities, you may employ the partial twinning test (*see* **Note 5**). Results of the partial twinning test may be obtained from the CCP4 program "detwin," or the CNS script "detect_twinning.inp." Results may also be obtained from the Merohedral Crystal Twinning Server in the same way as the perfect twinning test previously described, with the exception that you select "partial merohedral twinning test" at the top of the page. Having determined these parameters you can then attempt to recover the true crystallographic intensities from the twinned dataset (i.e., detwin) providing the twin fraction is less than about 40%, or proceed with structure determination with the twinned data and model the effects of twinning in structural refinement CNS, REFMAC (from CCP4), and SHELX *(28)* to handle twinned data.

3.4. Conclusion

The flow chart summarizes the recommended steps in crystal characterization used in preparation for structure determination. If the planned route of structure determination involves the use of isomorphous replacement, one should consider collecting a second native dataset to gage the reproducibility of the crystal's unit cell dimensions. If the two native crystals grown under the same conditions are not strongly isomorphous, then one can conclude that it will be difficult to produce an isomorphous derivative.

4. Notes

1. To find *Z*, you may simply sum up the number of symmetry operators for the space group, remembering to include all centering operations if present.
2. Both rotational and screw axes produce peaks in the self-rotation function and the two types of axes cannot be distinguished from each other based on the self-rotation function alone. This degeneracy comes from the fact that the self-rotation function examines only orientational relationships among Patterson self vectors (intramolecular vectors). If you recall from the nature of the Patterson map, the self vectors have been translated so that their tails lie at the origin (*see* **Subheading 2.2.**), obliterating any trace of the translational component of screw axes.
3. At this point it helps to remember that the self-rotation function plot is really a projection of a sphere. Think of the twofold rotation function peaks as points that lie



on the surface of the sphere. Imagine positioning a plane that contains these points as well as the center of the sphere. A great circle is formed by the intersection of this plane with the surface of the sphere. Draw the projection of this great circle on the self-rotation function plot.

4. The twinning can be either of two types, epitaxial or merohedral. Epitaxial (or non-merohedral) twinning is obvious from the appearance of multiple distinct reciprocal lattices in a diffraction image. When nonmerohedral twinning is observed, another untwinned crystal can be selected for data collection. In some cases it is possible to have higher forms of twinning with more than two orientations of crystal domains, but these are rarely reported. A crystal domain is simply a subset of the crystal that may be considered a single crystal if it were isolated from its neighboring domains.
5. The partial twinning statistic can yield a false-positive result when a twofold NCS axis is aligned with a potential twin operator. Before concluding that partial twinning is at play, check that the perfect twinning test indicates twinning or rule out the possibility of NCS by packing density limitations.

Acknowledgments

The author thanks Todd O. Yeates and Duilio Cascio for valuable comments and suggestions.

References

1. Vonrhein, C. and Schulz, G. E. (1999) Locating proper non-crystallographic symmetry in low-resolution electron-density maps with the program GETAX. *Acta Cryst*. **D55,** 225–229.
2. Matthews, B. W. (1968) Solvent content of protein crystals. *J. Mol. Biol*. **33,** 491–497.
3. Dodson, E. J., Winn, M., and Ralph, A. (1997) Collaborative computational project, number 4: providing programs for protein crystallography. *Methods Enzymol*. **277,** 620–633.
4. Kantardjieff, K. A. and Rupp, B. (2003) Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Science* **12,** 1865–1871.
5. Matthews, B. W. (1985) Determination of protein molecular weight, hydration, and packing from crystal density. *Methods Enzymol*. **114,** 176–187.
6. Westbrook, E. M. (1985) Crystal density measurements using aqueous ficoll solutions. *Methods Enzymol*. **114,** 187–196.
7. Leung, A. K. W., Park, M. M. V., and Borhani, D. W. (1999) An improved method for protein crystal density measurements. *J. Appl. Cryst*. **32,** 1006–1009.
8. Patterson, A. L. (1934) A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.* **46,** 372–376.
9. Navaza, J. (2001) Implementation of molecular replacement in AMoRe. *Acta Cryst*. **D57,** 1367–1372.
10. Vagin, A. and Teplyakov, A. (1997) MOLREP: an automated program for molecular replacement. *J. Appl. Cryst*. **30,** 1022–1025.
11. Brunger, A. T., Adams, P. D., Clore, G. M., et al. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Cryst*. **D54,** 905–921.
12. Tong, L. and Rossmann, M. G. (1997) Rotation function calculations with GLRF program. *Methods Enzymol*. **276,** 594–611.
13. Rossmann, M. G. and Blow, D. M. (1962) The detection of subunits within the crystallographic asymmetric unit. *Acta Cryst*. **15,** 24–31.
14. Rossmann, M. G. (2001) Molecular replacement–historical background. *Acta Cryst*. **D57,** 1360–1366.
15. Buehner, M., Ford, G. C., Moras, D., Olsen, K. W., and Rossmann, M. G. (1973). D-glyceraldehyde-3-phosphate dehydrogenase: three-dimensional structure and evolutionary significance. *Proc. Natl Acad. Sci. USA* **70,** 3052–3054.
16. Goodsell, D. S. and Olson, A. J. (2000) Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct*. **29,** 105–153.
17. Blundell, T. L. and Srinivasan, N. (1996) Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Natl. Acad. Sci. USA* **93,** 14,243–14,248.

18. Padilla, J. E. and Yu, T. (2002) Self-assembling symmetric protein materials. In: *Biopolymers Vol. 7: Polyamides and Complex Proteinaceous Materials I,* (Steinbuchel, A. and Fahnestock, S. R., eds.), Wiley-VCH, Weinheim, Germany, pp. 261–284.
19. Vagin, A. and Teplyakov, A. (2000) An approach to multi-copy search in molecular replacement. *Acta Cryst*. **D56,** 1622–1624.
20. Yeates, T. O. (1997) Detecting and overcoming crystal twinning. *Methods Enzymol*. **276,** 344–358.
21. Chandra, N., Acharya, K. R., and Moody, P. C. E. (1999) Analysis and characterization of data from twinned crystals. *Acta Cryst*. **D55,** 1750–1758.
22. Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B., and Steitz, T. A. (1999) Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit. *Nature* **400,** 841–847.
23. Yang, F., Dauter, Z., and Wlodawer, A. (2000) Effects of crystal twinning on the ability to solve a macromolecular structure using multiwavelength anomalous diffraction. *Acta Cryst*. **D56,** 959–964.
24. Hugot, M., Bensel, N., Vogel, M., et al. (2002) A structural basis for the activity of retro-Diels-Alder catalytic antibodies: evidence for a catalytic aromatic residue. *Proc. Natl. Acad. Sci. USA* **99,** 9674–9678.
25. Larsen, N. A., Heine, A., de Prada, P., et al. (2002) Structure determination of a cocaine hydrolytic antibody from a pseudomerohedrally twinned crystal. *Acta Cryst*. **D58,** 2055–2059.
26. Dunitz, J. D. (1964) The interpretation of pseudo-orthorhombic diffraction patterns. *Acta Cryst*. **17,** 1299–1304.
27. Declercq, J. -P. and Evrard, C. (2001) A twinned monoclinic form of human peroxiredoxin 5 with eight molecules in the asymmetric unit. *Acta Cryst*. **D57,** 1829–1835.
28. Herbst-Irmer, R. and Sheldrick, G. M. (1998) Refinement of twinned structures with SHELXL97. *Acta Cryst*. **B54,** 443–449.
29. Kahlenberg, V. (1999) Application and comparison of different tests on twinning by merohedry. *Acta Cryst*. **B55,** 745–751.
30. McRee, D. E. (1993) *Practical Protein Crystallography.* Academic, San Diego, CA.
31. Lee, S., Sawaya, M. R., and Eisenberg, D. (2003) Structure of superoxide dismutase from *Pyrobaculum aerophilum* presents a challenging case in molecular replacement with multiple molecules, pseudo-symmetry and twinning. *Acta Cryst*. **D59,** 2191–2199.
32. Barends, T. R. and Dijkstra, B. W. (2003) *Acetobacter turbidans* α-amino acid ester hydrolase: merohedral twinning in $P2_1$ obscured by pseudo-translational NCS. *Acta Cryst.* **D59,** 2237–2241.
33. Padilla, J. E. and Yeates, T. O. (2003) A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning. *Acta Cryst*. **D59,** 1124–1130.
34. Rudolph, M. G., Kelker, M. S., Schneider, T. R., et al. (2003) Use of multiple anomalous dispersion to phase highly merohedrally twinned crystals of interleukin-1b. *Acta Cryst.* **D59,** 290–298.

35. Xie, P. T. and Hurley, T. D. (1999) Methionine-141 directly influences the binding of 4-methylpyrazole in human σσ alcohol dehydrogenase. *Protein Sci*. **8,** 2639–2644.
36. Campobasso, N., Mathews, I. I., Begley, T. P., and Ealick, S. E. (2000) Crystal structure of 4-methyl-5-β-hydroxyethylthiazole kinase from *Bacillus subtilis* at 1.5 Å resolution. *Biochem*. **39,** 7868–7877.
37. Tulip, W. R., Varghese, J. N., Baker, A. T., et al. (1991) Refined atomic structures of N9 subtype influenza virus neuraminidase and escape mutants. *J. Mol. Biol.* **221,** 487–497.
38. Stehle, T. and Harrison, S. C. (1997) High-resolution structure of a polyomavirus VP1-oligosaccharide complex: implications for assembly and receptor binding. *EMBO J*. **16,** 5139–5148.
39. Lenzen, C. U., Steinmann, D., Whiteheart, S. W., and Weis, W. I. (1998) Crystal structure of the hexamerization domain of N-ethylmaleimide-sensitive fusion protein. *Cell* **94,** 525–536.
40. Rigden, D. J., Alexeev, D., Phillips, S. E., and Fothergill-Gilmore, L. A. (1998) The 2.3 Å X-ray crystal structure of *S. cerevisiae* phosphoglycerate mutase. *J. Mol. Biol.* **276,** 449–459.
41. Zhou, T., Kurnasov, O., Tomchick, D. R., et al. (2002) Structure of human nicotinamide/nicotinic acid mononucleotide adenylyltransferase. Basis for the dual substrate specificity and activation of the oncolytic agent tiazofurin. *J. Biol. Chem*. **277,** 13,148–13,154.
42. Fraaije, M. W., van den Heuvel, R. H., van Berkel, W. J., and Mattevi, A. (1999) Covalent flavinylation is essential for efficient redox catalysis in vanillyl-alcohol oxidase. *J. Biol. Chem*. **274,** 35,514–35,520.
43. Braden, B. C., Velikovsky, C. A., Cauerhff, A. A., Polikarpov, I., and Goldbaum, F. A. (2000) Divergence in macromolecular assembly: X-ray crystallographic structure analysis of lumazine synthase from *Brucella abortus*. *J. Mol. Biol.* **297,** 1031–1036.
44. Gill, H. S., Pfluegl, G. M., and Eisenberg, D. (2002) Multicopy crystallographic refinement of a relaxed glutamine synthetase from *Mycobacterium tuberculosis* highlights flexible loops in the enzymatic mechanism and its regulation. *Biochemistry* **41,** 9863–9872.
45. Grant, R. A., Filman, D. J., Finkel, S. E., Kolter, R., and Hogle, J. M. (1998) The crystal structure of Dps, a ferritin homolog that binds and protects DNA. *Nat. Struct. Biol.* **5,** 294–303.
46. Précigoux, G., Yariv, J., Gallois, B., Dautant, A., Courseille, C., and Langlois d'Estaintot, C. (1994) A crystallographic study of haem binding to ferritin. *Acta Cryst.* **D50,** 739–743.

7

Molecular Replacement

Eric A. Toth

Summary

As more protein structures are solved, the likelihood that current structural investigations will involve proteins for which there exists no homologous structure continually decreases. The extraction of phase information from diffraction experiments is one of several great barriers that crystallographers must overcome on the path to structure solution. One means to overcome this obstacle, the technique of molecular replacement, uses the structural similarity between proteins with similar sequences to give a good first estimate of the phases for the diffraction data of the protein of interest. The programs that execute this technique currently come in many flavors, from traditional Patterson-based methods, to stochastic searches in greater than three dimensions, to maximum likelihood-enhanced molecular replacement, each possessing unique advantages that can shake loose a recalcitrant solution. As crystallographers aim to solve larger macromolecular complexes that more faithfully depict the actors in cellular events, having existing phase information for parts of those biological machines will reinforce the technological advancements in data collection and structure solution that have already produced mammoth structures like the ribosome, yielding an ever-clearer picture of the inner workings of biology.

Key Words: Crystal; molecular replacement; phasing; X-ray; Phaser; AMoRe; Molrep; CNS; EPMR; Qs; SOMoRe; COMO; CCP4; rotation search; translation search.

1. Introduction

Taking advantage of the structural similarity of proteins with high-sequence homology is a conceptually straightforward way to overcome the phase problem in protein crystallography *(1)*. Unfortunately, this method, known as molecular replacement, is not always as easy to put into practice as it is to conceptualize. That being said, the continual development of the method over the years, along with a massive increase in affordable computational power, has allowed crystallographers to solve difficult molecular replacement problems that had once been intractable. Crystallographers now have at their disposal traditional Patterson-based approaches, six-dimensional (6D) searches, phase-based searches, and

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

maximum-likelihood molecular replacement, allowing them to utilize the strong points of each method. This cartel of methods, combined with structural genomics efforts cranking out many structures of "orf12345" and "conserved hypothetical protein X" affords the modern crystallographer the opportunity to embark on a rigorous and desperate attempt to avoid experimental phasing.

1.1. The Mechanics of the Search

The basic concepts behind molecular replacement are that proteins with similar sequences also have a similar three-dimensional structure and that the phases from a similar structure would be a good first estimate of the phases for the structure that is to be determined. Therefore, there must be an orientation of the known structure that, when placed in the unit cell of the unknown structure, will give a maximum overlap between the two ***(2)***. This can be described by a simple rotation and translation of the model from its current (arbitrary) position (usually chosen to bring its center of mass to the origin):

$$\vec{x}' = [C]\vec{x} + \vec{t} \tag{1}$$

where $[C]$ is a matrix describing the rotational parameters that bring the search molecule into maximum coincidence with the target molecule, and $\vec{t}$ is the translation vector that maximizes their overlap. The simplest way to think about achieving the maximum overlap between the target and known structures is to think about a brute force search. This involves placing the search model in the unit cell of the target structure, searching all possible rotations and translations, and calculating the quality of the overlap for each rotation and translation. Because we do not have any phases, and, therefore, any map for the target structure, the quality of the overlap would be determined by how well the observed structure factors (F_{obs}) agree with those calculated from the rotated and translated molecule (F_{calc}). In our simplistic brute force approach, this would require one fast Fourier transform (FFT) for each point evaluated. If we assume a medium-sized protein with three 100 Å cell edges (searched in 1-Å increments) and searching all of the rotation space without regard to symmetry in 1° increments, this would require $(360)^3 \times (100)^3$ or approx 4.7×10^{13} FFTs. That calculation is intractable today, and it was more so when the method was developed. This left a simple idea without a practical implementation, so to speed up the search, developers took the "divide and conquer" approach. If one takes our inelegant search from above and separates the rotation and translation searches, there will be 47 million points to search in rotation space, and 1 million in translation space. This is doable, and with much more sophisticated programming methods and mathematical representations, can be a very fast calculation. The key to this "divide and conquer" approach is finding a means of evaluating the overlap between the two molecules that is *independent* of translation, and then another

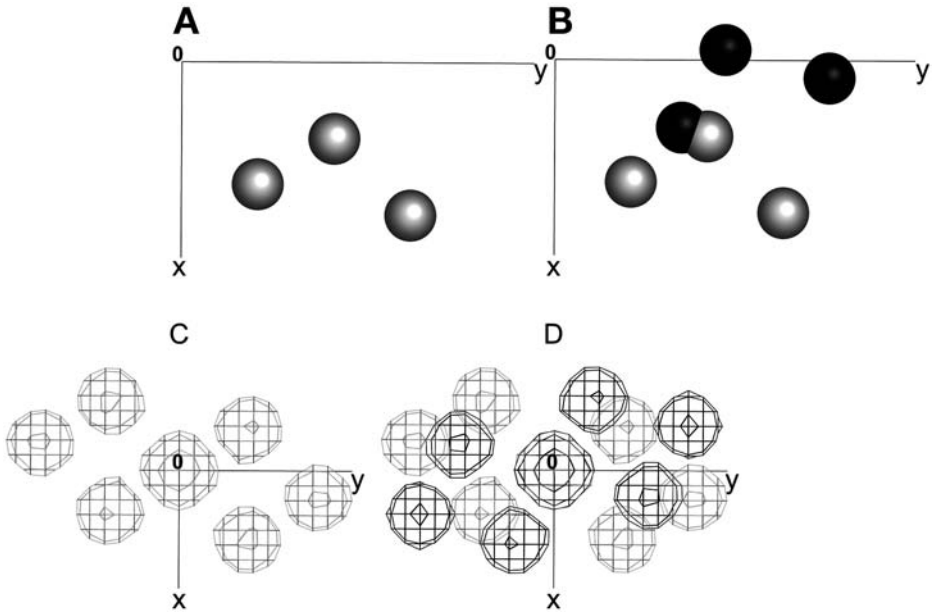


Fig. 1. The rotation-only dependence of the Patterson function. **(A)** A molecule comprised of three equal atoms (gray). 0 represents the origin. **(B)** The same gray molecule from **(A)**, along with the same molecule rotated by 30° about z (the axis comes out of the page). **(C)** The region of the Patterson function for the molecule in **A** close to the origin. **(D)** The Patterson function of the black molecule in **B** close to the origin superposed on the Patterson function of the same region for the gray molecule. The relationship between the two clearly involves a simple rotation.

measure that is *dependent* on the translation of the search model. Fortunately, the Patterson function *(3)* satisfies both these criteria by using different subsets of Patterson vectors for each search. A Patterson map contains all of the vectors between all atoms in the unit cell. Some of these vectors will be *intramolecular*, that is between atoms within the same molecule, and some will be *intermolecular*, or between different molecules in the unit cell. As you can probably already see, the intramolecular vectors have no translational component to them because they describe individual molecules and not how they are distributed in the unit cell, and that the intermolecular vectors contain all of the information about translation. Therefore, if one can separate these two sets of vectors, one can search for the correct *orientation* by maximizing the overlap of the observed and calculated intramolecular vectors *(2)* and the correct *translation* by maximizing the overlap between the observed and calculated intermolecular vectors. These calculations can be done in either real or reciprocal space, but are almost exclusively done in reciprocal space in the interest of speed.

The rotation-only dependence of the Patterson function of a molecule is illustrated in **Fig. 1**. If one considers a three-atom molecule that is shown in gray in the panel (**Fig. 1A**), the cluster of self-Patterson peaks shown in **Fig. 1C** is pretty simple. If one rotates this three-atom molecule by some arbitrary amount (in this case 30°, black in panel [**Fig. 1B**]), it is easy to see that the cluster of self vectors rotates by the same amount (**Fig. 1D**), and the relationship between the two Pattersons is pretty clear. Take note that the molecule was rotated such that its center of mass no longer coincides with its gray progenitor, yet their Patterson self vectors have a common center of mass, i.e., the origin. This demonstrates that there is no translational component to this set of Patterson vectors. Thus, measuring their overlap measures the correctness of the orientation of the search model. From this real space representation of the Patterson function, it might or might not be obvious that placing a sphere at the origin with a radius just large enough to encompass the self Patterson vectors and then defining everything outside that sphere to be zero effectively isolates these vectors for the rotation search. Performing this same operation in reciprocal space involves taking the transform of a sphere with the appropriate radius centered at the origin and multiplying the coefficients that describe the reciprocal space Patterson overlap by the coefficients of the transformed spherical interference function. This effectively eliminates any terms in the summation that would describe intermolecular vectors for the two Patterson functions, assuming that your protein is spherical. In practical terms, enough cross vectors are eliminated (or enough self-vectors remain) to adequately describe the rotational relationship between the search and target molecules. This is fairly obvious from the original R function described by Rossman and Blow ***(2)***:

$$R = (U/V^3)\sum_{\vec{h}} \sum_{\vec{h}'} \left|F(\vec{h})\right|^2 \left|F(\vec{h}'[C]^{-1})\right|^2 G(\vec{h}+\vec{h}') \qquad (2)$$

The guts of this equation is the product of the Patterson coefficients for the target and rotated search models, with the weighting function G (*see* **Note 1**).

Once the molecule is properly oriented, its precise location in the unit cell must be determined. This can be accomplished by maximizing the agreement between the set of intermolecular Patterson vectors (i.e., those that arise from vectors between molecules related by crystallographic symmetry) from the search and target structures. Because this calculation has to take into account the crystallographic symmetry, it can take on many space group specific forms, but the general form of the translation function is ***(4)***:

$$T(t) = \sum_{\vec{h}} \left|F_{obs}(\vec{h})\right|^2 \vec{F}_M(\vec{h})\vec{F}_M(\vec{h}[A])e^{-2\pi i\vec{h}\cdot\vec{t}} \qquad (3)$$

In this case $[A]$ is the crystallographic rotation matrix that relates two symmetry mates in the search structure. This function should be evaluated for each $[A]_i$

for that space group (containing *i* symmetry operators), thereby examining each independent pair of molecules in the unit cell (at a minimum, one must examine a combination of pairs that uniquely identifies the three components of the translation vector). Because the noise level of the translation function is higher in higher symmetry space groups, the redundancy provided by checking each independent pair increases the chances for success. It is also useful to exclude all self vectors from the calculation *(4)*, as they tend to degrade the signal-to-noise ratio (*see* **Notes 2** and **3**).

1.2. Evaluation of Possible Solutions

The three principal means by which solutions are evaluated are by the correlation coefficient (CC) and R-factor statistics, and by the chemical feasibility of their packing arrangement within the unit cell. The CC is generally regarded as a more useful statistic than the R-factor, but both can be implausibly good in the presence of packing violations, leading to potentially wrong solutions taken in favor of the correct one. Some programs calculate the packing explicitly, i.e., by checking the number of close contacts between C_α atoms of symmetry mates and disallowing solutions that have over a predetermined number of clashes *(5)*, whereas some programs use the translation function formulation by Harada and Lifchitz *(6)*, which downweighs solutions with bad packing by normalizing the original fast translation function. It is generally a good idea to make sure that the CC and R-factor have improved at each step of the molecular replacement process, and that the packing is chemically reasonable. If you do not trust the program to evaluate the packing for you, do it by eye. One sign that molecular replacement is *not* converging upon a solution is a lack of improvement in the R-factor or CC in the translation step. Because even a perfectly oriented molecule that is mistranslated will have only a modest correlation between its structure factors and the observed ones (because its phases do not much resemble the correct ones), the absence of a significant improvement in these statistics means that a solution is not at hand. The translation function is very sensitive to even small errors in the orientation of the search molecule, so if one fails to search enough rotation function peaks, a perfectly identifiable solution might be missed. **Figure 2** shows an example of a minor rotational difference that can preclude ready determination of the correct translation. Orientation errors of 3° in each α, β, and γ resulted in the correct translation scoring lower than incorrect solutions, whereas the absence of these errors yielded a readily identifiable translation search solution. Keep in mind that this is just one test case. The amount of orientational error that can be tolerated by the translation function is case specific. Rigid-body refinement *(7,8)* after the translation function can tidy up minor mistakes in orientation and translation. Once again, this is judged by a lower R-factor and a higher CC.

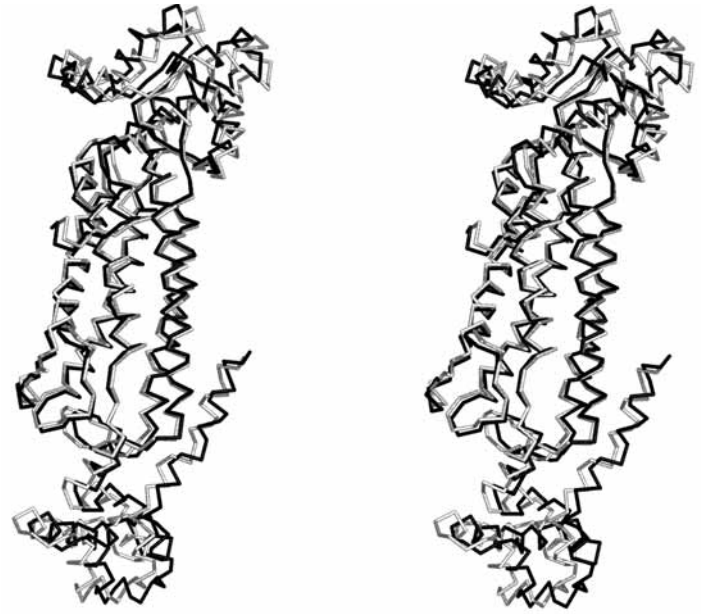


Fig. 2. A minor rotational difference can preclude ready determination of the correct translation. A stereo representation of two solutions to the rotation function, one of which leads to a correct translation function solution and one that does not. The difference in orientation is slight, involving only 3° differences in each α, β, and γ.

2. Traditional or Patterson-Based Molecular Replacement

2.1. Parameters That Affect the Rotation Search

Most programs require, or at least allow, user input regarding certain parameters that can be crucial for the success or failure of the rotation search. These include the resolution limits used, the radius of integration, and the size of the model unit cell. Other factors that affect the success rate of the rotation search are the space group symmetry and the presence or absence of noncrystallographic symmetry (NCS). This section will review these parameters and then how to use some of the programs that implement the Patterson-based molecular replacement approach in the pages that follow.

2.1.1. Resolution Limits

The basic idea behind choosing the resolution limits for a rotation search is to exclude information that is not likely to help finding a solution. One portion of data to be excluded is very low-resolution data, which contains a lot of information regarding the solvent, and the overall shape of the molecule from a "blobbic" perspective. These data do not describe anything unique

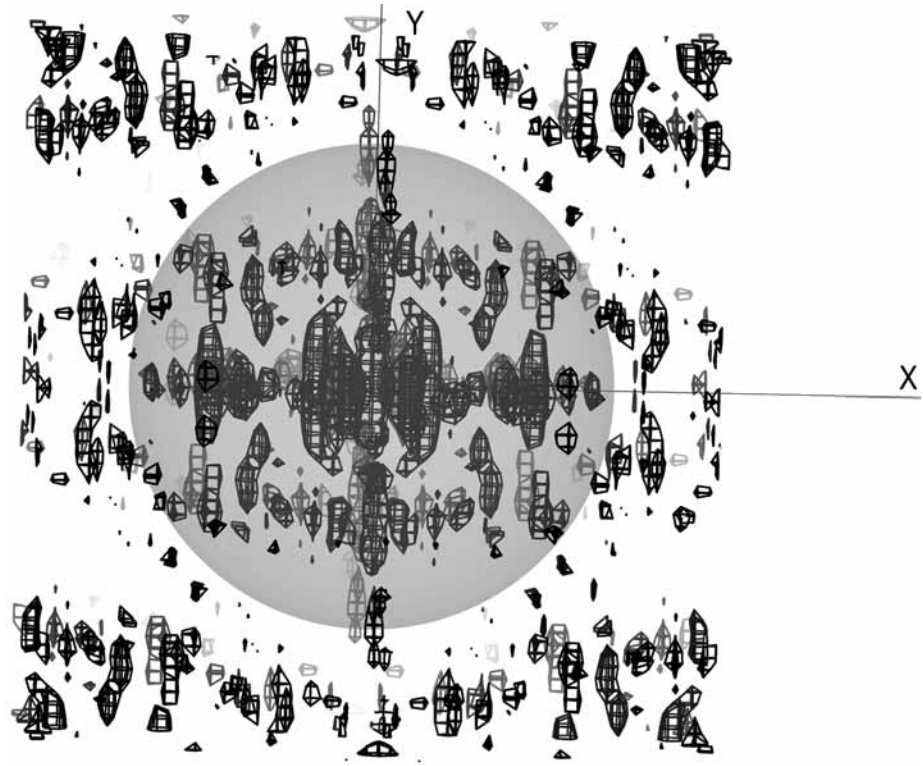


Fig. 3. Example of a spherical mask around the self-Patterson vectors of a molecule. The mask was calculated using data from 10 to 6 Å, a 1.3σ cutoff, and a 40-Å spherical radius.

enough to help determine the rotational parameters and contribute only noise to the search. High-resolution data, which contain information about features of the search structure that *are likely to be different* than the target structure, should also be excluded. Depending on the complexity and/or enormity of the problem, high-resolution limits might be set to prevent overtaxing computing resources (or the experimenter's limited patience; *see* **Note 4**).

2.1.2. Radius of Integration

By setting the radius of integration, the investigator seeks to include as many self vectors and exclude as many cross vectors as possible from the Patterson map of the target structure, with the emphasis being on the latter goal, as introduced noise is, in most cases, more deleterious than information lost via the spherical interference function. **Figure 3** shows an example of a spherical mask around the self-Patterson vectors of a molecule. Note how the mask covers most

of the density surrounding the origin. In general, using a sphere with a radius approx 75% of the expected radius of one's molecule should encompass enough self-Patterson vectors to discriminate a solution in a rotation function while at the same time eliminating nearly all of the intermolecular vectors. However, if one's molecule has a very elongated shape, the 75% criterion might not yield the best radius of integration. In these cases, and in many other difficult molecular replacement problems, the radius of integration might have to be optimized to give the best separation of the true solutions from the false ones.

2.1.3. Model Unit Cell

Although some may differ in how important the cell dimensions of the model are, the key is to make this cell large enough to avoid any intermolecular vectors from models in adjacent unit cells being incorporated into the structure factor calculation. Making the cell bigger simply increases the computational requirements and should not affect the results in any appreciable way, unless of course the initial choice of model unit cell was too small. The angles are always 90° and for simplicity, most people use a cubic cell, but it is not a requirement.

2.1.4. Crystal Symmetry and NCS

Simply put, the higher the symmetry of one's crystal, the more difficult molecular replacement will be. Because the calculation of the model Patterson function is performed in an artificial P1 cell, it can at most account for $1/N_{sym}$ of the molecules in the unit cell (where N_{sym} is the number of symmetry operations for that space group). Imagine how different the magnitude of the Patterson self vectors might be for a molecule with no symmetry (P1) and the same molecule in which 24 copies (cubic symmetry) are used to generate the observed Patterson vectors. The higher the symmetry, the larger the origin peak will be, and consequently the closer to the noise level the nonorigin self-vector peaks (i.e., all the ones that are important for this calculation) will be. This makes the overlap function much harder to interpret. NCS can also work to the detriment of the rotation function, but not if the NCS is precisely known. In this case, it becomes a powerful aide to structure solution by restricting the search space to those pairs, and other, of solutions that obey the NCS ***(9)***. Most programs implement this as a "locked rotation function" for closed groups of NCS operators ***(10–12)***.

2.2. AMoRe

Automated molecular replacement (AMoRe) ***(13,14)*** is a part of the CCP4 ***(15)*** package of crystallographic software (www.ccp4.ac.uk). This can be run through the graphical user interface provided with the CCP4 package (CCP4i) in a highly automated (auto-amore) or more user-interactive mode in which the individual steps are carried out one at a time. It can also be run the old-fash-

ioned way (i.e., using command scripts). The steps for running the program are: sorting, tabling, rotation search, translation search, rigid-body refinement, and reorientation and will be described next.

2.2.1. Sorting and Tabling

The sorting step just extends the F_{obs} to cover a hemisphere of reciprocal space and then reformats them for use with the program. This step can also be used to reformat structure factors as a table, e.g., when using an electron density map as a search target or when preliminary phases are to be used for a phased translation search. The tabling function creates a structure factor table for use in the rotation, translation, and rigid-body fitting steps. This table is used for the interpolation of F_{calc} in all subsequent steps, thereby increasing the speed of the program. Before the structure factor calculation, the model is shifted such that its center of mass lies at the origin and rotated such that its principal inertia axes are parallel to *a*, *b*, and *c*. This allows the program to determine the minimal box that will surround the molecule and use this as a basis for determining the enormous unit cell (default = 4*minimal box) used to create the structure factor table. The XYZOUT flag for this program will output the rotated and translated coordinates for later use.

2.2.2. Using the Rotation Function to Determine the Correct Orientation

AMoRe uses the fast rotation function derived by Crowther ***(16)*** and improved by Navaza ***(17–19)***. This expands the rotation function in spherical harmonics and thereby allows the entire rotation function to be calculated by a single FFT (once the harmonics are calculated, which does take time). The program outputs the rotation angles and several diagnostic indicators (CC_F, the correlation coefficient between F_{obs} and F_{calc}, RF_F, the R-factor, CC_I, the correlation between intensities, and CC_P, the correlation between observed and model Pattersons) (*see* **Table 1**).

Note how well separated (as judged by CC_F and CC_I) the correct orientation(s) are from the incorrect solutions, and yet how modest the correlation between F_{obs} and F_{calc} is as a result of the phase shift caused by their incorrect placement in the unit cell. Also, it is clear that the Patterson correlation (PC) is not a great discriminator relative to the other correlation statistics. This is an artificially good test case, and most often the distinction is not so clear, necessitating a search of many orientations in the translation search (*see* **Notes 5** and **6**).

2.2.3. The Translation Search, Rigid-Body Fitting, and Reorientation

The program offers several different flavors of translation function, including the Crowther–Blow fast translation function ***(4)***, the Harada–Lifchitz translation function ***(6)***, the Correlation Coefficient search ***(20)*** (most sensitive, but slowest),

Table 1
Example Rotation Search Output From AMoRe

	ITAB	ALPHA	BETA	GAMMA	TX	TY	TZ	CC_F	RF_F	CC_I	CC_P	Icp
SOLUTIONRCD	1	173.99	49.35	277.75	0	0	0	27.2	52.4	46.6	21.9	1
SOLUTIONRCD	1	6.3	49.26	261.63	0	0	0	27.2	52.3	46.6	21.5	3
SOLUTIONRCD	1	19	15.37	246	0	0	0	14.7	56.2	28.9	19.4	11
SOLUTIONRCD	1	161	14.87	294	0	0	0	14.5	56.3	28.5	19.4	12
SOLUTIONRCD	1	100.37	6	175.61	0	0	0	14.4	56.2	29.6	20.5	13
SOLUTIONRCD	1	2.85	80.42	267.03	0	0	0	13.9	56.5	28	19.5	14
SOLUTIONRCD	1	176.07	80.42	273.35	0	0	0	13.9	56.5	28.7	19.7	15
SOLUTIONRCD	1	169	15.81	98.5	0	0	0	13.8	56.5	27.8	18.9	16
SOLUTIONRCD	1	13.5	15.77	79	0	0	0	13.7	56.6	28.4	19	17
SOLUTIONRCD	1	278.73	0	0	0	0	0	13.5	56.5	26.9	19.4	18
SOLUTIONRCD	1	98.1	0	0	0	0	0	13.4	56.5	26.9	19.4	19
SOLUTIONRCD	1	99.58	0	0	0	0	0	13.4	56.7	26.4	19.4	20
SOLUTIONRCD	1	81.79	0	0	0	0	0	13.3	56.5	26.4	19.3	21
SOLUTIONRCD	1	0.49	45.79	89.81	0	0	0	13.2	56.4	21.6	19.4	22
SOLUTIONRCD	1	260.72	0	0	0	0	0	13.1	56.6	26.2	19.3	24
SOLUTIONRCD	1	10.83	56.92	86.01	0	0	0	13	56.6	26.3	18.9	25
SOLUTIONRCD	1	101.04	13.69	173.78	0	0	0	13	56.3	28.8	20.7	26
SOLUTIONRCD	1	82.62	13.69	1.85	0	0	0	13	56.3	28.8	21.1	27

and a phased translation search *(21,22)*, where the phases come either from a partial solution (i.e., one of *n* molecules in the ASU) or an experimental source like multiwavelength anomalous diffraction or multiple isomorphous replacement (MIR). AMoRe does not explicitly examine the packing in a molecular replacement search, but it does output the minimal distance between symmetry mates, which can be used as a rough packing discriminator. When searching for one molecule, AMoRe searches (and this is true for all programs of this type) only the minimal volume required to give a unique solution, otherwise known as the Cheshire cell *(23)*. The symmetry of the Cheshire cell (and therefore the limits of the search) is inversely related to the crystal symmetry. For example, in P1 the Cheshire cell has infinite symmetry (i.e., all points are equivalent), meaning no translation search is needed, in P2 the search covers $0 \leq x \leq 0.5$, $0 \leq z \leq 0.5$, with *y* being arbitrary, and in P222, the search extends from 0 to 0.5 in *x*, *y*, and *z*. Once the first molecule is placed, the origin is fixed and subsequent searches must cover the entire unique volume relative to a lattice point (i.e., the whole unit cell, unless it is a centered space group). The translation search from our above example, using the Correlation target, is as follows (*see* **Table 2**).

Now the two closely related solutions are clearly better than all of the false ones and the correlation is now quite good for the correct solution. Rigid-body fitting of the search model can be used to optimize the rotational and translational parameters. In AMoRe, the difference between the squared amplitudes of the model and observed structure factors is minimized with respect to scale, B-factor, and rotation and translation parameters *(24)*. Whether or not this is a good idea given the desire to overcome model bias and the absence of a free subset of the observed data in this calculation is an open question, and one that is not going to be answered here. Note that all structure refinement packages have a rigid-body refinement protocol in which a free subset can be used to assess its usefulness. Once the final orientation and position has been determined, one needs the coordinates in order to move on to subsequent steps. The "reorientate" stage requires the input data file with the new cell parameters and the shifts applied to the original model in the tabling stage, plus the rotation and translation output from the previous stage of the program. The output rotation and translation can then be applied to the original model with the CCP4 program pdbset.

2.3. Combined Molecular Replacement

Because the previous sections covered many topics relevant to running molecular replacement packages in general, this section will focus on the unique aspects of combined molecular replacement (COMO; *[5]*) and the other programs described in this section. COMO (http://como.bio.columbia.edu/tong/Public/Como/como.html) can be run through a Tcl/TK interface or punch cards (command scripts). It is a stand-alone program and not part of a larger software suite. The COMO *(25)* approach is not extremely different from

Table 2
Example Translation Search Output From AMoRe

	ITAB	ALPHA	BETA	GAMMA	TX	TY	TZ	CC_F	RF_F	CC_I	PKCOUNT	DMIN
SOLUTIONTF1_1	1	173.99	49.35	277.75	0.1981	0.4264	0.0065	71.7	34.4	75.6	1	18.8
SOLUTIONTF1_2	1	6.3	49.26	261.63	0.1979	0.0741	0.4929	69.8	35.5	73.8	1	18.9
SOLUTIONTF1_9	1	13.5	15.77	79	0.0088	0.02	0.255	21.1	55.7	26.9	1	2.7
SOLUTIONTF1_7	1	176.07	80.42	273.35	0.3791	0.2471	0.2609	20.8	54.7	27.4	13	29.4
SOLUTIONTF1_8	1	169	15.81	98.5	0.0087	0.4797	0.2449	20.7	55.6	26.2	1	2.7
SOLUTIONTF1_11	1	98.1	0	0	0.474	0.4195	0.2743	20.7	55	25.6	14	10.3
SOLUTIONTF1_13	1	81.79	0	0	0.026	0.4202	0.2257	20.7	55.1	25.5	8	10.3
SOLUTIONTF1_18	1	82.62	13.69	1.85	0.4638	0.1245	0.3238	20.5	54.5	27.2	2	23.2
SOLUTIONTF1_6	1	2.85	80.42	267.03	0.3797	0.2537	0.2389	20.4	55.2	26.8	9	29.3
SOLUTIONTF1_5	1	100.37	6	175.61	0.0175	0.0863	0.2231	20.1	54.9	27	2	10
SOLUTIONTF1_10	1	278.73	0	0	0.0254	0.0805	0.2743	20.1	55	25.1	18	10.2
SOLUTIONTF1_3	1	19	15.37	246	0.3082	0.328	0.2345	20	55.5	25.2	11	46.6
SOLUTIONTF1_17	1	101.44	13.69	173.78	0.4647	0.3766	0.1853	19.9	54.6	26.8	3	23.4
SOLUTIONTF1_15	1	260.72	0	0	0.474	0.0794	0.2257	19.6	55.1	24.4	20	10.3
SOLUTIONTF1_4	1	161	14.87	294	0.3088	0.1725	0.2654	19.5	55.5	24.8	11	46.4
SOLUTIONTF1_12	1	99.58	0	0	0.1037	0.242	0.2372	19.1	55.1	24.2	20	25.4
SOLUTIONTF1_14	1	0.49	45.79	89.81	0.48	0.3612	0.2771	19.1	58.2	21.4	11	10.3
SOLUTIONTF1_16	1	10.83	56.92	86.01	0.3223	0.0464	0.0362	18.5	55.6	24.9	1	16.9

the uncombined one, but it is more seamless. COMO does away with the idea that correct orientations will have a high rotation function score and instead subscribes to the notion that the best assessment of an orientation is by its behavior in the translation function. Therefore, multiple rotation angles (number specified by the user) are automatically checked in the translation function calculation until the correct solution is (hopefully) found.

The rotation function used is the Crowther fast rotation function ***(16)*** and the translation function is based on the Harada–Lifchitz translation function ***(6)***, and uses a large-term cutoff to improve its speed ***(26)***. The translation function searches with only about 10% of the largest terms and calculates the CC and R-factor with approx 50% of the reflections. Perhaps ignoring the weak reflections improves the signal-to-noise ratio in the translation search. Parameters like the radius of integration and the model unit cell are determined automatically by the program, thus requiring little input from the user save the name and format of the reflection file, the name of the search model file, and the space group and cell dimensions. The data determined by COMO can be input manually if desired, and multiple values of such parameters as the radius of integration (three radii) and resolution range (two sets of limits) can be tested and compared to improve the elimination of false solutions. The program incorporates a packing check in which only models that have a small number (preferably 0) of close C_α-C_α contacts are kept as potential molecular replacement solutions to be examined. The top solution(s) are optimized by rigid-body refinement and output as a PDB file. COMO will search for *n* copies of the input model, where *n* is specified by the user. These are searched for sequentially, with the top solution(s) from one round fixed in subsequent rounds of searching.

The most unique and advantageous feature of this program is the fact that it automatically searches neighboring grid points when the translation function finds a top solution as a means of finding the optimal rotational and translational parameters ***(5)***. This procedure allows the search to gravitate toward optimal solutions, even in some cases in the presence of substantial errors. In the difficult test case published by the authors, the program was able to migrate from an initial answer (for the fifth molecule, with four fixed, so plenty of initial phase info helped this search) that had errors of 26° in θ_1 and 3° in θ_2 to the correct solution. The program outputs a log file that looks like the following:

```
TFPOPC> The final top 5 solutions
TFPOPC> No  th1   th2   th3     x      y      z      CC     R  Cont TF RF Mol
TFPOPC>  1  14.1  12.0  118.1  0.817  0.438  0.055  31.9  43.6  0   1  1   0
TFPOPC>  2  11.2  12.0  120.9  0.817  0.438  0.055  31.8  43.7  0   1  6   0
TFPOPC>  3  16.9  12.0  115.3  0.817  0.438  0.055  31.0  44.0  0   1  2   0
TFPOPC>  4   8.4  12.0  123.8  0.817  0.438  0.055  30.6  44.4  0   1 18   0
TFPOPC>  5  11.2  15.0  123.8  0.817  0.438  0.055  30.2  45.0  0   1 36   0
```

continues

```
TFPOPC> 98.5% of possible solutions rejected by packing
SOLSEL> Solution number 1 selected
SOLSEL> Rotation   =   14.06     12.00     118.12
SOLSEL> Translation =  0.8167     0.4375     0.0547
RGCALC> Rigid-body optimization of this solution
RGCALC> Correlation coefficient before and after refinement  31.92   34.06
RGCALC> Optimized rotation angle      14.48     11.50     118.62
RGCALC> Optimized translation          0.8162    0.4359     0.0540
```

Once again the quality of the solution is judged by the CC (on Fs) and R-factor statistics.

2.4. CNS

Once again, most of the specifics detailed for molecular replacement calculations in CNS (http://cns.csb.yale.edu) would deal with running the CNS package *(27)* itself, so this section will explain the unique features of the program with regard to molecular replacement. CNS has two unique aspects to its implementation of molecular replacement. First, it employs two rotation functions not commonly used elsewhere, the real space *(28)* and direct rotation functions *(29)*. Second, CNS implements a PC refinement *(30,31)* procedure before the translation step which allows the search model to be broken into a variety of subfragments that are each allowed to move as rigid bodies.

2.4.1. The Real-Space and Direct Rotation Searches

The real-space rotation search is exactly what its name implies. The Patterson map for the model structure is rotated and compared with that for the target structure, and the rotation function evaluated via the overlap integral:

$$Rot(\Omega) = \int_U P_{obs}(u) P_{calc}(\Omega u) du \tag{4}$$

The area over which the integration proceeds (U) typically excludes a small sphere around the origin peak and includes a larger sphere surrounding the origin (this is identical to the integration sphere seen elsewhere in this chapter). This calculation is significantly slower than the Crowther fast rotation function, but is much faster than the direct rotation search. However, the increased speed of this calculation compared with its direct counterpart comes at a price, namely approximations are made to speed up the calculation. First, only a subset of the model Patterson peaks are rotated for the calculation, typically only the highest 3000. Second, rotation often leads to the resultant map having points of calculated density fall in between points on the grid. Because we can only assess the overlap using points for which the target Patterson has a value (i.e., *points on the grid*), the corresponding density values in the rotated Patterson

must be interpolated from the off-grid points. Interpolation can be quite error-prone, leading to inaccuracies in calculating the overlap.

The direct rotation function is also aptly named. This calculation involves rotating the model, calculating structure factors, normalizing them (normalized structure factors are called "E"s and do not have the same resolution-dependent falloff that "F"s have), and calculating the correlation coefficient between their squared amplitudes:

$$CC(\Omega) = \frac{\sum_H (|E_{H,obs}|^2 - \langle |E_{obs}|^2 \rangle)(|E_{H,\Omega,calc}|^2 - \langle |E_{\Omega,calc}|^2 \rangle)}{\left[\sum_H (|E_{H,obs}|^2 - \langle |E_{obs}|^2 \rangle)^2\right]^{1/2} \left[\sum_H (|E_{H,\Omega,calc}|^2 - \langle |E_{\Omega,calc}|^2 \rangle)^2\right]^{1/2}} \quad (5)$$

This calculation is computationally very expensive, but involves no approximations. If the real-space search were carried out without approximations, it would be equivalent to the direct rotation search. The direct rotation search typically has a better signal-to-noise ratio, but because it is about an order of magnitude slower, it is only used when the real space search fails.

2.4.2. PC Refinement

Once a list of candidate orientations has been obtained and before the translation search, CNS attempts to optimize both the rotational parameters and the makeup of the search model via PC refinement. Remember that the translation function is very sensitive to errors in the orientation, so the use of PC refinement to eliminate these errors is a good idea. In this procedure, the user specifies how to apportion the search model as rigid bodies. Typically, individual subdomains are the fragments used, although the user can break the model down to individual helices, and so on if so desired. This can overcome differences that are not as simple as individual domains being slightly misaligned. The target for this minimization is usually the correlation between normalized structure factors (same as **Eq. 5**). The PC refinement procedure differs from standard rigid-body refinement only in that it is performed in the absence of crystallographic symmetry. PC refinement can drastically reduce the noise in translation function maps, leading to an easier discrimination of the solution from false peaks (**Fig. 4**).

2.5. Molrep

Molrep (http://www.ysbl.york.ac.uk/~alexei/molrep.html) ***(32)*** is another automated molecular replacement package that comes as a part of the CCP4 suite. It contains many unique features, some for standard molecular replacement, some for phased searches (which will be discussed in **Subheading 3.**), and some for superposing models, but this section will only examine standard molecular replacement. One feature of Molrep that is not standard in most other

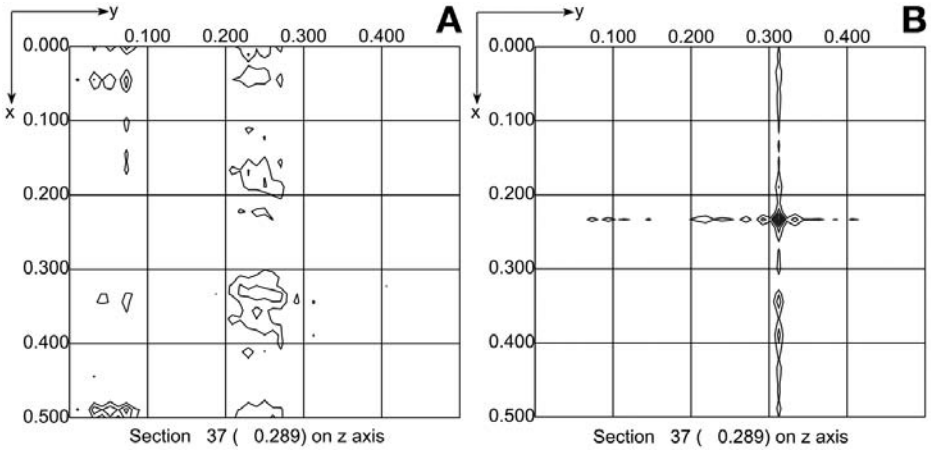


Fig. 4. PC refinement can drastically reduce the noise in translation function maps. **(A)** A section of the translation function for a molecule without PC refinement. There is no clear solution in this section. **(B)** The same section of the translation function after PC refinement optimization of the rotational parameters. Note the clear solution. The maps are scaled differently because of the extreme disparity in peak heights between the two.

programs is the option to use a map as a molecular replacement model. In most other programs, this requires some serious machinations. The most important feature of Molrep for standard molecular replacement is the multicopy search method developed by Vagin and Teplyakov *(33)*. The general method involves construction of a multicopy search model from an oriented (by a Crowther fast-rotation search) search model by using a "special translation function," (STF). The STF proceeds as follows, for example considering two monomers, referred to by the authors as a *dyad*. The goal of this translation function is to find, for every candidate orientation, the intermolecular vector that relates the two monomers (S_1 and S_2) of the dyad. This assumes that both members of the dyad are described by peaks in the input rotation function. The centers of mass for the two monomers are defined by the vectors $\vec{s}_1$ and $\vec{s}_2$, thus, the vectors $(\vec{s}_1 - \vec{s}_2)$ and $(\vec{s}_2 - \vec{s}_1)$ must be found. The evaluation of STF is via the real-space overlap of electron densities:

$$STF(\vec{s}_i - \vec{s}_j) = \int_{cell} P_{obs}(r) P_{calc}(\vec{s}_i - \vec{s}_j, r) dr \qquad (6)$$

P_{obs} (r) is the Patterson map for the observed data at a point r in the unit cell and $P_{calc}(\vec{s}_i - \vec{s}_j, r)$ is the Patterson function for that particular dyad. For exam-

ple, the coefficients used for the Patterson to determine the vector $(\vec{s}_1 - \vec{s}_2)$ are $F_1(\vec{h})F_2^*(\vec{h})$. The previously mentioned overlap function is essentially the same as that used to derive the phased translation function (**Subheading 3.**), lacking only the denominator of the correlation function. It is also of the same form as **Eq. 4**, the real-space rotation function, but differs in that it is carried out over the whole unit cell instead of a sphere centered at the origin and therefore incorporates translational information. Also, the coefficients that would yield the self vectors (i.e., $F_1(\vec{h})F_1^*(\vec{h})$ and $F_2(\vec{h})F_2^*(\vec{h})$) are eliminated from the calculation, making it a translation-only overlap function. Once the dyad vectors are determined, Molrep performs a conventional translation search to determine their position in the unit cell. Using this procedure, the search for any number of molecules in the asymmetric unit can be broken down into a series of dyad searches. This approach is entirely general and does not require that the NCS operators form a closed group. Once again the CC and R-factor are the statistics of choice.

3. Six(ish)-Dimensional Searches, Prior Phase Information, and Maximum Likelihood

3.1. Advantages and Disadvantages of 6D Searches vs Traditional Searches

Earlier in this chapter, the computationally intractable nature of full rotation and translation searches was referred to, at least when searched in fairly fine intervals (1° and 1 Å). Fortunately, this did not deter people from attempting 6D searches and developing novel ways to approach them. The old program BRUTE ***(20)*** simply searched the five or six dimensions necessary to find a solution, and was very slow, even on a coarse grid. Several groups have since found ways to make the simultaneous rotation and translation search fast enough to be completed in a reasonable amount of time (that is the good news), and they will be discussed in detail next. A technical note on the simultaneous searches is in order before we proceed. Because both the orientation and translation are searched simultaneously, all of the low-resolution data are used, and, thus, their accuracy is necessary for these methods to succeed. Take care to collect good low-resolution data, regardless of how you intend to solve your structure. One aspect that all of these methods have in common, is that they interpolate the structure factors of the rotated and translated model as a means of increasing the speed of the calculation ***(34)***. Unfortunately, this has the effect of eliminating useful modifications of the search model during the search (e.g., PC refinement). The accuracy of this interpolation is dependent on sampling the original structure factor calculation on a very fine grid. Once this initial FFT is

performed, any subsequent rotation and translation of the search model can be approximated by rotating the indices of the reciprocal lattice in the opposite direction, applying the appropriate phase shifts owing to translation, and interpolating the new structure factors from the off-lattice points that result from the rotation. The bad news is that they are still slow, especially when compared with the separate rotation and translation searches. The other disadvantage is that these methods ignore all information from the Patterson function and treat all orientations as equally likely, which is unrealistic. This increases the search space by including impossible packing arrangements, thereby exacerbating the speed problem. There are, however, some major advantages to performing the rotation and translation searches simultaneously. First, the sensitivity of the translation search to minor errors in the rotational parameters is alleviated because both parameters are optimized simultaneously. Second, the signal to noise is drastically increased over the separate searches because those orientations selected are not phase shifted because of undetermined translational parameters. Thus, the program is not searching for the best of a bad bunch of solutions (as exhibited by low CC scores of even perfectly oriented but mistranslated models) but rather searching for a solution that, when found, will give a good correlation between calculated and observed structure factors. Finally, if the molecule has an irregular shape or is densely packed in the crystal, it will no longer be feasible to separate the self and cross Patterson vectors, thus leading to a lot of noise in both the rotation and translation searches. Because there is no need to separate these data in the 6D searches, they are not much affected by unusually shaped or close-packed molecules. I'll now discuss three programs that conduct six(ish)-dimensional searches: EPMR, Queen of Spades, and SOMoRe.

3.2. EPMR

EPMR (available at www.epmr.info *[35,36]*) uses a relative of the genetic algorithm that was originally used by Chang and Lewis *(37)* in a stochastic 6D search. It is available as a stand-alone program and requires little user input, especially in its default mode of operation. EPMR starts with a population of potential solutions generated by applying random perturbations to the starting orientation and translation. A stochastic tournament is then conducted to see which solutions survive to the next generation. These solutions are then used to form a new population of potential solutions by adding normally distributed (i.e., Gaussian) random perturbations to the orientation and translation of the parent population. This procedure is iterated for a user-specified number of generations and then the solution with the highest CC score is rigid-body refined and output as the "solution." A new, soon-to-be-released version of EPMR promises to have more advanced features than the currently used version

(*see* **Note 7**). Because the search is stochastic as opposed to deterministic, it is apt to settle on different solutions if run multiple times. However, if the same solution keeps appearing multiple times over many trials, it is likely to be correct. The correlation between observed and calculated normalized structure factors was found by the authors to be the most sensitive criterion for assessing a potential solution ***(36)***. A further advantage that this approach has over the "divide and conquer" approaches is its enhanced sensitivity when the search model is incomplete or has low sequence homology.

3.3. Queen of Spades

Queen of Spades ***(38,39)*** is another stochastic six(ish)-dimensional search program available (http://origin.imbb.forth.gr/~glykos/Qs.html) as a standalone package. The program can be run with as little information as the names of the data and search model files, and the number of molecules to be found. Instead of using an evolutionary search procedure to determine each generation's rotational and translational parameters, the authors use a variation on the inverse Monte Carlo method in which the target function is the R-factor or (1-CC) on either structure factors or intensities. The simulated annealing approach has been used for quite some time in structure refinement ***(40)***, and in one study proved useful in placing fragments in electron density maps ***(41)***. The program starts with a random orientation and position, and then applies a random perturbation and compares the target function values for the new orientation and position. If the target function value is lower, then the new orientation is accepted and further perturbed. If the target function value is higher, then the new position is accepted under the following conditions (e.g., for the R-factor):

$$e^{\left([R_{old} - R_{new}]/T\right)} > \xi \tag{7}$$

where T is a control parameter ("temperature," but in the metaphoric, not the physical sense) that decreases as the minimization advances, and ξ is a random number between 0.0 and 1.0. Note that any time that $R_{old} - R_{new}$ is positive (i.e., the target function decreases) the exponential will be greater than one and, therefore, meet the acceptance criterion.

The key to getting this method to give the correct answer is choosing the right annealing schedule and range, and the right move size (i.e., how much to perturb the current state to obtain a new trial structure). These parameters can be determined by the program or input by the user. The annealing schedules that are supported by the program are (1) keep the temperature constant or do no annealing, (2) a linear decrease of the temperature with time, (3) a logarithmic schedule in which at each step k, the temperature $T(k)$ is $(T_0/\log k)$, where T_0 is the starting temperature, and (4) the temperature of the system at each step is

adjusted to keep the fraction of moves against the target function gradient equal to a constant user-defined value. The move size can be kept at a constant resolution-dependent value or be dependent on the current state of the target function. The best combination of the previously mentioned parameters to use is going to be dependent on the particulars of the system under investigation. The program has recently been modified to support the simultaneous search for multiple models, either several copies of the same model or copies of several different proteins and/or nucleic acids *(42)*.

3.4. SOMoRe

SOMoRe *(43)* (http://www.caam.rice.edu/~djamrog/somore.html), unlike EPMR and Queen of Spades, takes a deterministic approach to solving a 6D molecular replacement search. The program first performs an exhaustive low-resolution search (either 10 or 8 Å) and uses the top *n* (default 1000) points from the coarse search to do local optimizations with data extending to 4-Å resolution. In both instances the target function is the CC between the observed and calculated intensities (there is the option of making the target function the R-factor or the CC between observed and calculated structure factors). The rationale behind this search scheme is that at low resolution, the landscape of the target function is likely to be much smoother and have fewer local minima, thereby making it possible that one in approx 1000 of the best-scoring grid points is close enough to the global minimum to reach it by a simple minimization procedure. The packing for each solution that has a good target function score is assessed explicitly as a function of close contacts either between C_α atoms or between all atoms. The mechanics of the program, as well as many of the keywords for input are nearly identical to those from Queen of Spades, making it easy to learn this program if you are already familiar with Queen of Spades. Using this approach, the authors were able to solve a molecular replacement problem with a model that was more truncated than that allowed by EPMR (and far more than that allowed by X-PLOR or AMoRe). The program is extremely slow, but because it is deterministic, it will find the right answer (if it exists) on each run, whereas the stochastic methods are faster, but might have success rates as low as 5%.

3.5. Searches With Pre-Existing Phase Information

Any available phase information, be it from experimental phasing (i.e., multiwavelength anomalous diffraction, MIR, etc.) or a partial molecular replacement solution can be of great assistance in solving molecular replacement problems. Experimental phase information in particular, even if poor, can contain enough unique features to make positioning of a search model feasible when other methods have failed. This section will consider two cases: (1) the use of a phased translation function to place a well-oriented search model that fails to

give a clear solution to the translation function and (2) positioning of a search model in experimental density using the program Molrep.

3.5.1. The Phased Translation Function

When a properly oriented model is translated to its correct position, the calculated electron density for that model should achieve maximum overlap with the electron density derived from the experimental-phase information. This can be assessed by the correlation coefficient between the two electron densities ***(22)***:

$$C(t) = \frac{\int_V \rho_{obs}(\vec{x})\rho_{calc}(\vec{x}-\vec{t})dx}{\left[\int_V \left(\rho_{obs}(\vec{x})^2 dx\right)\int_V \left(\rho_{calc}(\vec{x})^2 dx\right)\right]^{1/2}} \tag{8}$$

This is the same form as any simple correlation coefficient, but looks simpler because, since the F_{000} is omitted from the map calculation, $\bar{\rho}$ is zero for both the experimental and model densities. Read and Schierbeek also derived a reciprocal space formulation of the phased translation function ***(22)***:

$$C(t) = \frac{V^{-1}\sum_{\vec{h}} m_p \left|F_{obs}(\vec{h})\right|\left|F_{calc}(\vec{h})\right| e^{i(\alpha_p - \alpha_{calc})} e^{-2\pi i\vec{h}\cdot\vec{t}}}{1/V\left[\sum_{\vec{h}} m_p \left|F_{obs}(\vec{h})\right|^2 \sum_{\vec{h}} \left|F_{calc}(\vec{h})\right|^2\right]^{1/2}} \tag{9}$$

This function can easily be evaluated by an FFT (e.g., the program FFT from the CCP4 suite ***[15]***). The observed amplitudes, phases, and figures of merit (m_p) must be expanded to space group P1 and the model amplitudes and phases must be calculated in the same P1 unit cell. The coefficients $[m_p|F_{obs}(\vec{h})||F_{calc}(\vec{h})|]$ and ($\alpha_p - \alpha_{calc}$) or ($-\alpha_p - \alpha_{calc}$) if the experimental phases have the wrong hand) are then the input amplitudes and phases for a P1 map calculation. The highest peak in the map should correspond to the translation vector that must be applied to the oriented model. Note that the α_p do not necessarily have to come from experimental phasing and can be a partial molecular replacement solution, in which case an estimate of the figure of merit for each reflection would be made based on the σ_A statistic.

Read and Schierbeek reported two instances in which the phased translation function greatly outperformed the standard translation function ***(22)***. In one case using MIR phases, the phased translation function was able to correctly identify the translation vector even when the model was misoriented by 6.9°, whereas the standard rotation function could only tolerate 3.5° of orientation error. In the second case, even with single isomorphous replacement phases (without solvent flattening, so the ambiguity was not even broken), the phased

translation function was able to easily identify the correct translation vector (2.9° orientation error), whereas the highest peak in the standard translation was an incorrect translation vector.

3.5.2. Placing Models in Experimental Density With Molrep

The placement of a search model in experimental density as implemented by Molrep is its most unique feature. In the procedure designed by the authors, the experimental-phase information is utilized at all stages of the search process instead of just at the translation step. The first step of the procedure is an attempt to find the approximate center of mass of the protein in the electron density. The authors utilize a spherically averaged phased translation function, which yields potential centers of mass for unoriented models ***(44)***. The form of the spherically averaged phased translation function is:

$$SAPTF(\vec{s}) = \int_0^a \hat{\rho}_{obs}(\vec{s},\vec{r})\hat{\rho}_{calc}(\vec{r})dr \qquad (10)$$

You'll notice that this looks like a modified version of an electron density overlap integral. The "^" symbol denotes spherical averaging and the procedure is as follows: for each point $\vec{s}_1$ in the experimental electron density, the density is averaged in a spherical volume with radius *a* and this density is compared with the density from the model spherically averaged within the same radius around its center of mass. A high score corresponds to overlapping regions of high density, which would presumably correspond to protein. Thus, the point $\vec{s}_1$ in the electron density is taken to be the center of mass of a potential solution. The authors have expanded the previously mentioned expression in spherical harmonics ***(44)***, making it calculable as an FFT. Once the potential centers of mass have been identified, the model is placed at these points and a phased rotation function is performed to identify the best orientation for the search model at each point. The phased rotation function is carried out in real space and consumes most of the CPU eaten by the program. Once the best orientation is found, the program does a standard phased translation ***(21)*** to tidy up any errors in the estimation of the center of mass of the model. This method, while not foolproof, has proven (in the author's experience) useful in placing models in some ratty electron density.

3.6. Maximum Likelihood Molecular Replacement in Phaser

The most likely model is one that maximizes the probability that the experimental observations were made. This simple statement is the overriding principle behind maximum likelihood. Over the past 10 yr, the statistical method of maximum likelihood has come to dominate most aspects of struc-

ture determination, most prominently experimental phasing and structure refinement. As the success of these efforts became apparent, the idea of using maximum likelihood to improve the sensitivity of molecular replacement, first proposed by Bricogne ***(45)***, became the next target in the sights of its proponents. Much of the theory behind implementation of maximum likelihood in molecular replacement is beyond the scope of this chapter, so I will attempt to incorporate enough to show how this approach differs from the traditional approaches.

The very short answer to the question "what is different about maximum likelihood molecular replacement?" is that the errors (i.e., lack of sequence identity and incompleteness) are handled explicitly by the rotation and translation target functions ***(46)***. Of course, the long answer involves describing the characteristics of the likelihood function and how these errors are treated. For either a 6D search or a translation search, the appropriate likelihood function would be any of the ones currently in use for structure refinement (discussed in Chapter 13). The only difference is that, instead of adjusting parameters for what is hopefully a pretty good model, the program will be evaluating models that have no resemblance to the truth vs models for which the calculated structure factors represent a reasonable approximation to the true structure factors. If this target function can discriminate between structures with minor differences like a few side chains that are missing or not positioned properly, discriminating correct from incorrect molecular replacement solutions should be relatively easy. However, calculating the log likelihood for each point in a 6D search or even a translation search is going to be computationally unattractive, as are brute force 6D or translation searches.

To "divide and conquer" maximum likelihood molecular replacement means deriving a likelihood function for a model with unknown position ***(47)***, and, therefore, with an unknown contribution from symmetry-related copies to each complex structure factor (i.e., the amplitudes of the contributions are known, but the phases are not, so the model structure factors cannot be calculated by a simple summation of each symmetry-related molecular transform). Therefore, a probability function for the calculated amplitudes needs to be used. In the current version of Phaser (http://www-structmed.cimr.cam.ac.uk/phaser/index.html), two types of distributions are used. One assumes that all copies of the molecule in the unit cell contribute independently to the calculated structure factor, and therefore models this probability as a two-dimensional Gaussian in the complex plane centered at its origin, much like the Wilson distribution. The other probability function is similar to the Rice distribution, and assumes that one contribution to the calculated structure factor dominates all others. Thus, the probability distribution is modeled as a two-dimensional Gaussian in the complex plane, centered at the end of the dominant component, called F_{big}. Like the direct

rotation function from **Subheading 2.**, this type of brute-force rotation search will be computationally slow, so the authors of Phaser have implemented a fast rotation function version of the likelihood rotation function (in addition to the brute force one) by expanding it in a Taylor series. The function that results from this treatment is what the authors term "a scaled and variance-weighted version of the Patterson overlap function used in the Crowther target" *(47)*. In other words, the model coefficients are modified based on the expected errors in the model arising from incompleteness and coordinate error, effectively smearing the atoms over their possible positions and then calculating the Patterson function. The program also employs a fast translation function equivalent of the translation likelihood function in addition to its brute-force kin. In a typical run of the program, the fast rotation and translation functions are used to identify potential solutions, which are then rescored using the full likelihood functions to determine the best solution. This represents a speedy compromise in which the accuracy of the likelihood function is utilized, but only for those spots in parameter space deemed worthy of further investigation.

4. Notes

1. The Fourier coefficients from the model are calculated by placing the model at the origin of a very large P1 cell, making the chances of cross vectors falling within a reasonably well-chosen sphere remote.
2. In P1, which has no symmetry, the translation function calculation is unnecessary.
3. When a screw axis ambiguity exists (e.g., $P6_1$ vs $P6_5$) the translation function should be performed in all possible members of that point group and centering type. The translation function should easily discriminate between the correct and incorrect space groups. If not, something is wrong.
4. Default resolution limits vary from program to program. A typical low-resolution cutoff would be 8 or 10 Å, so for lower resolution searches, either 10–6 or 8–6 Å are good limits to try. In normal cases, the high-resolution limit for the rotation search is 4 Å, but 3 Å can be used if the data warrant it.
5. If the top solution is not clearly separated from the rest, it is best to try as many peaks from the rotation function output that your patience will allow. The maximum number of output peaks in AMoRe is 99. With the current speed of computers, checking all 99 peaks in the translation function is not prohibitive, so you might as well check them all.
6. The two most commonly used angular systems are eulerian and spherical polar. In the eulerian system (α, β, and γ), the rotations are carried out sequentially about the axes determined by the convention being used by the author of the program. Typically, this is *zyz*, which means that first one rotates by α degrees about *z*, then β degrees about the new *y*, and then γ degrees about the new *z*. It is important to know which convention is being used if you try to write something that performs the rotations outside of the program. Otherwise, you will not be able to reproduce that program's results. The spherical polar angular system (ϕ, ψ, and κ) is such that

ϕ and ψ describe the axis of rotation (ϕ is the rotation in the *xy*-plane and ψ is the rotation out of the *xy*-plane toward *z*) and κ describes the angle of rotation.

7. The newest version of EPMR will allow the automatic evaluation of multiple search models, either by running each sequentially or by having them compete with each other during the search. It will have a similar treatment for different space groups. In addition to the normal sequential search, the new version of EPMR will allow for a simultaneous search for multiple molecules. The searches can be carried out in either reciprocal or real space. The program will be designed such that its compilation easily allows use of multiple processors (e.g., in a cluster).

Acknowledgments

Thanks to Tommy Joe Hollis for getting me roped into this endeavor.

References

1. Hoppe, W. (1957) Die 'Faltmolelciilmethode'—eine neue Merhode zur Bestimmung der Kristallstruktur bei ganz oder teilweise bekannter Molekiilstruktur. *Acta Crystallogr.* **10,** 750–751.
2. Rossmann, M. G. and Blow, D. M. (1962) The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallogr.* **15,** 24–31.
3. Patterson, A. L. (1934) A fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.* **46,** 372–376.
4. Crowther, R. A. and Blow, D. M. (1967) A method of positioning a known molecule in an unknown crystal structure. *Acta Crystallogr.* **23,** 544–548.
5. Jogl, G., Tao, X., Xu, Y., and Tong, L. (2001) COMO: a program for combined molecular replacement. *Acta Crystallogr. D. Biol. Crystallogr.* **57,** 1127–1134.
6. Harada, Y., Lifchitz, A., Berthou, J., and Jolles, P. (1981) A translation function combining packing and diffraction information: An application to lysozyme (high-temperature form). *Acta Crystallogr. A.* **37,** 398–406.
7. Scheringer, C. (1963) Least-squares refinement with the minimum number of parameters for structures containing rigid-body groups of atoms. *Acta Crystallogr.* **16,** 546–550.
8. Sussman, J. L., Holbrook, S. R., Church, G. M., and Kim, S.-H. (1977) A structure-factor least-squares refinement procedure for macromolecular structures using constrained and restrained parameters. *Acta Crystallogr. A.* **33,** 800–804.
9. Rossmann, M. G. (1990) The molecular replacement method. *Acta Crystallogr. A.* **46,** 73–82.
10. Rossmann, M. G. (1972) The locked rotation function. *J. Mol. Biol.* **64,** 246–249.
11. Tong, L. A. and Rossmann, M. G. (1990) The locked rotation function. *Acta Crystallogr. A.* **46,** 783–792.
12. Tong, L. (2001) How to take advantage of non-crystallographic symmetry in molecular replacement: 'locked' rotation and translation functions. *Acta Crystallogr. D. Biol. Crystallogr.* **57,** 1383–1389.

13. Navaza, J. (1994) AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A.* **50,** 157–163.
14. Navaza, J. (2001) Implementation of molecular replacement in AMoRe. *Acta Crystallogr. D. Biol. Crystallogr.* **57,** 1367–1372.
15. Collaborative Computational Project, N. (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D. Biol. Crystallogr.* **50,** 760–763.
16. Crowther, R. A. (1972) The fast rotation function. In: *The Molecular Replacement Method,* (Rossmann, M. G., ed.), Gordon and Breach, New York, NY, pp. 173–178.
17. Navaza, J. (1987) On the fast rotation function. *Acta Crystallogr. A.* **43,** 645–653.
18. Navaza, J. (1990) Accurate computation of the rotation matrices. *Acta Crystallogr. A.* **46,** 619–620.
19. Navaza, J. (1993) On the computation of the fast rotation function. *Acta Crystallogr. D. Biol. Crystallogr.* **49,** 588–591.
20. Fujinaga, M. and Read, R. J. (1987) Experiences with a new translation-function program. *J. Appl. Cryst.* **20,** 517–521.
21. Colman, P. M. and Fehlhammer, H. (1976) The use of rotation and translation functions in the interpretation of low resolution electron density maps. *J. Mol. Biol.* **100,** 278–282.
22. Read, R. J. and Schierbeek, A. J. (1988) A phased translation function. *J. Appl. Cryst.* **21,** 490–495.
23. Hirshfeld, F. L. (1968) Symmetry in the generation of trial structures. *Acta Crystallogr. A.* **24,** 301–311.
24. Castellano, E. E., Olivia, G., and Navaza, J. (1992) Fast rigid-body refinement for molecular-replacement techniques. *J. Appl. Cryst.* **25,** 281–284.
25. Tong, L. (1996) Combined molecular replacement. *Acta Crystallogr. A.* **52,** 782–784.
26. Tollin, P. and Rossmann, M. G. (1966) A description of various rotation function programs. *Acta Crystallogr.* **21,** 872–876.
27. Brunger, A. T., Adams, P. D., Clore, G. M., et al. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D. Biol. Crystallogr.* **54,** 905–921.
28. Huber, R. (1965) Die automatisierte Faltmolekiilmethode. *Acta Crystallogr.* **19,** 353–356.
29. DeLano, W. L. (1995) The direct rotation function: patterson correlation search applied to molecular replacement. *Acta Crystallogr. D. Biol. Crystallogr.* **51,** 740–748.
30. Brunger, A. T. (1990) Extension of molecular replacement: a new search strategy based on patterson correlation refinement. *Acta Crystallogr. A.* **46,** 46–57.
31. Yeates, T. O. and Rini, J. M. (1990) Intensity-based domain refinement of oriented but unpositioned molecular replacement models. *Acta Crystallogr. A.* **46,** 352–349.
32. Vagin, A. and Teplyakov, A. (1997) MOLREP: an automated program for molecular replacement. *J. Appl. Cryst.* **30,** 1022–1025.
33. Vagin, A. and Teplyakov, A. (2000) An approach to multi-copy search in molecular replacement. *Acta Crystallogr. D. Biol. Crystallogr.* **56,** 1622–1624.
34. Huber, R. and Schneider, M. (1985) A group refinement procedure in protein crystallography using fourier transforms. *J. Appl. Cryst.* **18,** 165–169.

35. Kissinger, C. R., Gehlhaar, D. K., and Fogel, D. B. (1999) Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr. D. Biol. Crystallogr.* **55,** 484–491.
36. Kissinger, C. R., Gehlhaar, D. K., Smith, B. A., and Bouzida, D. (2001) Molecular replacement by evolutionary search. *Acta Crystallogr. D. Biol. Crystallogr.* **57,** 1474–1479.
37. Chang, G. (1997) Molecular replacement using genetic algorithms. *Acta Crystallogr. D. Biol. Crystallogr.* **53,** 279–289.
38. Glykos, N. M. and Kokkinidis, M. (2000) A stochastic approach to molecular replacement. *Acta Crystallogr. D. Biol. Crystallogr.* **56,** 169–174.
39. Glykos, N. M. and Kokkinidis, M. (2001) Multidimensional molecular replacement. *Acta Crystallogr. D. Biol. Crystallogr.* **57,** 1462–1473.
40. Brunger, A. T., Kuriyan, J., and Karplus, M. (1987) Crystallographic R factor refinement by molecular dynamics. *Science* **235,** 458–460.
41. Subbiah, S. and Harrison, S. C. (1989) A simulated annealing approach to the search problem of protein crystallography. *Acta Crystallogr. A.* **45,** 337–342.
42. Glykos, N. M. and Kokkinidis, M. (2004) Molecular replacement with multiple different models. *J. Appl. Cryst.* **37,** 159–161.
43. Jamrog, D. C., Zhang, Y., and Phillips, G. N., Jr. (2003) SOMoRe: a multi-dimensional search and optimization approach to molecular replacement. *Acta Crystallogr. D. Biol. Crystallogr.* **59,** 304–314.
44. Vagin, A. A. and Isupov, M. N. (2001). Spherically averaged phased translation function and its application to the search for molecules and fragments in electron-density maps. *Acta Crystallogr. D. Biol. Crystallogr.* **57,** 1451–1456.
45. Bricogne, G. (1992) *CCP4 Study Weekend: Molecular Replacement,* (Wolf, W., Dodson, E. J., and Gover, S., eds.), Daresbury Laboratory, Warrington, UK, pp. 62–75.
46. Read, R. J. (2001) Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallogr. D. Biol. Crystallogr.* **57,** 1373–1382.
47. Storoni, L. C., McCoy, A. J., and Read, R. J. (2004) Likelihood-enhanced fast rotation functions. *Acta Crystallogr. D. Biol. Crystallogr.* **60,** 432–438.

8

Phase Determination Using Halide Ions

Miroslawa Dauter and Zbigniew Dauter

Summary

A short soak of protein crystals in cryosolution containing bromides or iodides leads to incorporation of these ions into the ordered solvent shell around the protein surface. The halide ions display significant anomalous signal, bromides in the vicinity of the absorption edge at 0.92 Å, and iodides at longer wavelengths, e.g., provided by the copper sources. Bromides can, therefore, be used through multiwavelength anomalous diffraction or single-wavelength anomalous diffraction (SAD) techniques and iodides through SAD or multiple isomorphous replacement (MIRAS) phasing. The halide cryosoaking approach involves very little preparative effort and offers a rapid and simple way of solving novel protein crystal structures.

Key Words: Anomalous scattering; MAD; SAD; phasing; halides; bromides; iodides.

1. Introduction

The idea of introducing halide ions to protein crystals originated from the observation that in the crystals of lysozyme, obtained from the standard solution containing 1 *M* of NaCl, several chloride ions reside in the ordered solvent region around the protein molecule *(1)*. Chlorine is a relatively light element and, although it has sometimes been used as a vehicle for phasing *(2–4)*, its heavier analogs, bromine and iodine, are much better suited as anomalously scattering heavy atoms for phasing crystal structures *(5)*. Subsequent trials confirmed that heavier halides, bromides and iodides, can also be introduced to protein crystals, if their salts are present in the crystallization medium *(6)* or in the cryoprotecting solution *(7,8)* and this approach has been proposed as one of the general methods of solving protein crystal structures *(9)*. A number of novel protein structures have recently been solved by this approach, and some examples are listed in **Table 1**.

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

Table 1
Some Novel Structures Solved With Soaked Halides

Protein	Number of residues	Resolution phas/ref (Å)	Method	Ions	Soak	Time (s)	PDB code	Reference
R11 of Man6P/IGFII receptor	286	2.2/1.75	SIRAS	16 I	1 *M* KI	60	1E6F	***(25)***
TIM, *Caenorhabditis elegans*	550	2.0/1.7	SAD	12 I	0.5 *M* NaI	600	1MO0	***(26)***
GRIP1 PDZ6-peptide	97	1.9/1.5	MAD	3 Br	1 *M* NaBr	30	1N7E	***(27)***
Drosophila NLP-core	540	2.8/1.5	MAD	8 Br	0.5 *M* NaBr	20	1NLQ	***(28)***
Noggin	345	2.4/2.4	MAD	10 Br	1 *M* NaBr	30	1M4U	***(29)***
APS kinase	844	1.9/1.43	MAD	13 Br	0.75 *M* NaBr	45	1M7G	***(30)***
KTN	294	3.1/2.85	MAD	4 Br+8 I	1 *M* NaBr/NaI	60	1LSU	***(31)***
NC1 bovine eye lens	2740	2.2/2.0	MAD	33 Br	0.5 *M* KBr	60	1M3D	***(32)***
S100A3	202	2.7/1.7	MIRAS	2 I	0.5 *M* KI	30	1KSO	***(33)***
Interleukin-22	358	1.92/1.9	SIRAS	10 I	0.125 *M* NaI	180	1M4R	***(34)***
Phenylalanine hydroxylase	297	2.4/1.74	MAD	7 Br	0.33 *M* NaBr	50	1LTU	***(35)***
NC1 human placenta	1372	2.5/1.9	MAD	25 Br	1 *M* NaBr	30	1LI1	***(36)***

Doc1/Apc10 subunit	442	2.2/2.2	SAD/SIRAS	5 Br	1 *M* LiBr	30	1GQP	(37)
Transcriptional repressor	142	2.2/1.5	MIRAS	2 I/3 Br	0.35 NaI/ 0.5 NaBr	n/a	1IRQ	(38)
Salmonella SicP-SptP	730	2.5/1.9	MAD	31 Br	2 *M* NaBr	30	1JYO	(39)
β-Defensin-1	72	1.5/1.2	SAD	6 Br	0.5 *M* KBr	30	1IJU	(40)
IGF-1	70	1.8/1.8	MAD	1 Br	1 *M* NaBr	30	1IMX	(41)
Peroxiredoxin 5	161	1.7/1.5	MAD	5 Br	1 *M* NaBr	30	1HD2	(42)
TonB	152	2.0/1.55	MAD	4 Br	1 *M* KBr	50	1IHR	(43)
Frizzled CRD	762	2.0/1.9	MAD	9 Br	0.5 *M* NaBr	40	1IJX	(44)
Thiamine pyrophosphokinase	638	2.0/1.8	MAD	12 Br	1 *M* NaBr	45	1IG0	(45)
PSCP	375	1.8/1.0	SAD	9 Br	1 *M* NaBr	30	1GA6	(46)
IPP5C	347	2.0/2.0	SIRAS	5 I	0.15 *M* KI	n/a	1I9Y	(47)
Acyl protein pyrophosphokinase	464	1.8/1.5	SAD	22 Br	1 *M* NaBr	20	1FJ2	(48)
β-Defensin-2	164	2.0/1.35	MAD	9 Br/9I	0.25 *M* KBr/KI	60	1FD3	(49)
GAF domain YKG9	224	2.8/1.9	MAD	7 Br	0.5 *M* NaBr	45	1F9M	(50)
NEIL1	364	2.3/2.1	MIRAS	8 I	0.25 *M* NaI	300	1TDH	(51)

Only novel structures solved with soaked halides are included in the table. Both Br MAD/SAD and I SAD/SIRAS can be equally successful, in spite of the fact that the Br soaks seem to be more popular.

Both bromine and iodine are known as useful elements for heavy-atom derivatization of macromolecules. In particular, bromouracil is almost isostructural with thymine, and has often been used for multiwavelength anomalous diffraction (MAD) phasing of crystal structures of nucleotides *(5)*, in analogy to the role of selenomethionine in MAD phasing of protein crystals (*see* Chapter 5 of volume 1). Bromine has one electron more than selenium, and its anomalous scattering properties are similar with an X-ray absorption edge at 0.92 Å (selenium at 0.98 Å), easily achievable at most MAD-capable synchrotron beam lines. Iodine has for a long time been used as a classical heavy atom for multiple isomorphous replacement (MIR) experiments, e.g., after iodination of tyrosines with *N*-iodosuccinimide *(10)*. Its X-ray absorption edges are not easily accessible, but iodine with its 53 electrons displays the anomalous scattering effect of about seven electron units at the Cu Kα wavelength of 1.54 Å.

2. Materials

The only material necessary after crystals of the native protein have been obtained is the halide (bromide or iodide) salt of alkali metal, lithium, sodium, or potassium. A variation of the halide-soaking approach has been proposed *(11,12)* where cesium or rubidium halides were used, providing an additional anomalous signal from these two cations, also bound at the protein surface (*see* **Note 1**). Another proposed variation involves triiodides *(12,13)*, requiring elemental iodine dissolved in KI.

3. Methods

3.1. Soaking Procedure

The soaking procedure is very simple and consists of submerging the native crystal for a short time in the mother liquor containing simultaneously both a cryoprotectant and an appropriate halide salt, before freezing it for diffraction data collection.

The halide ions diffuse into protein crystals very quickly. Experiments with variable soaking times have shown that soaking times as short as 10–15 s are sufficient. Longer soaking times may degrade the crystal diffraction power, but sometimes led to crystalline phase transition (*see* **Note 2**), or actually extended the resolution limit of diffraction *(14)*.

The concentration of salt in the soaking solution plays a more important role than a prolonged soaking time. A high halide concentration increases the number of sites and their occupancy. However, not all crystals tolerate high concentrations (up to 1 *M*) of halide salts and their crystalline order may rapidly deteriorate. The optimal concentration can only be evaluated empirically. If the crystals visibly crack or shatter when observed under the microscope, or if the initial diffraction image shows signs of significant deterioration, the concentra-

tion of halide salt in the cryosolution should be diminished. The first trials can be performed with 1 *M* concentration, but successful phasing has been obtained with salts diluted to 0.2 *M*. The triiodide soaks require much more diluted solutions, in the order of 10–20 m*M* of KI_3 ***(12)***.

It may be advantageous to preserve the content and concentration of the initial mother liquor and add the measured amount of the halide in the form of a solid salt. If the crystallization liquor contains a high concentration of another salt, such as ammonium sulfate, it may be beneficial to substitute part of that salt with a halide, preserving the overall ionic strength of the solution and decreasing the adverse competitive effect of other anions.

Some salts at very high concentration may serve as cryoprotectants ***(15)*** and it may be worth checking whether the crystal can be successfully frozen if the concentration of halide salt is increased to greater than 2 *M*.

3.2. Phasing

In general, the structure solution of diffraction data from halide-soaked crystals may proceed along well-established protocols for utilization of the anomalous signal in MAD, single-wavelength anomalous diffraction (SAD), or MIRAS techniques, and any of the suitable programs can be used. However, in contrast to the popular MAD approach based on selenomethionines, the number of identifiable anomalous scatterer sites cannot be predicted. The halide-soaked crystals always have a large number of sites with variable occupancies, but only those sites with significant occupancy are useful for the evaluation of protein phases. In general, the number of useful sites is roughly proportional to the surface area of the protein molecule, but other factors (*see* **Note 3**) obviously play important roles.

The anomalous signal of bromides can only be utilized if diffraction data are collected at the synchrotron beam line, where the X-ray wavelength can be adjusted to the vicinity of the Kα absorption edge of bromine (0.92 Å). The fluorescence spectrum recorded from the crystal soaked in the bromide salt will always show the Br Kα absorption edge, resulting from the excess of bromides in the cryosolution, regardless of the number of bound bromide sites. The presence of bound bromides (or iodides) can be confirmed from the clear anomalous signal present in the diffraction data. The anomalous signal of iodine is more pronounced at longer wavelengths; its f″ contribution at the Cu Kα wavelength is about seven electron units, so that it can be used with data collected at the laboratory X-ray sources, not only at synchrotron facilities.

3.2.1. Location of Anomalous Scatterers Sites

As usual in all methods based on heavy or anomalously scattering atoms, the first step in crystal structure solution is the location of these atoms, i.e., halide

ions in the approach discussed here. This can be done with either Patterson or direct methods, utilizing the anomalous or isomorphous (dispersive) differences, as appropriate (*see also* Chapters 10–12). If only single-wavelength data are available, the anomalous (Bijvoet) differences have to be used. If in addition a native dataset exists, the classic SIRAS (single isomorphous replacement with anomalous scattering) method can be applied, and the halide sites can be identified from the anomalous or isomorphous differences. If MAD data are collected on bromide-soaked crystal, various combinations of anomalous and dispersive differences can be utilized, as well as the properly estimated diffraction contributions of anomalous atoms, F_A.

Because the number of expected halide sites cannot be predicted, they can be selected, e.g., by comparing peaks in several putative direct methods solutions. Equivalent sites (*see* **Note 4**) present in several independent direct methods trials can be accepted with confidence. Usually after the first round of phasing it is possible to identify more sites from the appropriate (anomalous or isomorphous) difference Fourier synthesis. In practice, weaker sites can be iteratively added as long as they increase the phasing power and enhance the interpretability of the resulting electron density maps.

3.2.2. Evaluation of Protein Phases

After a number of halide sites are identified, it is possible to evaluate the protein phases, perform phase improvement (solvent flattening and related density modification procedures), and calculate the initial electron density map. This step can be performed separately, or can be a part of the integrated software system, such as SOLVE ***(16)***, autoSHARP ***(17)***, CNS ***(18)***, SHELXD/E ***(19,20)***, or BnP ***(21)***. At the stage of protein phase evaluation, in particular after density modification, there should be a contrast in the phasing statistics, particularly in the map interpretability, between the two enantiomorphs.

Obviously, if the first automatic attempt is successful, and leads to an interpretable electron density map, there is no need for any further proceedings. If the initial map is not satisfactory for either enantiomer, the phasing procedure can be modified and repeated. The possible changes involve inclusion of more heavy atom sites, a different resolution limit, or changing the program used.

If the significance of measured anomalous differences does not extend to the full-resolution limit of diffraction data, it may be beneficial to perform the phasing procedure at a lower resolution limit, and subsequently extend the phased data at the density modification step ***(22)***. Various programs use different algorithms, some employ simpler but quicker approaches, some are more elaborate but slower. It is usual to start from a quick approach, and switch to more sophisticated program if the other attempts failed. However, the individual crystallographer's experience with a particular program is often important in obtaining final success.

4. Notes

1. In addition to the use of bromides and iodides, it has been proposed that heavier alkali metal cations, Rb *(11)* and Cs *(8,12)*, soaked into protein crystals can also be used for phasing. The procedure is analogous to the use of halides. The anomalous scattering properties of Rb are similar to those of Br (with the X-ray absorption edge at 0.87 Å) suitable for MAD phasing. The properties of Cs are analogous to I, with a substantial anomalous signal at a wavelength of 1.54 Å and longer.
2 It has been observed that soaking native crystals in concentrated solutions of salts sometimes causes crystal lattice transitions. The orthorhombic $P2_12_12_1$ crystals of PSCP *(23)* underwent transformation to the hexagonal space group $P6_122$ after short soaking in 1 *M* solution of NaBr. The crystals of human peroxiredoxin 5 *(24)* in a twinned monoclinic form after 30 s soak in 1 *M* NaBr changed to the tetragonal form, diffracting to higher resolution.
3. In general the larger the solvent accessible protein surface is, the more identifiable halide sites can be expected. However, the number of highly occupied sites may depend on several factors, such as the concentration of the halide (and other salts) and the chemical composition and pH of the solution, the isoelectric point of the protein, the presence of the positively charged residues as well as hydrophobic patches at the protein surface, and others. It is therefore not possible to predict how many halide sites can be expected in each individual case.
4. The comparisons of putative sites from separate direct methods solutions have to take into account the space group symmetry operations as well as the possible origin shifts and inversion of the enantiomer.

References

1. Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G., and Sheldrick, G. M. (1999) Can anomalous signal of sulfur become a tool for solving protein crystal structures? *J. Mol. Biol.* **289,** 83–92.
2. Loll, P. (2001) *De novo* structure determination of vancomycin aglycon using the anomalous scattering of chlorine. *Acta Cryst.* **D57,** 977–980.
3. Lehmann, C., Bunkoczi, G., Vertesy, L., and Sheldrick, G. M. (2002) Structures of glycopeptide antibiotics with peptides that model bacterial cell-wall precursors. *J. Mol. Biol.* **318,** 723–732.
4. Lehmann, C., Debreczeni, J. E., Bunkoczi, G., et al. (2003) Structures of fou crystal forms of decaplanin. *Helv. Chim. Acta* **86,** 1478–1487.
5. Hendrickson, W. A. and Ogata, C. M. (1997) Phase determination from multiwavelength anomalous diffraction measurements. *Methods Enzymol*. **276,** 494–523.
6. Dauter, Z., and Dauter, M. (1999) Anomalous signal of solvent bromides used for phasing of lysozyme. *J. Mol. Biol.* **289**, 93–101.
7. Dauter, Z., Dauter, M., and Rajashankar, K. R. (2000) Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. *Acta Cryst.* **D56**, 232–237.
8. Nagem, R. A. P., Dauter, Z., and Polikarpov, I. (2001) Protein crystal structure solution by fast incorporation of negatively and positively charged anomalous scatterers. *Acta Cryst.* **D57,** 996–1002.

9. Dauter, Z. and Dauter, M. (2001) Entering a new phase: using solvent halide ions in protein structure determination. *Structure* **9,** R21–R26.
10. Brzozowski, A.M., Derewenda, U., Derewenda, Z.S., et al. (1991) A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. *Nature* **351,** 491–494.
11. Korolev, S., Dementieva, I., Sanishvili, R, Minor, W., Otwinowski, Z., and Jachimiak, A. (2001) Using surface-bound rubidium ions for protein phasing. *Acta Cryst.* **D57,** 1008–1012.
12. Evans, G. and Bricogne, G. (2002) Triiodide derivatization and combinatorial counter-ion replacement: two methods for enhancing phasing signal using laboratory Cu Kα X-ray equipment. *Acta Cryst.* **D58,** 976–991.
13 Evans, G., Polentarutti, M., Djinovic-Carugo, K., and Bricogne, G. (2003). SAD phasing with triiodide, softer X-rays and some help from radiation damage. *Acta Cryst.* **D59,** 1429–1434.
14. Harel, M, Kasher, R., Nicolas, A., et al. (2001) The binding site of acetylcholine receptor as visualized in the X-ray structure of a complex between alpha-bungarotoxin and a mimotope peptide. *Neuron* **32,** 265–275.
15. Rubinson, K. A., Ladner, J. E., Tordova., M., and Gilliland, G. L. (2000) Cryosalts: suppression of ice formation in macromolecular crystallography. *Acta Cryst.* **D56,** 996–1001.
16. Tewilliger, T. C. (2002) Automated structure solution, density modification and model building. *Acta Cryst.* **D58,** 1937–1940.
17. Bricogne, G., Vonrhein, C., Flensburg, C., et al. (2003) Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Cryst.* **D59,** 2023–2030.
18. Brünger, A. T, Adams, P. D., Clore, G. M., et al. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Cryst.* **D54,** 905–921.
19. Schneider, T. R. and Sheldrick, G. M. (2002) Substructure solution with SHELXD. *Acta Cryst.* **D58,** 1772–1779.
20. Sheldrick, G. M. (2002) Macromolecular phasing with SHELXE. *Z. Krist.* **217,** 644–650.
21. Weeks, C. M., Blessing, R. H., Miller, R., et al. (2002) Towards automated protein structure determination: *BnP*, the *SnB*-PHASES interface. *Z. Krist.* **217,** 686–693.
22. Debreczeni, J. E., Bunkoczi, G., Ma, Q., Blaser, H., and Sheldrick, G. M. (2003) In-house measurement of the sulfur anomalous signal and its use for phasing. *Acta Cryst.* **D59,** 688–696.
23. Dauter, Z., Li, M., and Wlodawer, A. (2001) Practical experience with the use of halides for phasing macromolecular structures: a powerful tool for structural genomics. *Acta Cryst.* **D57,** 239–249.
24. Declercq, J. -P., and Evrard, C. (2001) A twinned monoclinic crystal form of human peroxiredoxin 5 with eight molecules in the asymmetric unit. *Acta Cryst.* **D56,** 1829–1835.

25 Usón, I., Schmidt, B., von Bülow, R., et al. (2003) Locating the anomalous scatterer substructures in halide and sulfur phasing. *Acta Cryst.* **D59,** 57–66.
26. Symersky, J., Li, S., Carson, M., and Luo, M. (2003) Structural genomics of *Caenorhabditis elegans*: triosephosphate isomerase. *Proteins* **51,** 484–486.
27. Im, Y. J., Park, S. H., Rho, S. -H., et al. (2003) Crystal structure of GRIP1 PDZ6-peptide complex reveals the structural basis for class II PDZ target recognition and PDZ domain-mediated multimerization. *J. Biol. Chem.* **278,** 8501–8507.
28. Namboodiri, V. M. H., Dutta, S., Akey, I. V., Head, J. F., and Akey, C. W. (2003) The crystal structure of *Drosophila* NLP_core provides insight into pentamer formation and histone binding. *Structure* **11,** 175–186.
29. Groppe, J., Greenwald, J., Wiater, E., et al. (2002) Structural basis of BMP signalling inhibition by the cystine knot protein noggin. *Nature* **420,** 636–642.
30. Lansdon, E. B., Segel, I. H., and Fisher, A. J. (2002) Ligand-induced changes in adenosine 5'-phosphosulfate kinase from *Penicillium chrysogenum*. *Biochemistry* **41,** 13,672–13,680.
31. Roosild, T. P., Miller, S., Booth, I. R., and Choe, S. (2002) A mechanism of regulating transmembrane potassium flux through a ligand-mediated conformational switch. *Cell* **109,** 781–791.
32. Sundaramoorthy, M., Meiyappan, M., Todd, P., and Hudson, B. G. (2002) Crystal structure of NC1 domains. Structural basis for type IV collagen assembly in basement membranes. *J. Biol. Chem.* **277,** 31,142–31,153.
33. Mittl, P. E., Fritz, G., Sargent, D. F., Richmond, T. J., Heizmann, C. W., and Grutter, M. G. (2002) Metal-free MIRAS phasing: structure of apo-S100A3. *Acta Cryst.* **D58,** 1255–1261.
34. Nagem, R. A. P., Colau, D., Dumoutier, L., Renauld, J. -C., Ogata, C., and Polikarpov, I. (2002) Crystal structure of recombinant human interleukin-22. *Structure* **10,** 1051–1062.
35. Erlandsen, H., Kim, J. Y., Patch, M. G., et al. (2002) Structural comparison of bacterial and human iron-dependent phenylalanine hydroxylases: similar fold, different stability and reaction rates. *J. Mol. Biol.* **320,** 645–661.
36. Than, M. E., Henrich, S., Huber, R., et al. (2002) The 1.9 Å crystal structure of the noncollagenous (NC1) domain of human placenta collagen IV shows stabilization via a novel type of covalent Met-Lys cross-link. *Proc. Natl. Acad. Sci. USA*, **99,** 6607–6612.
37. Au, S. W. N., Leng, X., Harper, J. W., and Barford, D. (2002) Implications for the ubiquitination reaction of the anaphase-promoting complex from the crystal structure of the Doc1/Apc10 subunit. *J. Mol. Biol.* **316,** 955–968.
38. Muruyama, K., Orth, P., de la Hoz, A., Alonso, J. C., and Saenger, W. (2001) Crystal structure of ω transcriptional repressor encoded by *Streptococcus pyogenes* plasmid pSM19035 at 1.5 Å resolution. *J. Mol. Biol.* **314,** 789–796.
39. Stebbins, E. C. and Galan, J. E. (2001) Maintenance of an unfolded polypeptide by a cognate chaperone in bacterial type III secretion. *Nature* **414,** 77–81.
40. Hoover, D. M., Chertov, O., and Lubkowski, J. (2001) The structure of human β-defensin-1. *J. Biol. Chem.* **276,** 39,021–39,026.

41. Vajdos, F. F., Ultsch, M., Schaffer, M. L., et al. (2001) Crystal structure of human insulin growth factor-1: detergent binding inhibits binding protein interactions. *Biochemistry* **40,** 11,022–11,029.
42. Declercq, J. -P., Evrard, C., Clippe, A., Stricht, D. V., Bernard, A., and Knoops, B. (2001) Crystal structure of human peroxiredoxin 5, a novel type of mammalian peroxiredoxin at 1.5 Å resolution. *J. Mol. Biol.* **311,** 751–759.
43. Chang, C., Mooser, A., Plückthun, A., and Wlodawer, A. (2001) Crystal structure of the dimeric C-terminal domain of TonB reveals a novel fold. *J. Biol. Chem.* **276,** 27,535–27,540.
44. Dann, C. E., Hsieh, J. -C., Rattner, A., Sharma, D., Nathans, J., and Leahy, D. J. (2001) Insights into Wnt binding and signalling from the structures of two Frizzled cysteine-rich domains. *Nature* **412,** 86–90.
45. Baker, L. -J., Dorocke, J. A., Harris, R. A., and Tim, D. E. (2001) The crystal structure of yeast thiamine pyrophosphatase. *Structure* **9,** 539–546.
46. Wlodawer, A., Li, M., Dauter, Z., et al. (2001) Carboxyl proteinase from *Pseudomonas* defines a novel family of subtilisin-like enzymes. *Nature Struct. Biol.* **8,** 442–446.
47. Tsujishita, Y., Guo, S., Stolz, L. E., York, J. D., and Hurley, J. H. (2001) Specificity determinants in phosphoinositide dephosphorylation: crystal structure of an archetypal inositol polyphosphate 5-phosphatase. *Cell* **105,** 379–389.
48. Devedjiev, Y., Dauter, Z., Kuznetsov, S. R., Jones, T. L. Z., and Derewenda, Z. S. (2000) Crystal structure of the human acyl protein thioesterase I from a single X-ray data set to 1.5 Å. *Structure* **8,** 1137–1146.
49. Hoover, D. M., Rajashankar, K. R., Blumenthal, R., et al. (2000) The structure of human β-defensin-2 shows evidence of higher order oligomerization. *J. Biol. Chem.* **42,** 32,911–32,918.
50. Ho, Y. -S. J., Burden, L. M., and Hurley, J. H. (2000) Structure of the GAF domain, a ubiquitous signalling motif and a new class of cyclic GMP receptor. *EMBO J.* **19,** 5288–5299.
51. Doublié, S., Bandaru, V., Bond J. P., and Wallace, S. W. (2004) The crystal structure of human endonuclease VIII-like 1 (NEIL1) reveals a zincless finger motif required for glycosylase activity. *PNAS* **101,** 10,284–10,289.

9

The Same But Different

Isomorphous Methods for Phasing and High-Throughput Ligand Screening

Mark A. Rould

Summary

Isomorphous difference methods allow rapid and detailed visualization of localized changes in macromolecular structures, whether as a result of mutation or the binding of ligands. Practical aspects of isomorphous methods and differential crystallography are presented, particularly in their application to the phasing of new structures by multiple isomorphous replacement and the detection and characterization of ligand binding. Techniques for maintaining isomorphism between crystals to maximize the differential signal are covered, as are the computational steps involved in generating difference electron density maps. Frontier applications such as determining single-site ligand-binding affinities crystallographically, high-throughput screening of combinatorial compound libraries, *in crystallo* competition assays, and inferring protein function via exogenous ligand-binding screens are discussed.

Key Words: Isomorphous methods; difference Fourier; MIR phasing; heavy-atom derivatives; protein–ligand interactions; high-throughput screen; combinatorial libraries; pharmaceutical design; binding affinity; competition assay; structure–function relationships.

1. Introduction

Some methods are so simple and so exquisitely powerful that it is assumed that apprentices of a discipline will acquire them as a matter of course. Over time, methods once considered obvious by their practioners are lost as new recruits swarm to the latest technologies. Such is the case with isomorphous difference methods in contemporary crystallography.

Aside from their familiar role in locating additional heavy atoms when phasing new structures, isomorphous difference methods are proving useful in a wide range of applications. The high signal-to-noise ratio achievable with isomorphous difference methods, coupled with the rapidity with which the experiments

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

can be executed and results obtained, makes them ideal for finding and characterizing ligands, either individually or in mixtures such as combinatorial compound libraries. Differential crystallography can be used not only to determine the conformation of a bound ligand and its mode of interaction with the target macromolecule, but less obviously to determine the thermodynamic affinity of the ligand for each target site. Competition assays can be carried out in the crystalline state, titrating one ligand's affinity for a particular site against another's. Visualizing the structural effects of mutations is another area where differential methods can be applied, to better understand catalytic mechanisms or the molecular basis of disease.

These and other applications of isomorphous methods will be described in this chapter, primarily from a practical perspective. Well established as well as frontier methods will be presented to encourage greater use of differential crystallography in the pursuit of structural science.

2. Materials

1. Heavy-atom reagents (Hampton Research [Aliso Viejo, CA], Strem Chemicals [Newburyport, MA]).
2. Nine-well (depression) plates (Hampton Research).
3. Cellophane packing tape (Hampton Research).

3. Methods

3.1. Preparation of Isomorphous Crystals

Whether screening for heavy-atom derivatives or comparing mutant and native proteins, the same overall considerations apply. The key, of course, is to maintain isomorphism, primarily by treating all crystals identically. The environment should be kept as constant as possible from crystal to crystal, both chemically and physically. This includes temperature, pH, and concentrations of all solvents and solutes, other than the derivatizing agent or other ligands intentionally added. With this in mind, there are two general approaches to performing isomorphous experiments: either working *in situ* with the crystal while it is in the mother liquor from which it grew, or finding a "surrogate" mother liquor in which the crystal is indefinitely stable and cryoprotected.

In situ approaches involve working directly with the hanging or sitting drop that yielded the crystal, and generally require less effort. They are well suited for working with a series of crystals of protein variants to which nothing further will be added. If the crystals require addition of a cryoprotectant, then long-term equilibration (on the order of hours) of the drop plus cryoprotectant is preferable to "quick-dips" in cryoprotectant, which can reduce reproducibility and isomorphism. Heavy- or anomalous-atom derivatives can also be pre-

pared in this way, but with an increased risk of loss of isomorphism compared with crystals transferred to a defined stabilizing and cryoprotecting solution bearing the additional compounds or agents (*see* **Note 1**).

Stabilizing/cryoprotectant solutions may take a little effort to find, but increase the routinely achievable level of isomorphism between crystals, particularly for high-throughput applications. When multiple crystals grow in each hanging or sitting drop, stabilizing solutions allow all of the crystals to be harvested and used for individual experiments. Stabilizing solutions are often made of the same components as the crystal's mother liquor (usually *sans* the protein) but with an increased concentration of precipitant to compensate for the absence of the protein, and ideally contain a cryoprotectant (*see* **Notes 2–4**). For stabilizing solution trials, nine-well plates work well. The wells of the nine-well plates can be sealed for up to several days with the same cellophane packing tape that is commonly used to seal high-throughput crystal screening trays.

The most important characteristic of a suitable stabilizing solution is that the crystals remain indefinitely stable in it; that is, that one can take a crystal after it has been in the solution for several days and collect quantitatively the same data from it as from a crystal in the solution for only a few hours (i.e., after equilibration). The diffraction limit of a properly stabilized crystal will meet or exceed that of crystals mounted directly from the mother liquor from which they grew. It is worth collecting a few complete native datasets from crystals that have been in the solution for varying amounts of time, to learn the time required for equilibration, and to get a baseline value for the degree of reproducibility that one can obtain. Whether crystals are stabilized or used *in situ*, a little effort to maximize reproducibility of the native or parent crystals will be rewarded with more accurate multiple isomorphous replacement (MIR) phases and cleaner isomorphous difference maps (*see* **Note 5**).

3.2. Heavy-Atom Derivatization for Phasing

Strategies abound for choosing among the "traditional" heavy-atom derivatizing agents *(1,2)*. Conditions that have yielded successful derivatization have also been compiled *(3–5)*. Although in the past it had been difficult (and expensive) to amass a collection of derivatizing agents, kits are now commercially available that contain small quantities of a broad sampling of historically useful compounds. Rather than recapitulating the strategies for ranking these compounds for particular cases, the following covers more recent cutting-edge techniques that expedite, and make more enjoyable, the search for isomorphous heavy-atom derivatives.

3.2.1. Fast Soaks at High Concentrations

Perhaps the most useful of recent revelations regarding heavy-atom derivatization is that the process is remarkably fast in many cases. Along these lines,

high on the list of derivatizing schemes should always be the short, concentrated halide soaks *(6)* (Chapter 8), as well as the related triiodide soaks *(7)*. The same quick-soak approach has been found to increase the probability of obtaining isomorphous derivatives even when using conventional heavy-atom reagents *(8)*. Short soak times, on the order of seconds to minutes rather than hours or days, lead to greater retention of isomorphism while yielding sufficiently high occupancies for phasing. Longer soak times apparently give the reagent more time to snuggle deeper into crevices between adjacent macromolecules and thereby disrupt the lattice. The same observation applies to addition of other ligands, such as substrates, analogs, or inhibitors.

3.2.2. Multiple Compounds Simultaneously Per Crystal

Most heavy-atom searches proceed by infusing one single heavy-atom compound at a time into native crystals. Presumably there is a fear that if more than one type of heavy atom is introduced at once that it will not be possible to determine which type of atom binds at each site. In fact, for nearly all MIR applications, it does not matter which heavy atom(s) bind to a given site, because to a very good approximation, all heavy atoms look the same after adjustment of occupancy and atomic B-factor. Even when the diffraction amplitudes have been placed on an absolute scale, most heavy-atom refinement programs allow the occupancy to take on any value (beyond the physically constrained range of 0 to 1). For example, the isomorphous (nonanomalous) scattering of a fully occupied lutetium ion can be well modeled as a cesium ion with an occupancy of 1.3. For MIR purposes, the identity of the bound heavy atoms is irrelevant. (This disregard for heavy-atom identity is not yet suitable for phasing by multiwavelength anomalous dispersion [MAD] however, in which the values of f′ and f″ for a given element play a key role in deriving phases, and mixtures greatly complicate the mathematics.) The practical impact is that several heavy-atom compounds can be soaked simultaneously into one crystal, as long as the compounds do not react with one another.

3.2.3. Rapid Detection of Useful Derivatives

Methods have been developed rather recently to prescreen heavy-atom reagents for their ability to bind a given macromolecule. Electrophoretic gels *(9)* and particle-induced X-ray emission *(10)* are among the techniques used to detect binding of heavy atoms, greatly expediting the screening process. The gel-based methods require no special instrumentation, or even crystals. A very small volume of protein solution is treated with a heavy-atom reagent, and changes in the protein's migration rate on native gels (relative to untreated protein) indicate a potential derivative. When crystals are difficult to obtain, or protein is scarce, this prescreening technique is invaluable (*see* **Note 6**).

When crystals are plentiful, perhaps the most direct way to determine whether or not one has a useful derivative is to collect the diffraction data, given the ease with which these data can be collected at the home source (*see* Chapter 5) and the speed and simplicity of the software for finding heavy-atom sites (*see* Chapter 11.) A simple derivatization strategy is to soak a crystal in a stabilizing/cryoprotectant solution containing the highest possible concentration of a mixture of heavy-atom compounds for a few minutes to a couple hours. Nine-well plates (sealed with tape for the duration of the soak) work well for this purpose and allow several soaks to be tried in parallel. Crystals surviving this treatment can be frozen away for data collection, otherwise the heavy-atom cocktail can be diluted with an equal volume of the stabilizing/cryoprotectant solution and the process repeated. Depending on crystal quality and space group, multiple potential derivatives can be screened per day at low-to-medium resolution, and those crystals whose heavy-atom sites can be located can be recollected at higher resolution.

3.2.4. Looking Outside the Box

Although the heavy atom itself directly ligates the macromolecule in most derivatizations, in some cases other portions of the compound are responsible for binding. Alkyl and aromatic metal reagents often bind in this manner. If one knows the natural ligands to a given macromolecule, heavy atom-substituted analogs are also potential derivatizing agents. Quite serendipitously, while searching for exogenous (nonnatural) ligands to various proteins using mixtures of iodine-containing organic compounds, our group discovered that members of these iodocompound cocktails bind to a few sites on just about every protein tested. The iodo group of the compound is the primary determinant for some of the binding events, but in others the organic portion anchors the molecule and the tethered iodine acts more as a beacon (owing both to its 53 electrons and relatively strong anomalous signal at Cu Kα wavelengths) indicating a bound ligand. **Figure 1** shows an example of these iodocompounds decorating the motor protein kinesin. Such cocktails of organo-iodide or organo-metallic compounds complement the fast halide soaks, and may find general utility both for phasing and for identifying empirical "hot spots" of ligand binding on macromolecular targets.

3.3. Processing Isomorphous Data

Whether one is using isomorphous methods to phase a new structure, or to study mutants or ligand-binding properties of an already-solved protein, the initial processing is very similar. One of the strengths of isomorphous methods is the speed with which they deliver useful answers. Within minutes of collecting and reducing isomorphous datasets, one can know whether one has a new derivative,

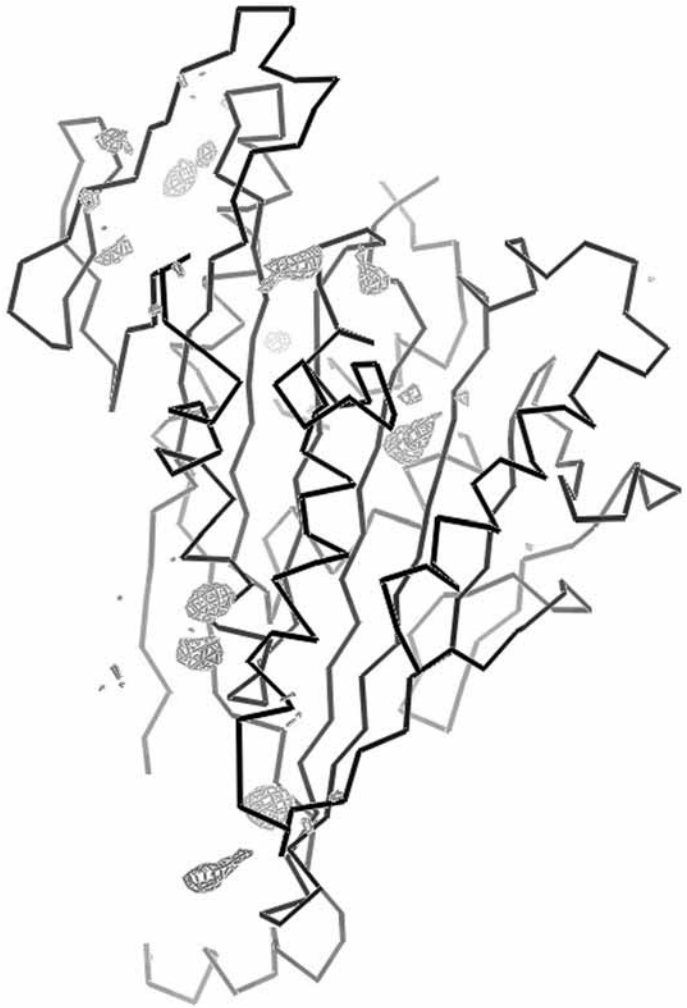


Fig. 1. Contrast enhancing and phasing agents: constellation of a mixture of iodine-containing compounds bound to kinesin. Several iodinated compounds were infused simultaneously into a kinesin crystal for several minutes and visualized via an isomorphous difference Fourier map. Iodines (or other heavy atoms) as substituents of compound libraries enhance detection of weak ligand binding, and afford an additional source of reagents for isomorphous phasing.

or, if phases are available, view an isomorphous difference Fourier map that indicates ligand binding and any significant changes in structure. The theory underlying isomorphous methods is given elsewhere *(11)*; we will restrict ourselves to practical aspects here.

3.3.1. Computational Methods in Differential Crystallography

In overview, to generate an isomorphous difference Fourier map, one applies the Fourier transform to the (signed) difference between the observed structure factor amplitudes of the two isomorphous crystals to be compared, with phases usually taken from the back-transformation of a refined isomorphous model:

Isomorphous difference map = FT[(Fobs,variant – Fobs,parent) * exp(i * Phase)] (1)

where Fobs,parent are the observed amplitudes from the native, wild-type, or reference crystal; Fobs,variant are the observed amplitudes from isomorphous crystals of the ligand-bound or mutated form of the macromolecule, scaled to the Fobs,parent; and the phases here are expressed in radians. All of the necessary steps are carried out by routines within crystallographic suites such as CCP4 ***(12)*** or CNS ***(13)***. Regardless of the suite chosen, the steps carried out are essentially the same:

1. Convert the diffraction datasets to the format used by the crystallographic suite.
2. Bring the two sets of amplitudes to the same scale, usually by applying a linear and resolution-dependent scale to the Fobs,variant amplitudes, or by applying local scaling ***(14,15)***. In the process, generate statistics (e.g., R_{iso}, or "cross-R" values) for the comparison of the two datasets:

$$R_{iso} = \frac{\text{sum}(|\text{Fobs,variant} - \text{Fobs,parent}|)}{0.5*\text{sum}(\text{Fobs,variant} + \text{Fobs,parent})} \quad (2)$$

where the sums are over all H K L indices for which both variant and parent amplitudes have been measured. Sometimes this value is reported using intensities rather than amplitudes (i.e., R_{iso} on I rather than R_{iso} on F).

3. If the overall differences appear excessively large (cross-R > 50%) and if alternate indexing is a possibility for this space group, apply the appropriate transformation to the Miller indices to make the indexing consistent with the other datasets (i.e., reindex) and repeat the scaling process. The REINDEX routine in CCP4 can be used for this. The reindexing matrix is a (hand-preserving) point-group symmetry operator of the reciprocal lattice that is not present in the Laue symmetry of the space group. An example of how this matrix is determined is given in **Note 7**.
4. Pair corresponding amplitudes from the two datasets. From the equation above for the difference map it is clearly important that only paired reflections (for which both Fobs,variant and Fobs,parent are measured) be used.
5. Discard amplitude differences that are statistically larger than reasonably expected. These are usually owing to an error in measurement of one or the other amplitude. A reflection might have been partially occluded or coincident with a spurious reflection (from ice for example.) A checkerboard-like ripple running throughout a difference map usually indicates the presence of a large difference that should be rejected. Threshold values for rejection of large differences are set as parameters within most program suites.

Table 1
Sequence of Steps for Generating Isomorphous Difference Fourier Maps via CNS or CCP4[a]

	CNS	CCP4i
Convert/reformat data	to_cns	Data reduction, import merged data
Merge datasets	merge.inp	Experimental phasing, CAD
Scale datasets	scale.inp	Experimental phasing, Scaleit (fhscale)
Difference statistics	analyze.inp	(generated by scaleit)
Back-transform model to give phases	model_phase.inp	Reflection data utilities, calculate Fs and phases
Make isomorphous difference map	fourier_map.inp	Map and mask utilities, run FFT, create map

[a]The CNS routines are found in the "inputs/" subdirectory of a standard installation. Major and minor menu headings are given for the CCP4i graphical user interface.

6. Generate phases by taking the inverse Fourier transform (i.e., back-transform) of an isomorphous model. In most cases the model (PDB file) to be used for generating phases is the refined model for the parent structure, although this need not be the case. A model that has been refined against a diffraction dataset that is isomorphous with the variant and parent datasets (R_{iso} less than about 20%) will also work. Alternatively, to completely eliminate any possible model bias, experimentally derived phases (e.g., MIR or MAD phases, with or without density modification restraints applied) can be used.
7. Apply the Fourier transform to the signed differences between paired Fobs,variant and Fobs, parent with phases from the previous step to give the isomorphous difference Fourier map.

Table 1 lists the steps for carrying out this procedure for making isomorphous difference Fourier maps via CCP4 ***(12)*** or CNS ***(13)***.

When making the map, choice of resolution limits can have an impact, particularly when the two structures are not entirely isomorphous. A simple empirical high-resolution limit to use is based on inspection of the average absolute isomorphous differences (i.e., the average numerator in the R_{iso} equation previously mentioned, not the R_{iso} value itself) as a function of resolution. The clearest difference map usually results when the high-resolution limit is chosen as the point at which the isomorphous differences begin to increase after reaching their lowest values.

Detection of binding of heavy atoms as a prelude to solving a new structure by isomorphous replacement follows the same steps as above up to the point of

taking the Fourier transform. At this point a difference Patterson map can be generated by taking the Fourier transform of the squared differences in amplitude with all phases set to zero, or an automated heavy-atom search program can be invoked. Several suites are available for rapidly solving and/or refining the heavy-atom substructure of a potential heavy-atom derivative, as discussed in Chapters 10–12.

3.3.2. Interpretating Isomorphous Difference Fourier Maps

The isomorphous difference Fourier map is interpreted analogously to the residual difference Fourier map, or "Fo-Fc" map, with which most crystallographers are familiar. Whereas the residual map shows errors or missing or excess atoms in a model, the isomorphous difference map compares electron density between two crystals, in a manner largely free of any model bias ***(11)***. The isomorphous difference map is currently as close as one can get to comparing two *structures*, rather than comparing two *models* that represent two structures. Given the typical 20–25% crystallographic residual of most models, and the greater than 0.2-Å root mean square coordinate error intrinsic to refined models, this difference map is generally a more accurate indicator of the true differences between structures.

If the isomorphous difference map is calculated as in **Eq. 1**, positive density, or peaks, occur at locations where the electron density in the variant crystal is greater than that in the parent crystal; negative density, or holes, result when the opposite is true. Atoms that shift more than their diameter give rise to peaks in the difference map where the atom is located in the variant, and holes where it is located in the parent. Just as with residual maps, the peak-hole pairs in difference density resulting from small shifts in an atom or side chain of the structure require more thought to interpret. For shifts less than the diameter of an atom, the distance between the extrema in the difference density map exaggerates the actual shift of the atom; that is, the atom in the parent structure is neither located at the minimum of the hole, nor is the atom in the variant structure located at the maximum of the peak. Calculation of the actual shift requires a more detailed analysis ***(11,16)***. Changes in occupancy of a moiety are revealed by a peak or hole centered on the moiety, whereas changes in B-factor similarly show a peak or hole but surrounded by a concentric spherical shell of opposite difference density. When the moiety is a heavy atom, the latter effect is difficult to discern from truncation errors in the Fourier transform. With practice, interpreting an isomorphous difference Fourier map becomes intuitive, and one soon finds that this map in itself reveals all the significant differences between the two structures. Even so, it is generally necessary to refine a model for each state, if only for the purpose of publication.

3.3.3. Refinement and Validation of Isomorphous Structures

Refinement of a new model often starts from a previously refined model of an isomorphous structure. For example, refinement of a mutant protein model might begin from a refined model of the wild-type. This prior information that the structures are related can be used to assist refinement of subsequent models, and at the same time imposes an important constraint on the choice of reflections that can be used for cross-validation of those subsequent models. We will begin by looking at one way that isomorphism can help the refinement.

The final R-factor for a model after all rebuilding and refinement is complete is typically around 20–25%. A significant portion of this residual error arises from the inadequacy of the model to represent the true structure (or ensemble of structures) in the crystal, rather than from measurement error. If we reasonably assume that this discrepancy between model and true structure is approximately the same for models of isomorphous structures, then we ought to be able to subtract this portion of the residual error in the refinement of subsequent models of isomorphous structures. This procedure, called Bayesian difference refinement ***(17)***, is implemented within the SOLVE crystallographic suite ***(18)*** and can be used in conjunction with any refinement program. One provides SOLVE with the final refined parent model in Brookhaven Protein Data Bank (PDB) format and the diffraction datasets for the parent crystal and the isomorphous variant crystal to be refined. The program returns a file with modified Fobs and σ values, which are to be used for refinement of the model of the variant structure starting from the model of the parent structure. The more isomorphous the structures, the greater the effect this procedure has on the acceleration of convergence of refinement, and the more accurately the difference between the refined models represents the difference between their structures.

Refinement of models for isomorphous structures requires careful consideration of the choice of reflections to be used for cross-validation. When refining a model starting from a previously refined model of an isomorphous structure, it is important that both refinements use the same set of reflections for cross-validation; that is, the "free-R" or "test" ***(19)*** reflections that are omitted from all steps of refinement should be the same for refinement of all subsequent models of isomorphous structures. The rationale for this requirement proceeds as follows. A reflection (called a "working" reflection) used in refinement introduces some degree of bias into the model being refined, which results in the discrepancy between the observed amplitude of that reflection and the amplitude calculated from the model being smaller than it really should be. The model, after refinement against the "working" set of reflections, is thus "overbuilt" or "overmodeled" with respect to those reflections. The "test" set of reflections, whose members are never used in refinement, does not introduce such bias into the model and is used to validate the model. The discrepancy between observed and

calculated amplitudes of these "test" reflections more correctly reflects the true error in the model. Because the observed diffraction amplitudes between isomorphous structures are very similar, then choosing a reflection that was once used as a "working" reflection to serve as a "test" reflection in a refinement starting from a previously refined isomorphous model has already biased the model toward that reflection, and thus cannot truly serve as a "test" reflection.

Failure to use the same set of reflections for cross-validation will give an initial free R-factor that is unusually low, but that tends to decrease at a slower rate and often stalls at a higher final value than for previous models. This is because the "test" reflections are used by most contemporary refinement programs to estimate the true discrepancy between the current model and the correct structure, and to adjust global parameters accordingly. By "fibbing" to these refinement programs by choosing a new "test" set that includes previous "working" set reflections, the inherently lower overall free R-value incorrectly suggests that the current model is closer to the correct model than it should be, and, hence, the parameter shifts to be applied are miscalculated. Often one can atone for this transgression by using molecular dynamics (e.g., simulated annealing) at a high enough temperature to raise the free and working R-values to peak greater than 40%, after which the resultant model has largely lost its bias toward the new set of "test" reflections. Of course, if one always starts refinement from a nonisomorphous model, then one is free to choose new "test" and "working" sets of reflections, although the refinement will generally take longer.

Comparing models for two isomorphous structures, one might wonder whether small structural differences between them are real. There are at least two simple ways to address this important question. The most direct is to inspect the models overlaid on the isomorphous difference Fourier map itself. The differences between models are more likely to be correct if they are consistent with features in the difference map, such as peak-hole pairs. The second test involves swapping the models and the data against which they have been refined, and refining again. For example, if model A has been refined to convergence against dataset A, and model B has been refined to convergence against dataset B, then to validate the differences between the models refine model A against dataset B to give a new model A′, and likewise refine model B against dataset A to give a model B′. After the two refinements have converged, not only should model A′ closely resemble model B and model B′ closely resemble model A, but more importantly, the shift vector for any individual atom in going from its position in model A to its position in model B should be the same as that atom's shift vector in going from its position in model B′ to its position in model A′. The agreement for the shift in this one atom can be expressed mathematically as the dot product of these two vectors AB and B′A′, and the dot products averaged over all the atoms in the region of interest give a quantitative assessment of the validity of the observed shift.

Programs to assist in calculating these validation values are available at the crystal.uvm.edu website.

3.4. Applications of Isomorphous Methods

In this last section, some of the ways that isomorphous methods and differential crystallography can be applied to structural science are discussed, from the more tried-and-true applications to frontier methods.

3.4.1. Phasing of New (and Not so New) Structures

Although most macromolecular structures are currently solved using multiwavelength anomalous dispersion or molecular replacement methods, there are still many cases in which isomorphous replacement phasing is worth pursuing. MAD has the potential to be the fastest route to a macromolecular structure, if anomalously scattering atoms can be inserted into the protein or nucleic acid under investigation. For those proteins which are extracted in large quantities from natural sources, preparation of selenomethionine-substituted samples can be problematic: rabbits, fowl, and cattle do not take well to the inclusion of large amounts of selenomethionine in their diet. Even for readily overexpressed Se-Met-substituted cloned proteins (*see* Chapter 5 of volume 1) availability of synchrotron beamtime can be rate-limiting. In some cases, structures have been determined by MIR while waiting for beamtime scheduled weeks or months in the future. MIR enables one to take advantage of the home source rotating anode X-ray generator when other resources are less available.

When suitable search models exist, molecular replacement allows structures to be solved with the least amount of experimental data: amplitudes from a single crystal, without any experimental phase information (*see* Chapter 7). The price paid for this simplicity in data collection is an intrinsic bias toward the search model. Cumulatively the vast arsenal of programs available nearly assures a solution to the rotation and translation functions when a model for a similar structure is available; however, care is required to minimize the bias introduced by using phases derived from the search model to (re)build the new model. Problems with molecular replacement also arise when the known portion of the new crystal is too small a fraction of the asymmetric unit to phase the unknown component. For this case, and in the general case to reduce the intrinsic model bias, perhaps the best use of a molecular replacement solution is as a source of crude phases to identify heavy-atom sites in isomorphous derivatives via difference Fourier maps. Acquiring those isomorphous derivatives, and the bias-free electron density maps they give rise to, can be rapid and straightforward, and well worth the effort when it comes time to rebuild and refine the model.

3.4.2. Structural Analysis of Mutations

Most physiologically relevant proteins have a long list of known mutations, both naturally occurring and site directed, often with functional ramifications, and all-too-often associated with disease. Visualizing the structural effects of the mutation can be rather simple if a few tenths of a milligram of the mutant or variant protein are available, because mutants or variants often crystallize under the same or similar conditions as the parent or native protein. If protein supply is limited, or to expedite crystallization, microseeding with existing parent crystals may nucleate growth of isomorphous crystals (Chapter 7 of volume 1). This incurs a risk of biasing the variant's global conformation toward the parent's, a possibility that can be addressed by screening the variant for alternative crystallization conditions.

3.4.3. Characterization of Ligand Binding

The function of a macromolecule is defined by the ligands, large and small, with which it interacts. These ligands include substrates, products, inhibitors, allosteric regulators, or other proteins or nucleic acids. Understanding these ligands, their binding sites, conformation, and orientation relative to other ligands, as well as the structural changes in the target macromolecule their binding may induce, is accelerated when they can be directly visualized interacting with the macromolecule. Particularly for small ligands (or the relevant portion of larger ligands) that can be soaked or infused into pre-existing crystals, isomorphous methods can expedite the characterization of a large proportion of these macromolecule–ligand interactions. This process requires a source of phases, which often can be obtained by back transforming a refined model of an isomorphous structure. Within minutes of collecting and reducing the diffraction data, one can generate an isomorphous difference Fourier map, which clearly reveals whether or not anything has bound as well as provides details of the interaction, including any structural rearrangements of the target macromolecule. **Figure 2** gives an example of the quality of structural information that is "instantly" available, before any refinement of a model for the complex begins.

Perhaps as important as indicating the configuration of the protein-bound ligand, the isomorphous difference map shows in detail how the protein adjusts to accommodate the exogenous ligand. For example, in **Fig. 2** the isomorphous difference map reveals a tyrosine side chain to the right of the cytochalasin molecule that swings around to better accommodate the ligand. By providing a model of the protein as it is configured to bind ligands, even a single crystal structure with an exogenous ligand bound in the targeted site greatly improves the probability of successful design of an inhibitor by computational modeling ***(20)***.

Although it is always possible when soaking compounds into pre-existing crystals that crystal-packing forces may prevent the full expression of a conformational change induced by ligand binding (or relatedly, that these forces may

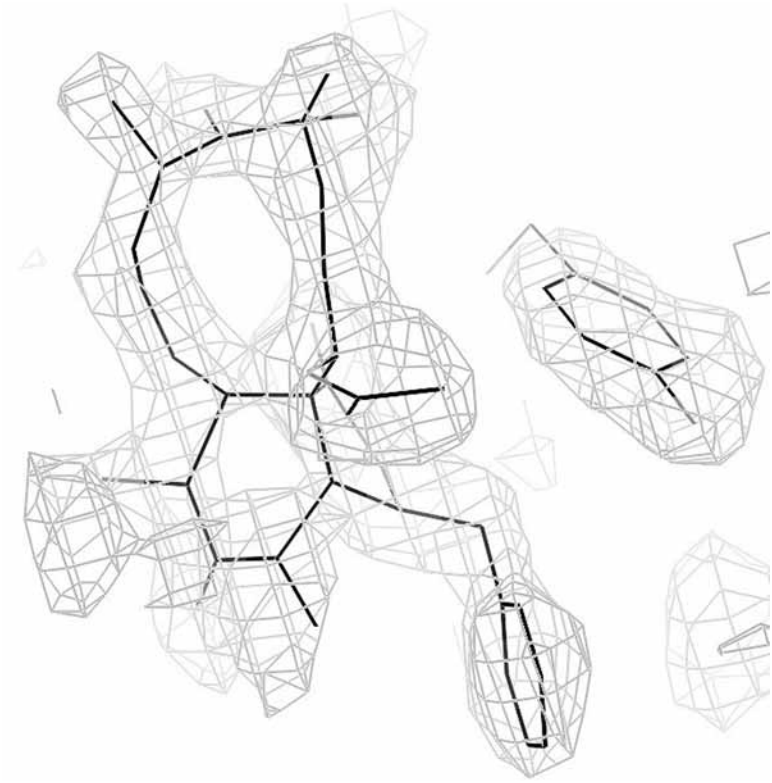


Fig. 2. Rapid characterization of ligand binding: isomorphous difference Fourier map of cytochalasin D bound to actin. Within minutes of data collection (to a resolution of 2.5 Å) this image was generated, free of any model bias, revealing the rings and substituents of this fungal metabolite bound to its target protein. It is worth noting that to maximize isomorphism the reference crystal used in the making of this isomorphous difference map contained the carrier solvent (ethanol) at the same concentration as was used for the cytochalasin D soak.

prevent possible reconfiguration of the ligand-binding pocket), one can always attempt to crystallize the macromolecule and ligand *de novo*. Except when the ligand is located at a crystal-packing interface, in which case the details of its interaction with the macromolecule are suspect anyway, cocrystals often can be grown under similar conditions as the parent (*see* **Note 8**).

From a random sampling of the literature regarding crystal structures of ligand-bound complexes and the corresponding data deposited in the Protein Data Bank *(21)*, it is clear that isomorphism between unliganded crystals of macromolecules and their ligand-bound forms is far more prevalent than most investigators realize. Retrospective application of the methods described in this

chapter will most likely lead to a re-interpretation of results and conclusions in more than a few cases.

3.4.4. Crystallographic Determination of Site-Specific Affinity Constants

In many respects, the macromolecular crystal is the ideal specimen for experimental thermodynamics. Each crystal contains an ensemble of about 10^{12} molecules in identical, well-defined environments, and the average state of that ensemble can be directly visualized crystallographically. Given the high solvent content of macromolecular crystals, and the extensive network of solvent channels running throughout them, the interior solute composition of the crystal quickly equilibrates with the exterior solution. Taken together, these features allow ligand-binding affinities to be determined empirically for any *individual binding site* on the macromolecule, in contrast to bulk-phase affinities which may involve multiple binding sites per macromolecule.

For cases in which the amount of ligand present is much greater than the amount of protein, the dissociation constant for a single site reduces to the simple equation:

$$(1/\text{occupancy}) = K_D * (1/[\text{Ligand}]) + 1 \qquad (3)$$

where the occupancy is the fraction of the asymmetric units with the ligand bound in the site of interest. Measurement of occupancy at various ligand concentrations should give a constant value for the K_D. Accurate determinations of the occupancy are thus central to this method. When high-resolution diffraction data are available, occupancy of the ligand can be determined by its direct refinement with standard crystallographic refinement programs; however, there is a strong coupling between B-factor and occupancy that will limit the accuracy of this approach.

Alternatively, the occupancy can be estimated by integrating the electron density for the ligand. The simplest incarnation of this method has been used to estimate the affinities of various cations for sites within the KcsA ion channel *(22)* (*see* **Note 9**). Crystals can be grown from solutions of various ligand compositions, or pregrown crystals can be transferred to stabilizing/cryoprotectant solutions of defined ligand composition with greater chance of retaining isomorphism. Displaced solvent molecules and the effects of symmetry at the ligand-binding site must also be taken into account in a more rigorous analysis, as well as activity coefficients and entropic effects as a result of partitioning of the ligand in the crystal and solution. There are the obvious limitations to this method: the binding sites must not be occluded by crystal packing, and if any gross conformational change in the macromolecule accompanies ligand binding, that too must be accommodated by the lattice. Even so, crystallographic thermodynamics is a promising area of investigation, with

obvious practical applications to the fields of medicinal chemistry and molecular toxicology.

3.4.5. To Find a Needle in a Haystack, Subtract the Haystack: Crystallographic High-Throughput Screens for Ligand Binding

From **Eq. 3**, any ligand present at a concentration equal to its dissociation constant for a binding site on the target macromolecule will be seen to occupy that site half of the time. If the target is challenged with multiple compounds simultaneously, those compounds that compete for the same site will bind in proportion to their affinity, and those that do not interact with either the site or other compounds will have little or no effect on the binding of the others. This is the principle underlying high-throughput screens for ligand binding by crystallography, with obvious applications to pharmaceutical lead discovery and toxicological screening. Crystals are soaked in a mixture of up to several thousand compounds simultaneously, and isomorphous difference methods are employed to reveal whether any compounds have bound. The resulting difference density represents the ensemble of bound ligands, each weighted by its relative affinity, less any displaced ordered solvent molecules. By analogy with phage display methods, we refer to this method as "crystal display." The information thus obtained can be used in at least two ways. One could identify individual compounds by iteratively dividing the mixture and using the crystallographic screen as a simple assay for the presence of compounds that bind. With somewhat more sophistication, one could use the difference electron density of the ensemble of bound compounds as a template (or envelope) for the design of a subsequent round of potential ligands, because the ensemble average density indicates where the target macromolecule has a predisposition to be bound. The ligand design problem is thus reduced to a computational search for compounds that individually fill the observed difference density envelope while maximizing interaction complementarily (hydrophobicity, hydrogen bonding, and electrostatic potential) with the target protein.

Combining "crystal display" with combinatorial organic chemistry overcomes the primary disadvantages of using combinatorial compound libraries in high-throughput screens. Other assay methods require that the compounds be either uniquely tagged or individually accessible in order to identify the active members. Via differential crystallography, one can either use the electron density of the ensemble to guide design of subsequent combinatorial libraries as previously described, or individual compounds can be identified iteratively by resynthesis *(23)*. For example, positive isomorphous difference density is shown in **Fig. 3** for member(s) of a 400-compound combinatorially synthesized library bound in the substrate cleft of the benchmark protein lysozyme.

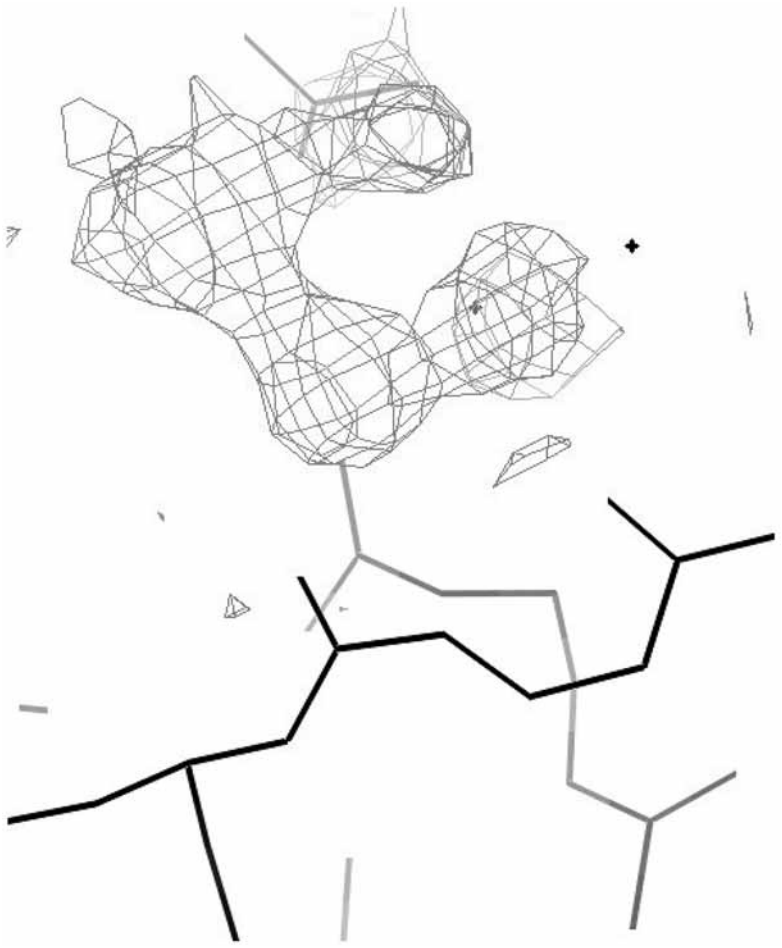


Fig. 3. Crystallographic screening of combinatorial compound libraries. Positive isomorphous difference density reveals the binding of member(s) of a 400-compound combinatorial library (without contrast-enhancing atoms) in the active site of lysozyme.

To the extent that heavy atoms (e.g., iodine or metals) can be incorporated in the combinatorial compounds, this greatly enhances the detection limit of bound compounds. The heavy atoms act much like contrast agents or atomic beacons, thus lowering the occupancy required to detect binding and allowing lower affinity compounds to be screened. The anomalous signal of most heavy atoms provides an additional route to verify the binding of exogenous ligands, via the anomalous difference Fourier map (*see* **Note 10**).

3.4.6. Function From Structure: Liganomics

Applications of high-throughput ligand-binding screens are not limited to the pharmaceutical industry. The obvious next stage of the structural genomics initiative is to use structures to determine function. Unfortunately, it is well established that structure does not uniquely identify function: the same basic protein scaffold can support multiple—sometimes dozens—of quite different functions or reactions. However, if one could determine the constellation of ligands that the protein binds, that is, which ligands bind where on the protein, in what orientation and conformation, then one would be in a much better position to ascribe a function to that protein. Although one could envision a myriad of methods for screening potential ligands, a high-throughput crystallographic screen is one of the few capable of rapidly providing answers to all of the previously listed questions. Ideally the true ligands, the substrates, products, transition state analogs, and allosteric regulators, would be among the ligands in the vast array of candidates to be screened. This would not likely be feasible, however, because there are far too many natural ligands to test exhaustively. A representative set of probe compounds must be used instead.

Fragments of a larger ligand sometimes bind in the same or similar location on the target protein as they do when the complete ligand binds *(24,25)*. If a database can be established that relates the true ligands of proteins of *known* function to the constellation of exogenous ligands that the protein is seen to bind via crystallographic screens involving a very large number of diverse probe compounds, then one might be able to infer the true ligands of a crystallized protein of *unknown* function by determining its ligand-binding constellation from among the same large set of probe compounds. This simple notion embodies the concept of structural *liganomics* for determining the function of a protein from the ligands that it is literally *seen* to bind. It is worth noting that this academic application of "crystal display" dovetails quite well with the pharmaceutical and toxicological applications. All parties would benefit from the sharing of their databases of ligand-binding constellations and the diffraction data from which those constellations were derived.

4. Notes

1. When adding a cryoprotectant, with or without extra compounds or derivatizing agents, to a hanging drop *in situ*, osmotic shock to the crystal can be reduced by placing the cryoprotectant solution next to (but not yet joined with) the drop bearing the crystal on the underside of the cover slip. After overnight equilibration by vapor diffusion, the cryoprotectant-bearing drop is dragged to the crystal drop, the cover slip is placed back over the reservoir, and the crystal is allowed to further equilibrate in the merged solution for a few hours before freezing.

2. A simple stabilizing/cryoprotectant solution that may find more general applicability particularly for high-salt precipitants is made by equilibrating concentrated polyethylene glycol 400 solutions (e.g., 50% [v/v]) with a saturated solution of the precipitant, including pH buffer and any other components of the crystallization mixture. After vigorous vortexing, the suspension is allowed to settle overnight and the (usually upper) polyethylene glycol 400 layer is extracted for use as a stabilizing/cryoprotectant solution. Recently, concentrated solutions of sodium malonate (and other organic salts such as tartrate) have been used with frequent success for cryoprotection and stabilization of crystals grown from other salts ***(26,27)***.
3. In the worst case, one can always make a stabilizing solution from the mother liquor from which the crystals are grown. This can be an expensive alternative, because it requires a high concentration (a nearly saturated solution) of the protein itself. For some projects, however, this may not be too high a cost, because only a small volume is needed either for harvesting multiple crystals from a single drop, or for preparing homogeneous heavy atom or ligand soaks (infusions). These "surrogate drops" of stabilizing solution should be the same as the final *equilibrated* mother liquor (with a slightly lower concentration of protein or precipitant to prevent growth of satellite crystals), rather than the initial concentrations existing when the crystallization is set up. Drops from the same conditions but that failed to yield crystals are one source of this mother liquor. Cryoprotectant can be blended in to give a homogeneous cryoprotecting and stabilizing solution. An added benefit of using artificial mother liquor as a stabilizing solution for heavy-atom derivatization is that the high concentration of protein buffers the stabilized crystal from the sudden addition of the derivatizing agent, the same as for *in situ* derivatization of crystals in the drop from which they are grown. For the same reason, more derivatizing agent may be required.
4. A variant of the above approach is to use a heterologous protein in place of the crystallized protein in the stabilizing solution. We have successfully used bovine serum albumin and lysozyme in the stabilizing solutions of more precious protein crystals.
5. Whether using *in situ* or harvested crystals, once a cryoprotectant is found it is worth trying to grow subsequent crystals in the presence of a low concentration of the cryoprotectant. Cryoprotectant molecules are frequently seen to bind to the protein, and may perturb the lattice slightly, increasing mosaicity. Growth with a little cryoprotectant allows these sites to be accommodated in the growing lattice, thus reducing the shock when more is added later.
6. For high-throughput phasing of proteins for which selenomethionine substitution is not possible, one could envision carrying out the gel-based heavy atom screen before crystallization trials. Subsequently one would screen for crystallization conditions of the protein pretreated with the derivatizing agent.
7. For example, for space groups P321 and P312, the hand-preserving symmetry of the reciprocal lattice is 622. One looks to space group P622 to generate all the hand-preserving point-group symmetry operators. The list of point-group symmetry operators in reciprocal space can be generated as the transpose of the real space

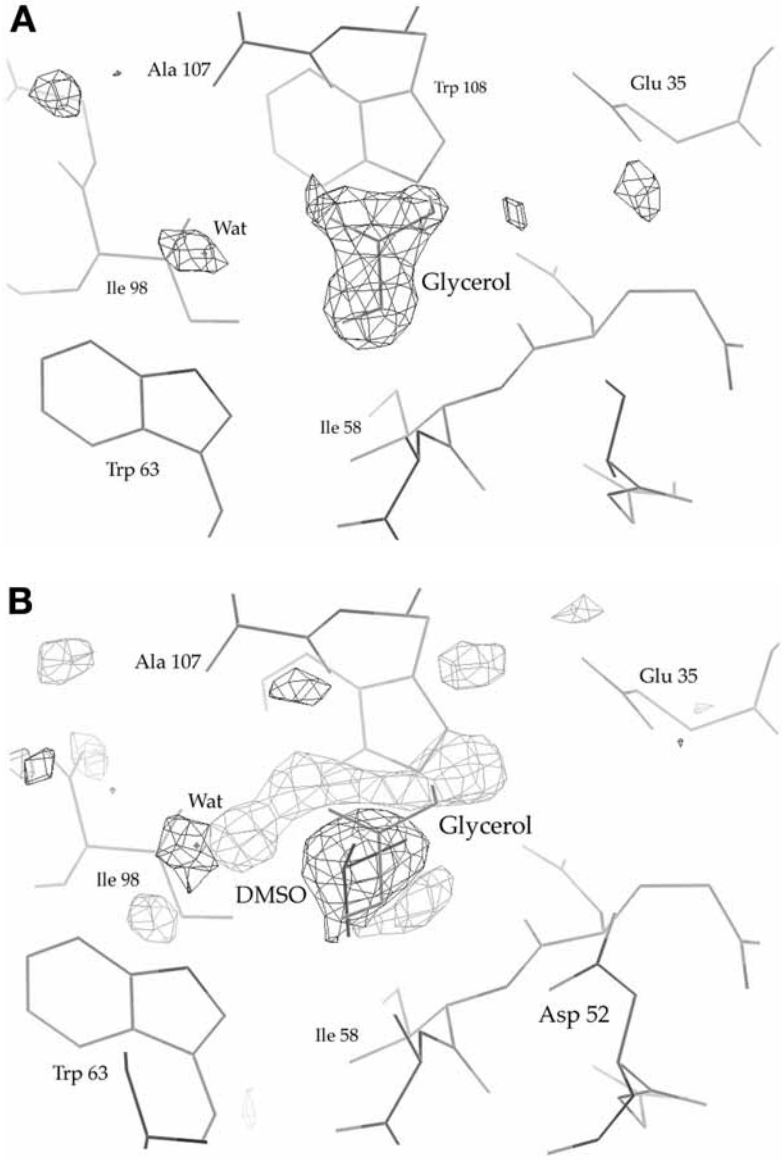
Fig. 4. *In crystallo* competition assays. **(A)** Residual difference Fourier map indicates a glycerol (cryoprotectant) molecule bound in the substrate-binding cleft of lysozyme. **(B)** After allowing a crystal to equilibrate in a stabilizing/cryoprotectant solution containing defined concentrations of glycerol and dimethyl sulfoxide (DMSO), an isomorphous difference Fourier map between this crystal and the glycerol-only reference shows the DMSO (positive isomorphous difference density, dark contours) displaces the glycerol (negative isomorphous difference density, light contours) and shifts the position of an adjacent water molecule. Crystallographic titration allows the relative affinities of two ligands for a single site to be estimated. (Molecular graphics in this chapter were made with vuSette zc [M. Rould].)

symmetry operators for the space group, dropping the translational components. For space group P321, the lattice transformation that takes H to –K, K to –H, and L to –L (i.e., 0 –1 0/–1 0 0/0 0 –1) is not among the Laue symmetries of this space group, and is thus the matrix that relates alternate indexing schemes. For space group P312, the transform that takes H to K, K to H, and L to –L (i.e., 0 1 0/1 0 0/0 0 –1) is the matrix. One should check that the determinant of the re-indexing matrix is positive, else the sense of Friedel mates will be inverted.

8. If crystals only can be grown with a ligand present, then one is often able to prepare a ligand-free crystal simply by soaking the ligand-bound crystal for a day or so in a stabilizing/cryoprotectant solution lacking ligands. Alternatively, one can challenge the ligand-bound crystals directly with other ligands and observe the difference in occupancy of the two ligands by isomorphous methods (**Fig. 4**). These *in crystallo* competition assays are of particular use in determining the binding affinity of a new ligand relative to a ligand whose dissociation constant has been well measured.
9. In the Zhou and MacKinnon reference, Fo-Fc maps are referred to incorrectly as *isomorphous* difference Fourier maps. Such maps with Fo-Fc coefficients are *residual* difference Fourier maps (a special case of which are the *omit* maps). The difference between these types of maps is substantial and should not be confused.
10. The anomalous difference Fourier map is generated analogously to the isomorphous map, with the signed difference between each primary reflection and its Friedel mate as the coefficients of the Fourier transform, with phases (experimental or calculated by back transformation of a model) retarded by 90°. If an anomalous map is specified in CCP4 or CNS, the phases are “automatically” retarded.

Acknowledgments

J. Connolly, B. Millard, E. Hayes, M. Sperber, Q. Wan, A. Bowser, and S. Flemer are the students and technicians who participated in data collection or



compound synthesis for the applications discussed in this chapter, with collaborators K. Trybus (kinesin, actin-cytochalasin) and J. Madalengoitia (combinatorial chemistry). The author would like to thank DOE-EPSCoR (DE-FG02-00ER45828, to S. S. Wallace), the Vermont Cancer Center (Lake Champlain Cancer Research Organization), and National Institutes of Health HL38113 (Trybus) for support of this work.

References

1. Blundel, T. L. and Johnson, L. N. (1976) *Protein Crystallography.* Academic Press, New York, NY.
2. Petsko, G. A. (1985) Preparation of isomorphous heavy-atom derivatives. *Methods Enzymol,* **114,** 147–156.
3. Rould, M. A. (1997) Screening for heavy-atom derivatives and obtaining accurate isomorphous differences. *Methods Enzymol.* **276,** 461–472.
4. Islam, S. A., Carvin, D., Sternberg, M. J. E., and Blundell, T. L. (1998) MAD, a data bank of heavy-atom binding sites in protein crystals: a resource for use in multiple isomorphous replacement and anomalous scattering. *Acta Cryst.* **D54,** 1199–1206.
5. Garman, E. and Murray, J. W. (2003) Heavy-atom derivatization. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 1903–1913.
6. Dauter, Z., Dauter, M., and Rajashankar, K. R. (2000) Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. *Acta Crystallogr. D Biol. Crystallogr,* **56,** 232–237.
7. Evans, G. and Bricogne, G. (2002) Triiodide derivatization and combinatorial counter-ion replacement: two methods for enhancing phasing signal using laboratory Cu Kalpha X-ray equipment. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 976–991.
8. Sun, P. D., Radaev, S., and Kattah, M. (2002) Generating isomorphous heavy-atom derivatives by a quick-soak method. Part I: test cases. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 1092–1098.
9. Boggon, T. J. and Shapiro, L. (2000) Screening for phasing atoms in protein crystallography. *Structure Fold. Des.* **8,** R143–R149.
10. Garman, E. F. and Grime, G. W. (2005) Elemental analysis of proteins by microPIXE. *Prog. Biophys. Mol. Biol.* **89,** 173–205.
11. Rould, M. A. and Carter, C. W., Jr. (2003) Isomorphous difference methods. *Methods Enzymol.* **374,** 145–163.
12. Collaborative_Computational_Project. (1994) The CCP4 suite: programs for protein crystallography. *Acta Cryst.* **D50,** 760–763.
13. Brunger, A. T., Adams, P. D., Clore, G. M., et al. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905–921.
14. Matthews, B. W. and Czerwinski, E. W. (1975) Local scaling: a method to reduce systematic errors in isomorphous and anomalous scattering measurements. *Acta Cryst.* **A31,** 480–487.
15. Blessing, R. H. (1997) LOCSCL: a program to statistically optimize local scaling of single-isomorphous-replacement and single-wavelength-anomalous-scattering data. *J. Appl. Crystallogr.* **30,** 176–177.
16. Freer, S. T. (1985) Classic (Fo-Fc) Fourier refinement. *Methods Enzymol.* **115,** 235–237.
17. Terwilliger, T. C. and Berendzen, J. (1996) Bayesian difference refinement. *Acta Cryst.* **D52,** 1004–1011.
18. Terwilliger, T. C. (2003) SOLVE and RESOLVE: automated structure solution and density modification. *Methods Enzymol.* **374,** 22–37.

19. Brunger, A. T. (1997) Free R value: cross-validation in crystallography. *Methods Enzymol.* **277,** 366–396.
20. Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A., and Vieth, M. (2004) Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **47,** 45–55.
21. Berman, H. M., Battistuz, T., Bhat, T. N., et al. (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 899–907.
22. Zhou, Y. and MacKinnon, R. (2004) Ion binding affinity in the cavity of the KcsA potassium channel. *Biochemistry* **43,** 4978–4982.
23. Shipps, G. W., Jr., Pryor, K. E., Xian, J., Skyler, D. A., Davidson, E. H., and Rebek, J., Jr. (1997) Synthesis and screening of small molecule libraries active in binding to DNA. *Proc. Natl. Acad. Sci. USA* **94,** 11,833–11,838.
24. English, A. C., Done, S. H., Caves, L. S., Groom, C. R., and Hubbard, R. E. (1999) Locating interaction sites on proteins: the crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins* **37,** 628–640.
25. Mattos, C. and Ringe, D. (2001) Proteins in organic solvents. *Curr. Opin. Struct. Biol.* **11,** 761–764.
26. Xing, Y. and Xu, W. (2003) Crystallization of the PX domain of cytokine-independent survival kinase (CISK): improvement of crystal quality for X-ray diffraction with sodium malonate. *Acta Cryst.* **D59,** 1816–1818.
27. Holyoak, T., Fenn, T. D., Wilson, M. A., Moulin, A. G., Ringe, D., and Petsko, G. A. (2003) Malonate: a versatile cryoprotectant and stabilizing solution for salt-grown macromolecular crystals. *Acta Cryst.* **D59,** 2356–2358.

10

Substructure Determination in Multiwavelength Anomalous Diffraction, Single Anomalous Diffraction, and Single Isomorphous Replacement With Anomalous Scattering Data Using *Shake-and-Bake*

G. David Smith, Christopher T. Lemke, and P. Lynne Howell

Summary

A general method for selenium and sulfur substructure determination using the *Shake-and-Bake* (*SnB*) algorithm as implemented in *SnB* in conjunction with anomalous difference E magnitudes is presented. The protocol can be used for Se-Met multiwavelength anomalous diffraction, seleneomethionine single anomalous diffraction (SAD), S-SAD and S/Se-single isomorphous replacement with anomalous scattering data and with minor modifications for other heavy atom derivatives, such as the location of halides.

Key Words: Anomalous scattering; substructure determination; *SnB*; *ab initio* phasing; single isomorphous replacement; SIR; multiwavelength anomalous diffraction; MAD; single anomalous diffraction; SAD; single isomorphous replacement with anomalous scattering; SIRAS.

1. Introduction

In 1954, Perutz et al. introduced the multiple isomorphous replacement (MIR) technique to protein crystallography *(1)*. The MIR technique has been used routinely to solve structures for many decades but suffers from a number of problems, namely that derivatization is often trial-and-error in nature and that the derivatized crystals typically show some degree of nonisomorphism or are of lower data quality than the original native crystals. These problems can largely be eliminated if the anomalous scattering signal from either native sulfur or from selenium in a recombinant selenomethionyl protein is utilized. The first protein structure determined using an anomalous signal was that of crambin *(2)*, which used the anomalous signal from sulfur with Cu $K\alpha$ radiation. Routine use of the anomalous scattering signal, however, did not become popular until

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

Hendrickson et al. in 1990 *(3)* demonstrated the utility of coupling the systematic incorporation of selenium in the form of seleneomethionine (Se-Met) with the multiwavelength anomalous diffraction method (MAD). This method (Se-Met MAD) is now the method of choice for most new protein structure determinations. In a typical Se-Met MAD experiment, a single frozen selenomethionine-derivatized protein crystal is used to collect data at three wavelengths: the selenium edge (minimum f′), the selenium peak (maximum f″), and a remote wavelength, often of higher energy. In order to obtain these specific wavelengths, radiation from a synchrotron must be used. In addition to the ability to tune the wavelength, synchrotron radiation provides increased X-ray flux, which translates into data of higher resolution, as well as an improvement in the signal-to-noise ratio.

To avoid the additional step of selenomethionine incorporation, there has been an increased interest in recent years in exploiting the sulfur anomalous signal. The absorption edge of sulfur occurs at 5.02 Å, a wavelength that precludes the use of the MAD technique. Anomalous data from sulfur has therefore been measured at a synchrotron at a wavelength of 1.77 Å and used in conjunction with the single anomalous diffraction (SAD) technique to solve a number of *de novo* protein crystal structures *(4–8)*. The value of the sulfur anomalous signal (f″) measurable at 1.77Å is 0.72 electrons. Although at 1.54 Å the sulfur anomalous signal is only 0.56 electrons, this signal has been measured using both synchrotron radiation *(9)* and Cu *K*α radiation *(10–13)* and used successfully to determine the structure of a number of proteins. A combination of sulfur and selenium anomalous data measured using Cu *K*α radiation has also been used in a single isomorphous replacement with anomalous scattering (S/Se-SIRAS) *(10)*. With the development of chromium-rotating anode targets it has recently been demonstrated that Cr *K*α radiation with a wavelength of 2.29 Å can also be successfully used for S-SAD phasing *(14)*. Regardless of the type of experiment performed (Se-Met MAD, S-SAD, Se-Met SAD, or S/Se-SIRAS) or the source of the radiation (synchrotron vs home source), the first essential step in the phasing process once data have been measured is the determination of the anomalous scattering substructure.

In the early days of MAD, the selenium atom substructure was usually determined by classical Patterson techniques, but as the substructures became larger these traditional methods became ineffective. Patterson methods as implemented in *CNS (15)* or *SOLVE (16)* have been used routinely to solve substructures with as many as 30 atoms in the substructure. The *Shake-and-Bake* (*SnB*) algorithm *(17)* as implemented in the direct methods programs, *SnB (18)* and *SHELXD (19)*, has proven to be a very powerful tool for determining the positions of large numbers of substructure atoms. The first application of *SnB* using renormalized anomalous difference E magnitudes *(20)* to solve a selenium atom

substructure was that of a known structure with eight selenium atoms *(21)*; the next application was that of a unknown protein structure containing 30 independent selenium atoms *(22)*. Based on these results and further analysis, optimal parameters for the *DREAR* data reduction package *(23)*, renormalization of the anomalous difference E magnitudes *(20)*, and *SnB* were developed *(24)*.

In the remainder of this chapter, we shall describe a general method for anomalous substructure determination using *SnB (18)* in conjunction with renormalized anomalous difference E magnitudes *(20)*. The protocol presented can be used for Se-Met MAD, S-SAD, Se-Met SAD, and S/Se SIRAS data and reflect our experiences with *S*-adenosylhomocysteine hydrolase *(22,24)* and *Escherichia coli* argininosuccinate synthetase *(10)*. The procedure can also be used with minor modifications for other heavy-atom derivatives.

2. Materials

For the methods described in this chapter, two key materials are required: protein crystals and X-rays. The protein crystals must contain an anomalous scatterer. This could be either native (sulfur methionine) or selenomethionine-derivatized protein, or could be a heavy–atom derivative or a metal ion that is naturally associated with the protein. The X-rays could be from either a Cu *K*α or Cr *K*α home source, or a tunable synchrotron X-ray source. The general method outlined here will focus on substructure determination for the Se-Met MAD or S-SAD cases, but will work for any substructure determination.

Once crystals are available and an X-ray source is chosen, their combination determines the possible methods. If synchrotron radiation and selenomethionine protein crystals are available then Se-Met MAD data collection is the method of choice. If synchrotron radiation and native protein crystals are available, then S-SAD at 1.77Å can be attempted. If Cu *K*α or Cr *K*α radiation and native and/or selenomethionine protein crystals are available, then S-SAD, S/Se-SIRAS, or Se-Met SAD may be attempted.

3. Methods

3.1. General Guidelines for Data Collection

Regardless of the type of experiment to be performed, or whether the experiment is to be conducted at home or at a synchrotron source, a cryocooled crystal is required (*see* **Note 1**). The goal of the experiment is to accurately measure Bijvoet pairs of reflections. Each of these Bijvoet pairs will differ by only a small amount, and it is this small difference that corresponds to the anomalous signal and will be used to determine the substructure. Thus, it behooves the experimentalist to perform as accurate an experiment as possible. One way in which accuracy can be improved is by the measurement of highly redundant data. There are certain to be outliers among the measured data, reflecting systematic

or random errors in the measurements. If two equivalent reflections differ significantly, it is not possible to determine which is in error. Even when three replicate measurements are present, elimination of an erroneous measurement can be problematic. Data reduction programs such as *d*TREK (25)* and *DENZO/SCALEPACK (26)* can and do remove outliers, but only when there is sufficient redundancy. An additional benefit of high redundancy is the improvement in the signal-to-noise ratio (*see* **Note 2**). The experiment should also be designed so that maximum completeness is attained. Another important consideration is the measurement of as much low-resolution data as is possible, even at the expense of high-resolution data. The low-resolution data will improve the quality of the maps, but more importantly these data are critical in the formation of sufficient numbers of triple invariants (*see* **Note 3**) to solve the substructure from the direct methods procedures.

3.2. Synchrotron Source

In the case of a MAD experiment performed at a synchrotron, at least a hemisphere of data should be measured in order to improve the accuracy of the signal and the signal-to-noise ratio (*see* **Note 2**). Although synchrotron time is limited, the experimentalist should ensure that more than just the minimum amount of data required, as predicted using various computer programs, is measured. If radiation damage is a concern or if multiple crystals are to be used to measure a single dataset, then the experimentalist should consider using inverse beam geometry or orienting the crystal such that a mirror plane is perpendicular to the rotation axis (*see* **Note 4**). An important consideration for performing a MAD experiment is to verify that the wavelength is adjusted to maximize and minimize f″ and f′, respectively (*see* **Note 5**).

3.3. Home Source

At a home source, the reduction of the X-ray flux as compared with a synchrotron source will require longer exposure times for each image and will necessarily result in a reduced signal-to-noise ratio. Thus, with a home source it is even more critical to measure highly redundant data (*see* **Note 6**).

3.4. Integration and Scaling

Throughout data processing the anomalous pairs should be processed separately. As always, the various parameters in the program package used to integrate and scale the redundant data should be tailored to the data. A careful look at the final data reduction statistics is important to ascertain at which point the high-resolution reflections are basically just noise. One would like to see $|F^2|$ greater than $3\sigma(|F^2|)$ for at least 90% of the data in the highest resolution shell. Although this is a more stringent criteria than would typically be used, statis-

tics such as these will insure that the magnitudes of the high resolution, renormalized anomalous difference structure factors (diffE's) will be significant with respect to their standard deviations.

3.5 SnB

A complete description of the *SnB* algorithm as implemented in the direct methods program, *SnB*, has been published (***[18,27]***; *see also* the website, http://www.hwi.buffalo.edu/SnB/) and will not be described in any detail here. The first thing that is required when running *SnB* is to create a "configuration" file, which is accomplished by invoking the "General Information" menu. Information such as unit cell constants, unit cell contents (*see* **Note 7**), and other information are entered here. Once the general information has been entered, one should proceed to the generation of E's ("Create E's" menu).

A very nice feature of the *SnB* program package is the inclusion of the *DREAR* data reduction suite ***(23)***, and the programs, *LOCSCL* ***(28)*** and *DIFFE* ***(20)***. The *DREAR* package consists of four programs, *SORTAV*, *BAYES*, *LEVY*, and *EVAL*. *SORTAV* reformats and sorts the data. *BAYES* applies Bayesian statistics ***(29)*** to improve the weak data, as well as to either eliminate or correct negative intensities; reflections that are negative and differ from zero by four or more standard deviations will be rejected and noted in the output-listing file. *LEVY* obtains estimates of the absolute scale factor and overall isotropic temperature factor, whereas *EVAL* calculates normalized structure factors for the entire set of data, including all anomalous pairs. The normalized structure factors that will be used to generate the renormalized anomalous difference structure factors can be calculated either in *EVAL* or in *BAYES* (*see* **Note 8**). It is very important to carefully examine the output-listing files from these data processing programs in order to exclude any erroneous data from subsequent calculations and to ascertain that all procedures worked properly. Both *BAYES* and *EVAL* provide comparisons of the E-statistics with the theoretical ones, and these should be in reasonable agreement. The largest E-values, sorted in descending order, should also be examined. In most cases, E-magnitudes should not exceed four. Values larger than six or eight are unreasonable, are most likely incorrect, and should be eliminated by deleting these reflections from the original reflection file and reprocessing the data through *DREAR* before proceeding further.

If desired, *LOCSCL* is the next program to be run, but it has been our experience that when accurate data are available, local scaling diminishes real differences between pairs of anomalous data. Thus, in most cases *LOCSCL* should be bypassed. The generation of the renormalized diffE is the final step in the procedure, and *DIFFE* accomplishes this task. In nearly all cases, default values can be used. However, when less than optimal data are being used, one or

more of the parameters (referred to as the difference E limits in the "Create E's" menu: the min(F)/Sig(F) to be included as well as t_{max}, x_{min}, y_{min}, z_{min}) may have to be relaxed in order to obtain enough reflections for subsequent phasing (*see* **Note 9**). The output listing of *DIFFE* contains a plot of diffE vs $\sin(\theta)/\lambda$. This plot should be reasonably flat with neither an upturn nor downturn at larger values of $\sin(\theta)/\lambda$. Such behavior usually suggests that the resolution range is too large and that inaccurate diffEs have been included.

At this point, the actual phasing process can begin ("Reflections and Invariants" menu). For an N atom substructure, it is usually suggested that 20–30N phases be used to generate 200–300N triple invariants. If too few phases are used, then it is likely that the required number of triple invariants cannot be obtained (*see* **Note 10**). With this exception, the default values for most *SnB* parameters both in the generation of reflections and invariants, and in the number of "Trials and Cycles" can be used. A data resolution minimum and maximum cutoff also needs to be applied at this juncture. No low-resolution cutoff should be applied to the data; as stated earlier (*see* **Note 3**) these are the reflections that are critical for the formation of sufficient triple invariants. Although not generally recommended, a high-resolution cutoff can be applied to the data. Truncation of data to 4 Å and even 5 Å has been shown to yield solutions *(24)*. However, it should be noted that the percentage success rate for lower resolution data is less than for higher resolution data, although the number of solutions per unit of CPU time is significantly improved (i.e., the solutions are obtained faster). Use of a high-resolution cutoff may in certain instances be advantageous but generally is not recommended. Finally, in the "Phase Refinement" and "Fourier Refinement" menus, the parameter-shift technique should be chosen as this has been shown in most cases to be superior to that of tangent refinement; "twice-baking" should be invoked as this may allow additional substructure atoms to be located. Also in the "Fourier Refinement" menu are the constraints that are applied, such as the number of peaks to select, the map resolution, the minimum interpeak distance, and the minimum distance between symmetry-related peaks. In all cases the default values provided can be used.

Before initiating and submitting the job, the number of trials and cycles to be run needs to be set in the "Trials and Cycles" menu (*see* **Note 11**). The job is initiated on the "Submit Job" menu. A number of options are available, including the ability to run the job on multiple processors or in batch mode. How the job is submitted will be highly dependent on the computing environment of the user. Once the phasing procedure is initiated, the user can follow the progress by observing the shape of the histogram of bins of R_{min} ("Evaluate Trials" menu). Although the large majority of the trials will result in nonsolutions, there will be a small subset of trials which correspond to correct solutions, and these can be identified by the bimodal distribution of R_{min} as shown in **Fig. 1**. The trace of

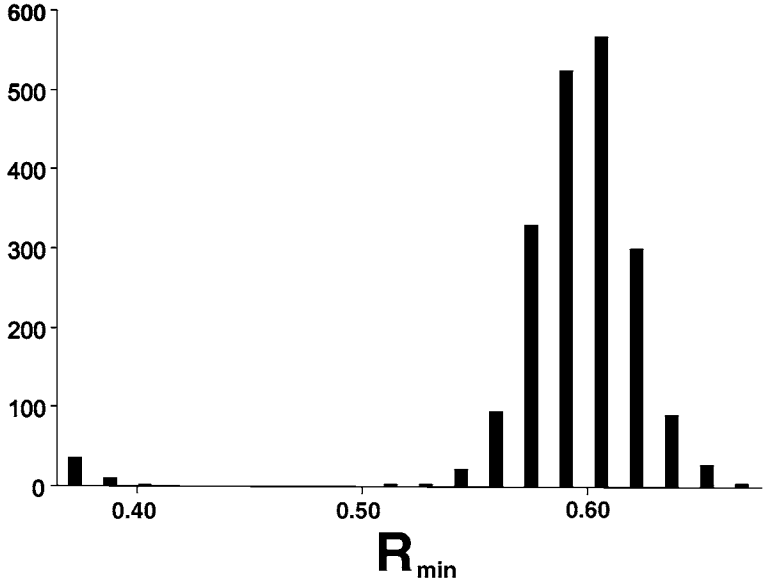


Fig. 1. Histogram illustrating the bimodal distribution of R_{min} for 2000 random atom trials for argininosuccinate synthetase.

R_{min} vs cycle number is also an indicator of a correct solution (*see* **Note 12**). In the case of an incorrect solution, R_{min} will decrease slightly but maintain a relatively large value. In contrast, R_{min} for a correct solution will exhibit a dramatic decrease at some arbitrary cycle number and will remain at this low value. Typical traces of R_{min} for a correct and incorrect solution are shown in **Fig. 2**.

Following the conclusion of the job, the user can examine the various sets of coordinates that correspond to trials with low values of R_{min} (*see* **Note 13**). The presence of multiple solutions is an advantage because peaks that occur in more than one solution are most likely correct. Because a set of N random atoms was the starting point for the phasing procedure, different solutions may correspond to different origins, be related by different symmetry elements, or be different enantiomorphs. Thus, a casual examination of the coordinates will reveal little. A stand-alone program (*see* **Note 14**) exists that will make comparisons of the sets of coordinates taking into consideration the various origins, symmetry elements, and enantiomorphic differences *(30)*.

Following the determination of the atom positions of the substructure, a variety of programs are available to produce the complete protein structure. In the case of argininosuccincate synthetase *(10)*, the complete protein structure was traced automatically using default values with *RESOLVE (16)*. *SnB* and the

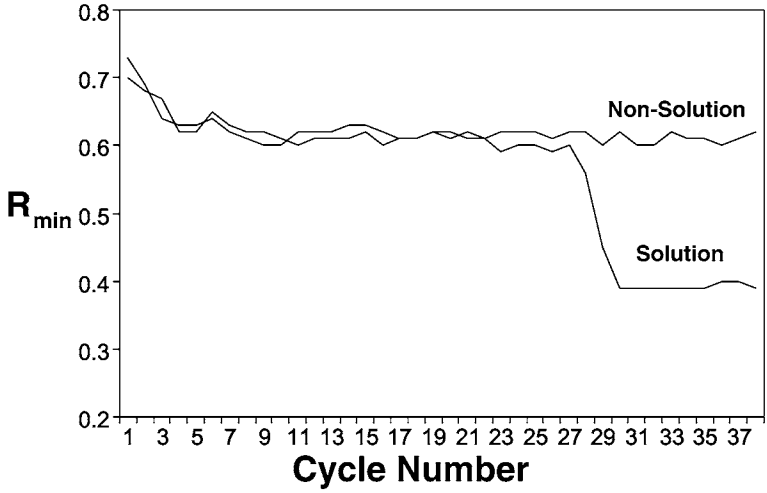


Fig. 2. Trace of R_{min} vs cycle number for a solution and a nonsolution.

PHASES (31) program package have recently been integrated to provide a complete phasing package. This program is called *BnP (32)* (*see* **Note 15**).

3.6. Conclusions

By exploiting the anomalous signal from a native sulfur containing protein or by systematically incorporating an anomalous scatterer into a protein, one of the major barriers to routine structure determination, namely the phase problem, has essentially been eliminated. The determination of the anomalous substructure is the key to this phasing process. We have presented a general protocol for the direct methods program, *SnB*, which is extremely successful at solving both small and large substructures. *SnB* has been used routinely to determine substructures of 30–50 selenium atoms, and in the case of ketopantoate hydroxymethyl transferase 160 out of 180 selenium atoms were located *(33)*.

4. Notes

1. At room temperature few protein crystals will survive longer than a minute of exposure to synchrotron radiation, and although radiation from a home source is much less intense, the increased duration of the experiment is also likely to result in severe damage to an unprotected crystal (*see* Chapter 1 on cryocooling).
2. The majority of the data reduction programs calculate the mean amplitude, $\langle|F_{\mathbf{h}}|\rangle = (1/n) \sum |F_{\mathbf{i}}|$, and the standard deviation of the mean, $\sigma(|F_{\mathbf{h}}|) = [(1/n\,(n-1) \sum (|F_{\mathbf{i}}| - |F_{\mathbf{h}}|)^2]^{1/2}$. As *n* increases, the standard deviation from the mean is reduced significantly. The standard deviation of the mean should not be confused with the standard deviation of the population, $\sigma(|F_{\mathbf{h}}|) = [(1/(n-1) \sum (|F_{\mathbf{i}}| - |F_{\mathbf{h}}|)^2]^{1/2}$, which does not exhibit such a dramatic decrease as *n* increases.

3. Triple invariants are three reflections whose Miller indices sum to zero ($h_1k_1l_1 + h_2k_2l_2 + h_3k_3l_3 = 0$). If the values of the Miller indices for one reflection are small, as would be the case for a low-resolution reflection, then it becomes much more likely that this reflection will interact with higher resolution data to form many triple invariants. In the case of selenomethionyl *S*-adenosylhomocysteine hydrolase ***(22)***, two sets of independent data were measured. The dataset with a low-resolution limit of 20 Å had only half the success rate in determining the selenium atom substructure as the dataset with a low-resolution limit of 50 Å.
4. Bijvoet pairs can be recorded simultaneously by alignment of the crystal with a mirror plane of diffraction symmetry perpendicular to the rotation axis, or Friedel pairs can be recorded in an "inverse beam" experiment. Inverse beam geometry measures Friedel pairs using both the forward and reverse directions of the incident X-ray beam. In a real experiment, diffraction images and their Friedel equivalents are recorded at crystal positions related by 180° rotation about any axis perpendicular to the incident beam, usually the data collection axis. For a complete set of data, small wedges of data, 180° apart, are therefore measured to simulate inverse beam geometry. The inverse beam experiment requires neither crystal symmetry nor crystal alignment, and is well suited to crystals mounted in random orientations.
5. The features of an X-ray absorption edge can be very sharp, with the energies of the inflection and peak absorption separated by as little as 2 eV. Therefore, in the MAD experiment it is critical to determine the edge and peak wavelengths experimentally by recording the absorption edge from the labeled macromolecule at the time of the experiment.
6. In the case of argininosuccinate synthetase (space group I222), 720º of data ($\Delta\phi$ = 0.5º) were measured, resulting in redundancies of approx 14.
7. The anomalous data that will be used to solve a substructure corresponds to only the substructure atoms, not the entire protein. Thus, in the case of a protein that contains either 25 methionine or selenomethionine residues, the contents of the unit cell would be 25 sulfur or selenium atoms.
8. Normalized structure factors can be calculated on the basis of the absolute scale factor and overall isotropic temperature factor obtained from a Wilson plot ($|E_{\mathbf{h}}| = K^{1/2} |F_{\mathbf{h}}|_{meas}/(\varepsilon_{\mathbf{h}}\Sigma_i f_{\mathbf{i}}^2)^{1/2} \exp [-B_{iso} (\sin\theta)^2/\lambda^2]$ or by calculating the mean $|F^2|$ in shells of $\sin(\theta)/\lambda$, approx 100 data per shell, and dividing each measured $|F_{\mathbf{h}}^2|$ by $\langle|F^2|\rangle$ for that shell, $E_{\mathbf{h}}^2 = |F_{\mathbf{h}}^2| / \langle|F^2|\rangle$. In cases where the high-resolution limit is less than 2.5–3.0 Å, the Bayesian E-values should be used.
9. Default values are built into the *SnB* interface and usually can be used as is. The purpose of t_{max} is to eliminate unreliably large values of ($||E_+| - |E_-||$) pairs that lie at the tails of the E-distribution; a value of six is usually suggested. x_{min} will eliminate those reflections where $E/\sigma(E)$ is less than some value, usually three. This insures that only the most significant data are included in the generation of diffEs. y_{min} eliminates those pairs for which $\Delta E/[\sigma^2(E_+) + \sigma^2(E_-)]^{1/2}$ is deemed to be not significant; a value greater than 1.0 is usually suggested. z_{min} eliminates those diffEs that are less than optimal; a value of 3.0 is usually suggested. However, if an insufficient number of diffEs are generated, z_{min} can be reduced to 2.75 or 2.5. A

complete description of the suggested values for these parameters can be found in Howell et al. *(24)*.

10. If not enough triple invariants are generated, *SnB* stops and informs the user of this fact. The best remedy is to increase the number of phases, decrease the number of triples, or a combination of both. It is important to generate as many triples as possible (up to the requested amount) given a specific number of phases. For a given triple invariant ($\phi_{\mathbf{h}} + \phi_{\mathbf{k}} + \phi_{\mathbf{l}} \approx 0$, where *h*, *k*, and *l* refer to sets of Miller indices), the probability that the phases of the three sets of Miller indices adds to zero is proportional to the product of the three E-values, $E_{\mathbf{h}}$, $E_{\mathbf{k}}$, and $E_{\mathbf{l}}$. When the probability is small, the phase distribution is flat and it is likely that the sum of the phases will be different from zero. Thus, it is important to use E-values with the largest magnitudes for the generation of the triple invariants.
11. The number of trials is by default set to 1000. This is many more trials than is usually required to find a solution. The job is therefore usually terminated once a sufficient number of solutions have been found. However, for large substructures, the number of trials may have to be increased to several thousand. Determining the correct number of cycles between real and reciprocal space to use per trial is critical, as if too few cycles are chosen one risks not finding a solution. However, as the number of cycles per trial increases so does the computational overhead. A good compromise appears to be 2N cycles, where N is the number of atoms in the substructure. For substructures smaller than 10, the number of cycles should never be less than 15 or 20.
12. In order to be able to examine all traces, the user must request the traces of all trials to be saved, not just the trace of the solution with the smallest R_{min} value ("Trials and Cycles" menu).
13. Only the coordinates of the random atom trial with the smallest value of R_{min} are saved. Coordinates of other trials that may be potential solutions can be generated once the trial number of a solution is known. In the "Evaluate Trials" menu, the "View Sorted Trials" button will list the trial numbers of solutions in descending order of R_{min}. Once the trial number of a small number of solutions is known, a single *SnB* trial is then initiated starting at a particular trial number, e.g., number of trials = 1, start at trial = 233 ("Trials and Cycles" menu), where trial 233 was previously identified as a solution. It is imperative that individual trials that reproduce a longer job be run on the same computer with the same random number seed, to ensure that the same solution is found.
14. The program, *NANTMRF*, is available from G. D. Smith at e-mail address, gdsmith@sickkids.ca.
15. The *SnB* setup menu in *BnP* for substructure determination is comparable to the "Reflection and Invariants," "Phase Refinement," "Fourier Refinement," and "Trials and Cycles" menus in *SnB*. The program also includes an automatic solution identification criterium that allows the job to be terminated when a solution is found and also incorporates *NANTMRF (30)* and a number of other features for comparing the peaks found in different solutions. *BnP* is available for downloading at http://www.hwi.buffalo.edu/BnP/. (Make sure to add the slash character "/"at the end of the URL.)

Acknowledgments

This work was supported in part by grants from the National Institutes of Health, EB002057 (G. D. S.) and the Canadian Institutes of Health Research, MT13337 (P. L. H.) and MOP 43998 (P. L. H.). P. L. H. is the recipient of a Canada Institutes of Health Research Investigator Award and C. T. L. was the recipient of studentship awards from the Canadian Institutes of Health Research, the University of Toronto, and the Hospital for Sick Children.

References

1. Green, D. W., Ingram, V. M., and Perutz, M. F. (1954) The Structure of hemoglobin. IV. Sign determination by the isomorphous replacement method. *Proc. R. Soc. London Ser. A* **225,** 287–307.
2. Hendrickson, W. A. and Teeter, M. M. (1981) Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature (London)* **290,** 107–113.
3. Hendrickson, W. A., Horton, J. R. and LeMaster, D. M. (1990) Selenomethionyl proteins produced for analysis of multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J*. **9,** 1665–1672.
4. Gordon, E. J., Leonard, G. A., McSweeney, S., and Zagalsky, P. F. (2001) The C_1 subunit of α-crustacyanin: the *de novo* phasing of the crystal structure of a 40 kDa homodimeric protein using the anomalous scattering from S atoms combined with direct methods. *Acta Cryst.* **D57,** 1230–1237.
5. Liu, Z. J., Vysotski, E. S., Chen, C. J., Rose, J .P., Lee, J., and Wang, B. B. (2000) Structure of the Ca2+-regulated photoprotein obelin at 1.7 Å resolution determined directly from its sulfur substructure. *Protein Science* **9,** 2086–2093.
6. Micossi, E., Hunter, W. N., and Leonard, G. A. (2002) *De novo* phasing of two crystal forms of tryparedoxin II using the anomalous scattering from S atoms: a combination of small signal and medium resolution reveals this to be a general tool for solving protein crystal structures. *Acta Cryst.* **D58,** 21–28.
7. Doan, D. N., and Dokland, T. (2003) Structure of the nucleocapsid protein of porcine reproductive and respiratory syndrome virus. *Structure* **11,** 1445–1451.
8. Lartigue, A., Greuz, A., Briand, L., et al. (2004) Sulfur-SAD crystal structure of a phermone binding protein from the honeybee *Apis mellifera L. J. Biol. Chem.* **279,** 4459–4464.
9. Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G., and Sheldrick, G. M. (1999) Can anomalous signal of sulfur become a tool for solving protein crystal structures? *J. Mol. Biol.* **28,** 83–92.
10. Lemke, C. T., Smith, G. D., and Howell, P. L. (2002) S-SAD, Se-SAD and S/Se-SIRAS using Cu *K*α radiation: why wait for synchrotron time? *Acta Cryst.* **D58,** 2096–2101.
11. Ramagopal, U. A., Dauter, M., and Dauter, Z. (2003) Phasing on anomalous signal of sulfurs: what is the limit? *Acta Cryst.* **D59,** 1020–1027.

12. Debreczeni, J. E., Bunkoczi, G., Ma, Q., Blaser, H., and Sheldrick, G. M. (2003) In-house measurement of the sulfur anomalous signal and its use for phasing. *Acta Cryst.* **D59,** 688–696.
13. Debreczeni, J. E., Bunkoczi, G., Girmann, B., and Sheldrick, G. M. (2003) In-house phase determination of the lima bean trypsin inhibitor: a low resolution sulfur-SAD case. *Acta Cryst.* **D59,** 57–66.
14. Yang, C., Pflugrath, J. W., Courville, D. A., Stence, C. N., and Ferrara J. D. (2003) Away from the edge: SAD phasing from the sulfur anomalous signal measured in-house with chromium radiation. *Acta Cryst.* **D59,** 1943–1957.
15. Brünger, A. T., Adams, P. D., Clore, G. M, et al. (1998) Crystallography and NMR system (*CNS*): a new software suite for macromolecular structure determination. *Acta Cryst.* **D54,** 905–921.
16. Terwilliger, T. C. and Berendzen, J. (1999) Automated MIR and MAD structure solution. *Acta Cryst.* **D55,** 849–861.
17. DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R., and Hauptman, H. A. (1994) Structure solution by minimal function phase refinement and Fourier filtering: theoretical basis. *Acta Cryst*. **A50,** 203–210.
18. Weeks, C. M. and Miller, R. (1999) The design and implementation of *SnB* v2.0. *J. Appl. Cryst.* **32,** 120–124.
19. Schneider, T. R. and Sheldrick, G. M. (2002) Substructure solution with *SHELXD*. *Acta Cryst.* **D58,** 1772–1779.
20. Blessing, R. H. and Smith, G. D. (1999) Difference structure factor normalization for heavy-atom or anomalous-scattering substructure determinations. *J. Appl. Cryst.* **32,** 664–670.
21. Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A., and Blessing, R. H. (1998) The use of *SnB* to determine an anomalous scattering substructure. *Acta Cryst.* **D54,** 799–804.
22. Turner, M. A., Yuan, C. -S., Borchard, R. T., Hershfield, M. S., Smith, G. D., and Howell, P. L. (1998) Structure determination of selenomethionyl S-adenosylhomocysteine hydrolase using data at a single wavelength. *Nat. Struct. Biol.* **5,** 369–375.
23. Blessing, R. H. (1987) Data reduction and error analysis for accurate single crystal diffraction intensities. *Cryst. Rev.* **1,** 3–58.
24. Howell, P. L., Blessing, R. H. Smith, G. D., and Weeks, C. M. (2000) Optimizing *DREAR* and *SnB* parameters for determining Se-atom substructures. *Acta Cryst.* **D56,** 604–617.
25. Pflugrath, J. W. (1999) The finer things in X-ray diffraction data collection. *Acta Cryst.* **D55,** 1718–1725.
26. Otwinowski, Z. and Minor, W. (1997) Processing of X-ray diffraction data collected in oscillation mode. In: *Methods in Enzymology, Vol. 276,* (Carter, C. W. and Sweet, R. M., eds.), Academic Press, New York, NY, pp. 307–326.
27. Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R., and Usón, I. (2001) Ab initio phasing. In: *International Tables for Crystallography, Vol. F,* (Rossmann, M. G. and Arnold, E., eds.), Kluwer Academic Publishers, Dordrecht, Germany, pp. 333–345.

28. Blessing, R. H. (1997) *LOCSCL*: a program to statistically optimize local scaling of single-isomorphous-replacement and single-wavelength-anomalous-scattering data. *J. Appl. Cryst.* **30,** 176–177.
29. French, S. and Wilson, K. (1978) On the treatment of negative intensity observations. *Acta Cryst.* **A34,** 517–525.
30. Smith, G. D. (2002) Matching selenium-atom peak positions with a different hand or origin. *J. Appl. Cryst.* **35,** 368–370.
31. Furey, W. and Swaminathan, S. (1997) *PHASES-95*: a program package for processing and analyzing diffraction data from macromolecules. In: *Methods in Enzymology, Vol. 277,* (Carter, C. W. and Sweet, R. M., eds.), Academic Press, New York, NY, pp. 590–620.
32. Weeks, C. M., Blessing, R. H., Miller, R., et al. (2002). Towards automated protein structure deteremination: *BnP*, the *SnB-PHASES* interface. *Z. Kristallogr.* **217,** 686–693.
33. von Delft, F. and Blundell, T. L. (2003) Structure of *E. coli* ketopantoate hydroxymethyl transferase complexed with ketopantoate and Mg^{2+}, solved by locating 160 selenomethionine sites. *Structure* **11,** 985–996.

11

Substructure Determination in Isomorphous Replacement and Anomalous Diffraction Experiments

Ralf W. Grosse-Kunstleve and Thomas R. Schneider

Summary

The determination of the substructure of heavy atoms or anomalous scatterers is a central step in experimental-phasing procedures. We give an overview of commonly used methods for substructure determination, including estimation of substructure structure factors, Patterson methods, direct methods, dual-space recycling procedures, and methods for substructure refinement and completion. This chapter also includes an annotated list of available program packages.

Key Words: Patterson methods; direct methods; heavy atoms; dual-space recycling; fast translation function.

1. Introduction

Traditionally, experimental phasing of macromolecular structures involves heavy-atom soaks and the collection of two or more datasets: the diffraction intensities of the *native* crystal and that of the *derivative(s)*. This is often referred to as single or multiple isomorphous replacement (SIR, MIR). In recent years, because of the growing availability of tunable synchrotron radiation, it has become very popular to use crystals containing anomalous scatterers. Experiments in which the anomalous signal is explicitly measured are known as single or multiple anomalous diffraction experiments (SAD, MAD) ***(1)*** or alternatively single anomalous scattering experiments. Isomorphous replacement and anomalous scattering can also be combined (SIRAS, MIRAS).

Heavy atoms, which may at the same time be anomalous scatterers, can be naturally present, for example iron in heme proteins, but more often they are introduced artificially. Extensive overviews of procedures for the preparation of heavy-atom derivatives are given by **refs. *2–5***. These procedures are still commonly used. Recently, Dauter ***(6)*** (*see* Chapter 8) introduced the method of

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

quick halide soaks. A more complex but powerful and now very popular method for the introduction of anomalous scatterers is the replacement of methionine residues in protein structures by selenomethionine *(7)*. Interestingly, the data-collection technology developed for selenomethonine experiments has now advanced to the point where even the small anomalous signal from the native sulfur atoms *(8,9)* in proteins or phosphorus atoms in oligonucleotides *(10)* can in favorable cases lead to successful structure determination. There have also been reports *(11)* of a selenomethionine derivative being used together with a native crystal in an isomorphous replacement experiment. This is unusual but entirely possible if the degree of nonisomorphism is sufficiently small.

Once crystals containing heavy atoms or anomalous scatterers are available, experimental phasing can be viewed as a divide-and-conquer technique where the larger problem of determining the complete macromolecular structure is divided into two steps:

1. Given experimental diffraction data from one or several crystals, approximate substructure structure factors corresponding to the substructure of heavy atoms or the anomalous scatterers alone are computed. The substructure is solved based on the substructure structure factors using methods originally developed for the solution of small molecules.
2. Using the known substructure in combination with the measured diffraction data, algebraic or probabilistic methods are used to extrapolate phases for the full structure.

Once initial phases are found, the structure solution process continues with density modification, model building, and refinement. In this chapter, we focus on the determination of the substructure and some aspects of substructure refinement.

2. Overview of Commonly Used Methods

Most computer programs in use implement a combination of several procedures. In the following, the most important building blocks are described.

2.1. Estimation of Substructure Structure Factors

2.1.1. Isomorphous Differences

Because the number of atoms in a native macromolecular structure is usually much larger than the number of additional heavy atoms in a derivative, it is a valid approximation to assume $F_H << F_{PH}$, where F_H are the structure factor amplitudes corresponding to the substructure only, and F_{PH} the structure factor amplitudes of the derivative. This approximation leads to (*[2]*, **Subheading 6.2.** therein):

$$F_H << F_{PH} \Rightarrow F_{PH} - F_P \approx F_H \cos(\varphi_{PH} - \varphi_H) \quad (1)$$

The cosine term takes on values between –1 and 1. Therefore the isomorphous differences $F_{PH} - F_P$ as extracted from diffraction experiments on the

derivative and the native crystal are lower estimates of the true substructure structure factor modulus F_H: the true F_H can be larger, but they cannot be smaller than the observed isomorphous differences.

2.1.2. Anomalous Differences

Similar considerations lead to the following equation for anomalous differences $F^+_{PH} - F^-_{PH}$ (***[2]***, **Subheading 7.6.** therein):

$$F''_H << F'_{PH} \Rightarrow F^+_{PH} - F^-_{PH} \approx 2F''_H \sin(\varphi_{PH} - \varphi_H) \quad (2)$$

Here F''_H are the imaginary contributions to the structure factors of the anomalous scatterers. F'_{PH} is the sum of the structure factors of the macromolecular structure and the real contributions of the anomalous scatterers. The sine term also takes on values between –1 and 1. Therefore, the anomalous differences taken between the measured Bivoet mates F^+_{PH} and F^-_{PH} are lower estimates of the imaginary contributions of the anomalous scatterers.

The ratio between anomalous differences measured at different wavelengths is, to a first approximation, a constant for all reflections from a given crystal. The degree to which experimental data follow this expectation can be measured by calculating standard correlation coefficients (e.g., **ref.** ***12***) between anomalous differences originating from different measurements. The magnitude of the correlation coefficient is a very useful indicator of data quality (**Fig. 1A**) (*see* **Notes 1–3**).

2.1.3. F_A Structure Factors

In the case of MAD experiments it is possible to compute better estimates of the structure factors corresponding to the substructure. Commonly these estimates are referred to as F_A structure factors. Various algorithms for the computations of F_A structure factors are available: MADSYS ***(1)***, CCP4 REVISE ***(13)***, SOLVE ***(14)***, XPREP (Bruker AXS, Madison, WI). For good MAD data, F_A structure factors usually lead to significantly more efficient determination of the substructure. It has also been observed that F_A structure factors enable the solution of a substructure that could not be solved from anomalous differences measured at any one wavelength. However, if the MAD data are affected by systematic errors, such as intensity changes resulting from radiation damage, it is also possible that the corresponding F_A structure factors are not suitable for the substructure determination. In this case it is advantageous to attempt substructure determination with the wavelength collected first (ideally at the peak of the anomalous signal) ***(15)*** (*see* **Notes 4–8**).

2.2. The Phase Problem

In the second and third decades of the 20th century early X-ray crystallographers worked out that the observed diffraction intensities are directly related to

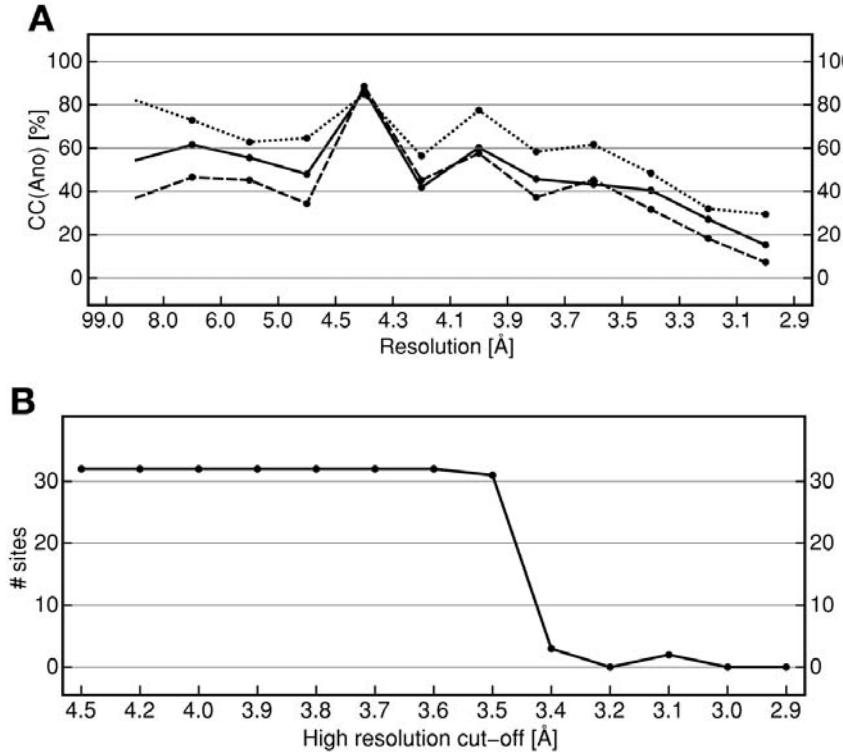


Fig. 1. **(A)** Plot of the correlation coefficient between signed anomalous differences measured at different wavelengths against resolution for the three-wavelength multiple anomalous diffraction (peak, inflection point, high-energy remote) dataset collected on the Se-Met substituted form of an acyltransferase *(56)*. The correlations shown are peak-remote (full line), inflection-remote (dashed line), and peak-inflection (dotted line). **(B)** Plot of the number of correct sites (as checked with SITCOM *[57]*) present in the top-scoring solution of the Se substructure of the acyltransferase as obtained by SHELXD vs high-resolution cut-off. The crystal structure contains 32 ordered Se-sites.

the Fourier transformation of the electron density of the crystal structure (not taking Lorentz factors, polarization factors, and other experiment-specific corrections into account):

$$\sqrt{I} \equiv |F| \propto FT(\rho) \tag{3}$$

Here I represents the observed intensities, $|F|$ the *structure factor amplitudes*, ρ is the electron density, and FT a Fourier transformation. The same relation more specifically:

$$F_h = |F_h|e^{i\varphi_h} = \frac{V}{N}\sum_x \rho(x)e^{2\pi ihx} \tag{4}$$

Here h is a Miller index, x the coordinate of a grid point in real space, N the total number of grid points, and V the volume of the unit cell. The complex structure factor F is also shown in the alternative representation as an amplitude $|F|$ combined with a phase (ϕ).

Obviously it is straightforward to compute the structure factor amplitudes from the electron density. Given complex structure factors it is equally easy to compute the electron density via a Fourier transformation:

$$\rho \propto FT^{-1}(F) \tag{5}$$

where FT^{-1} represents the inverse Fourier transformation. Unfortunately with current technology it is almost always impractical to directly measure both intensities and phases. Conventional diffraction experiments only produce intensities, the phases are not available. This is colloquially known as the phase problem of crystallography.

2.3. Techniques for Solving the Phase Problem

2.3.1. Patterson Methods

The Patterson function is defined as the Fourier transformation of the observed intensities:

$$Patterson \propto FT^{-1}(I) \tag{6}$$

This is a straightforward calculation requiring only the experimental observations as the input. Patterson ***(16)*** showed that the peaks in this Fourier synthesis correspond to *vectors between atoms* in the crystal structure.

2.3.2. Patterson Interpretation in Direct Space and in Reciprocal Space

In the classic textbook *Vector Space,* Buerger ***(17)*** demonstrates that under idealized conditions image-seeking procedures are capable of recovering the image of the electron density from the Patterson function. "Idealized conditions" essentially means fully resolved peaks in the Patterson function. In practice, this condition is only fulfilled for very small structures, but it still is possible to extract useful information from real Patterson maps. The basic idea is:

1. Postulate a hypothesis, for example a putative substructure configuration.
2. Test the hypothesis against the Patterson map.

The test involves the computation of vectors between the atoms of the putative substructure and the determination of the values in the Patterson map at the location of these vectors. This involves interpolation between grid points of the

map. The interpolated peak heights are usually the input for the computation of a Patterson score. Theoretically the minimum of all the peak heights found is the most powerful measure, but sum or product functions have also been used *(17)*. Nordman *(18)* suggests using the mean of a certain percentage of the lowest values.

It is also possible to work with the observed intensities in reciprocal space, without transforming them according to **Eq. 6**. Conceptually the procedure is even simpler:

1. Postulate a hypothesis, for example a putative substructure configuration.
2. Test the hypothesis against the observed intensities.

In this case the test involves the calculation of intensities for the putative structure and the evaluation of a function comparing these with the observed intensities, for example the standard linear correlation coefficient (e.g., **ref.** *12*). An advantage of this method is that it does not involve interpolations and therefore it should be intrinsically more accurate. However, the calculations are much slower than the computation of Patterson scores in direct space if done in the straightforward fashion suggested here. The key to making the reciprocal space approach feasible is the fast translation function devised by Navaza and Vernoslova *(19)*. We were able to show that the fast translation function is typically 200–500 times faster than the conventional translation function. The fast translation function was originally designed for solving molecular replacement problems, but it is also being used for the determination of substructures *(20,21)*.

2.3.3. Direct Methods

Direct methods were originally developed for the direct determination of phases without direct use of stereochemical knowledge. The fundamental approach is to start with a very small set of starting phases and to construct a more complete phase set by applying phase probability relationships. The expanded phase set in combination with the observed structure factors is used to compute an electron density map that is hopefully interpretable when stereochemical knowledge is taken into account.

The phase probability relations governing the phase extension procedure are usually based on the well-known tangent formula *(22)*. This formula is typically introduced as:

$$\tan(\varphi_h) = \frac{\sum_k |E_k E_{h-k}| \cos(\varphi_k + \varphi_{h-k})}{\sum_k |E_k E_{h-k}| \sin(\varphi_k + \varphi_{h-k})} \tag{7}$$

The tangent formula can be used to estimate an unknown phase φ_h given a set of known phases φ_k and φ_{h-k}. To avoid distraction, for the moment we will assume that the Es in this formula are analogous the structure factors F previ-

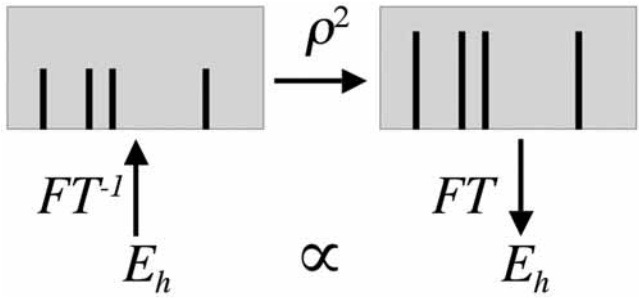


Fig. 2. The essence of direct methods. Normalized structure factors correspond to point atoms at rest. Squaring in direct space followed by a Fourier transformation leads to structure factors that are proportional to the original structure factors. The phases are identical.

ously introduced (the exact definition of the *E*s is given by **Eq. 11**). The derivation of the tangent formula employs the assumptions that the electron density is positive everywhere in the unit cell (positivity) and that all atoms are resolved (atomicity). To understand this it is useful to rewrite the tangent formula as a simpler, but mathematically equivalent, expression:

$$E_h \propto \sum_k E_k E_{h-k} \tag{8}$$

Comparison with the definition of the convolution (e.g., **ref.** ***23***) leads us to recognize:

$$\sum_k E_k E_{h-k} \equiv Convolution(E, E) \tag{9}$$

Application of the convolution theorem followed by application of **Eq. 5** leads to:

$$E_h \propto \sum_k E_k E_{h-k} = FT(FT^{-1}(E)^2) = FT(\rho^2) \tag{10}$$

This equation shows that the tangent formula uses positivity and atomicity to introduce a self-consistency argument as illustrated in **Fig. 2**. This argument is essentially the same one used in the derivation of the Sayre equation ***(24)***, which is slightly more complex than the tangent formula because it is formulated for atoms with Gaussian shapes rather than point atoms. This describes real crystal structures more closely, but in practice it is often more advantageous to eliminate the shape term and to work with normalized structure factors corresponding to point atoms. Normalized structure factors can be estimated from observed structure factor amplitudes $|F_h|$ by enforcing the expected average $\left\langle |E_h|^2 \right\rangle = 1$ in resolution shells:

$$|E_h|^2 = \frac{|F_h|^2}{\left\langle |F_h|^2 \,/\, \varepsilon_h \right\rangle} \tag{11}$$

To be precise, this equation yields estimates of the quasi-normalized structure factors. The term ε takes the multiplicity of the reflections into account and can be directly computed from the space group symmetry.

2.3.4. Dual-Space Recycling

The tangent formula alone often does not work efficiently for solving structures with many atoms ***(25)***. The most popular "direct methods" programs used in macromolecular crystallography today are the result of an evolution that transformed the pure phase-extension idea into complex multitrial search procedures. MULTAN ***(26)*** pioneered the multitrial approach but is still motivated by the phase-extension idea. RANTAN ***(27)*** and early versions of SHELX ***(28)*** mark the transition to random-seeded multitrial approaches that use the tangent formula in a recycling procedure to enforce self-consistency (**Fig. 2**). Shake-and-Bake ***(29)*** and more recent versions of SHELX ***(30)*** introduced the concept of dual-space recycling ***(31)***. Reciprocal-space phase manipulation based on the tangent formula, or the minimal function in the case of shake-and-bake, alternated with direct-space interpretation of Fourier maps. Shake-and-Bake picks peaks from the Fourier maps, taking a given minimum distance into account. The peaks are used in a structure–factor calculation to obtain new phases that are entered into the next cycle of phase manipulation. SHELXD ***(32)*** follows a similar approach but after selecting typically 1.3 times the expected number of sites N, eliminates 0.3 N of them in a random omit procedure. This imitates the common practice in macromolecular refinement of omitting part of a structure from a Fourier calculation in order to obtain an improved electron density in the next iteration. Picking peaks from an electron density to be used for the calculation of the next phase set is a very direct way of enforcing atomicity and taking into account the actual number of scatterers. The dual-space recycling procedure is very fast as it only works with a small fraction of the reflections. Typically 10–15% of the strongest E-values are used.

2.3.5. Direct Methods Recycling With Patterson Seeding

Conventional direct methods programs initialize the recycling procedure with random phases or random coordinates. In contrast, SHELXD ***(32)*** and HySS ***(21)*** use Patterson seeding to obtain better than random starting phases for the recycling procedure. The fundamental steps in the procedure are:

1. Generation of two-atom fragments. A given number of peaks are picked from a sharpened Patterson map. These are considered to be possible vectors between two

atoms of the substructure. However, at this stage only the relative orientation of the two atoms is known, but not the absolute position in the unit cell. SHELXD uses the Nordman *(18)* function to obtain scores for a large number of random translations of the two-atom fragments. HySS uses the fast translation function *(19)* to systematically score all translations on a uniform grid.

2. Extrapolation to the full substructure. Conceptually a third probe atom is systematically placed on the points of a uniform grid over the asymmetric unit while keeping a trial two-atom fragment fixed at a position that leads to a high score. SHELXD uses the Nordman function to compute scores for the position of the third atom. HySS uses the fast translation function for the same purpose. Points with the highest scores are added to the original two-atom fragment to generate the expected number of atoms.
3. Correction of defects. Typically, the structures obtained in the previous step contain a considerable number of misplaced atoms. Even the best solutions often have less than half of the atoms correctly placed. These defects are efficiently corrected using dual-space recycling (tangent formula expansions and random omission of peaks). The standard linear correlation coefficient between calculated and observed intensities is a very reliable score for ranking the final results of the dual-space recycling procedure.

2.4. Identification of Correct Solutions

Most search procedures including shake-and-bake and SHELXD are run for a preset number of trials, or until they are terminated manually. Various figures of merit, such as the Patterson figure of merit to measure the consistency between calculated and observed Patterson maps *(33)*, the value of the minimal function for the final set of phases *(29)*, or the correlation coefficient between observed and calculated E-values, can then be used to compare the results of different trials. However, a general problem with these figures of merit is that an absolute threshold for the discrimination between correct and incorrect solution cannot be given. Some searches yield trimodal distributions so that simply looking for outstanding figures of merit can also be misleading. Thus, additional criteria have to be employed. One such criterion is the presence of noncrystallographic symmetry in the substructure. This can be checked, for example, by using the "Patterson crossword table" *(33)* as implemented in SHELXD, by using a sophisticated combination of scores *(34)* or simply by inspecting the substructure on a graphical display. Another useful criterion for the correctness of a substructure that is suitable for automation is the consistency of substructures obtained from different trials. Therefore, the shake-and-bake suite *(35)* and recently the SHELX suite *(36)* include programs for comparing substructures that automatically take allowed origin shifts and change-of-hand operators into account. These programs are run externally from the actual search procedure. In HySS a similar procedure is fully integrated into the substructure

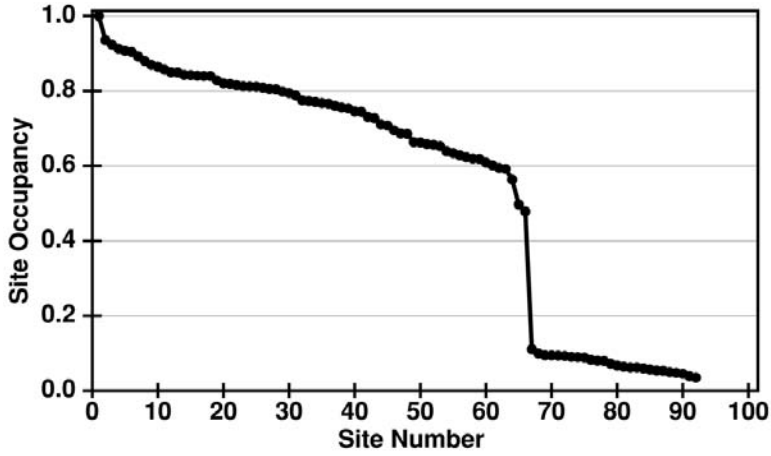


Fig. 3. Occupancy vs site number for a typical solution obtained for the Se-substructure of 2-aminoethylphosphonate transaminase *(58)* against FA data produced by XPREP. The occupancy of site 66 is 0.48 and the occupancy of site 67 is 0.11, clearly indicating that the substructure contains 66 sites.

search and is used in combination with the correlation coefficients to automatically terminate a search *(21)*.

Often it is not clear how many well-ordered substructure sites to expect exactly, in particular when working with heavy-atom soaks (*see* **Notes 9** and **10**). It is also quite common that a certain fraction of selenomethionine residues is not well ordered, for example residues near the termini. The SHELXD program employs refinement of occupancy factors to assist in determining the number of ordered sites. At the end of each trial a certain number of the highest peaks, typically 1.3 times the number of expected sites, are picked from the last E-map obtained in the dual-space recycling procedure. The occupancies of these sites are then refined against all observed substructure structure factors and the sites are sorted according to the refined values of their occupancies. In favorable cases the plot of the occupancy vs site number shows a sharp drop, indicating where the list of well-ordered sites ends (**Fig. 3**). Only sites that appear before this point should be used in subsequent steps.

Occupancy refinement is also used in the HySS program, but only after the determination of the sites common to the top two or three solutions (the consensus model), ranked by the correlation coefficient. The consensus model may contain more sites than expected because the number of peaks picked from the last E-map depends on the number of expected sites N. Typically, for $N < 30$ between N and 1.5 N peaks are picked. All these sites are subject to positional and occupancy refinement, and the resulting model sorted by occupancies in

decreasing order. Only the top N sites are retained. If this means that the model was truncated, it is refined a second time. Experience shows that the refined, truncated consensus model rarely contains a significant number of incorrect sites, if any.

2.5. Substructure Refinement and Completion

All programs currently in use for the determination of substructures will produce a model for the substructure that has the best agreement with a given set of substructure structure factors. However, for the determination of the best phases on which the initial electron density map for the complete structure will be based, it may be advantageous to use a substructure model that is consistent with all experimental data. For instance, in MAD phasing, the substructure should not only be in agreement with the F_A values, but more importantly should be consistent with the two or more original datasets collected at different wavelengths.

The process of refinement and completion of a substructure model against all available data in many aspects resembles the refinement of an entire macromolecular against native data: parameters of the model are optimized such that the agreement with the experimental data is improved and as the model becomes better, difference electron density maps reveal previously invisible parts of the structure to be included into the model. There are, however, a number of aspects that make substructure refinement substantially more complex than "native" structure refinement. As substructures contain significantly fewer atoms than structures of entire proteins, the data-to-parameter ratio is much higher, supporting the refinement of more parameters per site. Commonly refined parameters not only include coordinates and B-values, but also anisotropic B-values, occupancies, and dispersive and anomalous contributions to the scattering factor. Some of these parameters are highly correlated, e.g., the B-value and the occupancy of a site, and appropriate constraints and restraints have to be introduced to obtain reasonable results. Furthermore, the parameters have to be adjusted such that the agreement with several datasets is improved at the same time. This simultaneous use of several datasets introduces the problem of the relative scaling and weighting of the data from different sources.

When datasets from different crystals (or from the same crystals with different levels of radiation damage) are used, a further complication is introduced by the fact that structural differences between the different crystals, the so-called nonisomorphism, result in inconsistencies ("lack-of-closure") in the phasing calculations that are not trivial to handle. The difficulty of relative weighting of different sources of information also becomes apparent in the calculation of intermediate electron density maps to be used for completing the substructure model and a variety of sophisticated schemes have been suggested, one prominent example being the log-likelihood gradient map *(37)*.

Several programs, mostly based on maximum-likelihood principles, are available for carrying out heavy-atom refinement and phasing. These include CNS *(38)*, MLPHARE *(39,40)*, PHASES *(41)*, SHARP *(37)*, and SOLVE *(34)*.

With currently available experimental technology, data of very high quality can be collected (*see* **Note 11**) and in favorable cases it can be afforded to be less strict in the treatment of experimental errors for the sake of simplicity and speed. The recently implemented combination of SHELXD and SHELXE follows this route: based on the unmodified substructure model found by SHELXD, SHELXE rapidly calculates an initial phase set that is subsequently subjected to density modification, often yielding an interpretable electron density map within minutes.

3. Summary of Programs Available

ACORN: http://www.ccp4.ac.uk/. Primarily designed for solving protein structures at atomic resolution (1.2 Å or better), but tests of substructure determinations were also reported *(42)*. Starts with a known substructure, known fragments positioned with an extensive molecular replacement search, or randomly placed atoms. This is followed by application of Patterson superposition methods, Sayre phase refinement, and a special density modification procedure (dynamic density modification).

CNS heavy_search.inp: http://cns.csb.yale.edu/. Evaluation of direct-space and reciprocal-space Patterson interpretation, combined with Patterson correlation refinement. A robust but relatively slow method that has proven most useful for solving moderately sized structures from noisy data *(43)*, but it has also been tested successfully to solve structures with up to 66 sites *(44)*.

HySS: http://phenix-online.org/. Dual-space recycling with Patterson seeding. A highly automated procedure with a sophisticated, fully integrated validation mechanism that enables automatic termination as soon as the solution is clear. Particularly easy to use given anomalous diffraction data because no input files other than the reflection file are required and all common reflection file formats are processed directly (this includes files produced by XPREP.

RANTAN: http://www.ccp4.ac.uk/. Tangent refinement initialized with random phases. Used by autoSHARP (http://www.globalphasing.com/sharp/) (*see* Chapter 12).

Shake-and-Bake: http://www.hwi.buffalo.edu/SnB/. Dual-space recycling, alternatively using the tangent formula or the minimal function *(45)*. The complete *shake-and-bake* suite includes sophisticated procedures for the determination of normalized difference structure factors *(46)*, graphical visualization tools for monitoring the progress of the search, and a procedure for substructure comparisons *(35)*.

SHELX: http://shelx.uni-ac.gwdg.de/. The SHELX-suite of programs for macromolecular phasing includes modules for the determination of substructure structure factors (SHELXC), substructure solution (SHELXD), phase calculation, and density modification (SHELXE). The modules can be used individually or in an integrated manner via a graphical user interface (HKL2MAP).

SOLVE & RESOLVE: http://solve.lanl.gov/. SOLVE implements a tight integration of direct-space Patterson methods, difference Fourier analysis, and phasing *(34)*. One or two initial substructure sites are determined with Patterson superposition functions. The remaining sites are found by repeated analysis of isomorphous or anomalous difference Fourier maps. These fundamental building blocks are integrated into a high-level procedure that automates decision-making using a sophisticated scoring system. Includes all steps including the refinement of experimental phases. The RESOLVE program *(47)* implements statistical density modification and automated model building.

XPREP: http://www.bruker-axs.com/. Program for the evaluation of diffraction data. Among other features, it allows the estimation of substructure structures from various types of experiments, such as MAD, SAD, MIR, SIRAS, and others. The program contains procedures for detwinning anomalous data.

4. Notes

1. Careful evaluation of data quality is crucial. The structure factors representing a substructure are calculated as small differences between large experimentally measured values. Therefore, the importance of evaluating the data quality before starting the substructure solution process cannot be overstated. If measurements of anomalous differences at different wavelengths are available, truncating the data at the resolution where the correlation coefficient between anomalous differences falls below 30% is a good strategy (**Fig. 1A**). Including data of lower quality may be harmful and in extreme cases the inclusion of a 0.1-Å thick high-resolution shell can lead to failure of the substructure solution process (**Fig. 1B**). Another measure of data quality is the signal to noise for the substructure factors. However, for this measure absolute thresholds are difficult to define because the estimation of accurate standard deviations for substructure structure factors is difficult.

   In a single-wavelength experiment it may be advantageous to carry out additional steps in order to obtain two or more independent measurements of anomalous differences. For example, the crystal can be rotated around different axes on a multicircle diffractometer. A very rigorous approach is to repeat the same experiment on two different crystals and then to correlate the measurements (e.g., **ref.** *48*).
2. Some programs provide the facilities to use unmerged data (e.g., SOLVE and XPREP). It is often advantageous to provide unmerged data to these programs. In comparison to premerged data, unmerged data allow for a more robust statistical analysis. A major reason is that uncertainties can be estimated from a spread of measurements and outliers can be detected more reliably.

3. Collecting highly redundant data has been repeatedly shown to increase the chances of success *(10,49)*. In addition, higher redundancy results in more reliable statistics.
4. Radiation damage is the major obstacle to collecting data to high redundancy. If the crystal is heavily affected by radiation damage, a low-dose, high-redundancy dataset that extends to 3 Å can be much more useful for initial phasing than a high-dose, low-redundancy dataset that extends to 2 Å.
5. Radiation damage can be used as a phasing tool. The specific structural changes induced in a crystal structure when deliberately exposed to large doses of X-ray photons can be exploited for experimental phasing. Ravelli et al. *(50)* have successfully used a dataset from a fresh crystal as the native dataset, and the dataset from a "burnt" crystal as the derivative dataset in a pseudo isomorphous replacement experiment.
6. Collecting related reflections close in space and time (*see* Chapter 5) can make the difference between success and failure if radiation damage is a severe problem (e.g., **ref.** *51*).
7. Merohedral twinning *(52)* creates fewer problems than one would expect for the determination of the substructure as: (1) the strong reflections used in dual-space recycling procedures are statistically less affected by the twinning than general reflections, and (2) structure solution using Patterson methods against twinned data will produce only one twin component as no vectors between the twin component are present in the Patterson map *(53)*.
8. Treating a failed MAD experiment as a SAD experiment may be a viable approach. Sometimes, the systematic errors between the different wavelengths of a MAD experiment are too large to give useful results. Using selected single-wavelength data in a SAD approach may nonetheless solve the structure *(43)*.
9. The simultaneous presence of different kinds of anomalous scatterers can lead to complications because the wavelength-dependent change of the anomalous signal is very specific to each element type. The available substructure determination programs may not function properly because a number of commonly used approximations are based on the assumption that only one kind of anomalous scatterer is present.
10. If the number of substructure sites is not known, e.g., in the case of halide soaks (*see* Chapter 8), it is generally better not to overestimate the expected number sites, but to err on the low side. If too many spurious sites are included in the calculations most figures of merit cease to be good indicators. A rule of thumb could be "less than number of residues divided by 20," but often multiple different values have to be tried.
11. MAD phasing at atomic resolution is possible. Recent work on endoglucanase A *(54)* and aldose reductase *(55)* with experimental phases extending to 1.0 and 0.9 Å resolution, respectively, has shown that a well-refined heavy-atom substructure can provide the basis for the calculation of excellent phases to atomic resolution. In favorable cases the maps are virtually indistinguishable from refined maps.

Acknowledgments

We thank P. Adams for suggestions and for critically reading this manuscript. Work by RWGK was funded in part by the US Department of Energy under contract no. DE-AC03-76SF00098 and by National Institutes of Health

(NIH)/NIGMS under grant number 1P01GM063210. Work by TRS was funded in part by the European Union (QLRI-CT-2000-00398).

References

1. Hendrickson, W. A. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254,** 51–58**.**
2. Blundell, T. L. and Johnson, L. N. (1976) *Protein Crystallography,* Academic Press, London, UK.
3. McPherson, A. J. (1982) *The Preparation and Analysis of Protein Crystals,* Wiley, New York, NY.
4. Petsko, G. A. (1985) Preparation of isomorphous heavy-atom derivatives. *Meth. Enzym.* **114,** 147–156.
5. Carvin, D., Islam, S. A., Sternberg, M. J. E., and Blundell, T. L. (2001) The preparation of heavy atom derivatives of protein crystals for use in multiple isomorphous replacement. In: *International Tables for Crystallography: Crystallography of Biological Macromolecules, Vol. F,* (Rossman, M. G. and Arnold, E., eds.), Kluwer Academic Publishers, Dordrecht, Germany, pp. 247–255.
6. Dauter, Z. (2002) New approaches to high-throughput phasing. *Curr. Opin. Struct. Biol.* **12,** 674–678.
7. Doublié, S. (1997) Preparation of selenomethionyl proteins for phase feterminations. *Meth. Enzym.* **276,** 523–530.
8. Hendrickson, W. A. and Teeter, M. M. (1981) Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. *Nature (London)* **290,** 107–113.
9. Dauter, Z., Dauter, M., De La Fortelle, E., Bricogne, G., and Sheldrick, G. M. (1999) Can anomalous signal of sulfur become a tool for solving protein crystal structures? *J. Mol. Biol.* **289,** 83–92.
10. Dauter, Z. and Adamiak, D. A. (2001) Anomalous signal of phosphorus used for phasing DNA oligomer: importance of data redundancy. *Acta Cryst.* **D57,** 990–995.
11. Lemke, C. T., Smith, G. D., and Howell, P. L. (2002) S-SAD, Se-SAD and S/Se-SIRAS using Cu Kα radiation: why wait for synchrotron time? *Acta Cryst.* **D58,** 2096–2101.
12. Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986) *Numerical Recipes. The Art of Scientific Computing.* Cambridge University Press, Cambridge, UK.
13. Fan, H. F., Woolfson, M. M., and Yao, J. X. (1993) New techniques of applying multi-wavelength anomalous scattering data. *Proc. R. Soc. Lond*. **A442,** 13–32.
14. Terwilliger, T. C. (1994) MAD phasing: Bayesian estimates of FA. *Acta Cryst.* **D50,** 11–16.
15. Dauter, Z. (2002) One-and-a-half wavelength approach. *Acta Cryst*. 1958–1967.
16. Patterson, A. L. (1935) A direct method for the determination of the components of interatomic distances in crystals. *Z. Krist*. **90,** 517–542.
17. Buerger, M. J. (1959) *Vector Space,* Wiley, New York, NY.

18. Nordman, C. E. (1966) Vector space search and refinement procedures. *Trans. Am. Crystallogr. Assoc.* **2,** 29–38.
19. Navaza, J. and Vernoslova, E. (1995) On the fast translation functions for molecular replacement. *Acta Cryst.* **A51,** 445–449.
20. Grosse-Kunstleve, R. W. and Brunger, A. T. (1999) A highly automated heavy-atom search procedure for macromolecular structures. *Acta Cryst.* **D55,** 1568–1577.
21. Grosse-Kunstleve, R. W. and Adams, P. D. (2003) Substructure search procedures for macromolecular structures. *Acta Cryst.* **D59,** 1966–1973.
22. Karle, J. and Hauptman, H. A. (1956) A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22. *Acta Cryst.* **9,** 635–651.
23. Giacovazzo, C. (1992) *Fundamentals of Crystallography.* IUCr/Oxford University Press, Oxford, UK.
24. Sayre, D. (1952) The squaring method: a new method for phase determination. *Acta Cryst.* **5,** 60–65.
25. Woolfson, M. M. (1961) *Direct Methods in Crystallography.* Oxford University Press, Oxford, UK.
26. Germain, G., Main, P., and Woolfson, M. M. (1970) On the application of phase relationships to complex structures. II. Getting a good start. *Acta Cryst.* **B26,** 274–285.
27. Yao, J. X. (1981) On the application of phase relationships to complex structures XVIII. RANTAN-random MULTAN. *Acta Cryst.* **A37,** 642–644.
28. Sheldrick, G. M. (1985) *SHELXS86. Program for the Solution of Crystal Structures.* University of Göttingen, Göttingen, Germany.
29. Miller, R., Gallo, S. M., Khalak, H. G., and Weeks, C. M. (1994) SnB: crystal structure determination via shake-and-bake. *J. Appl. Cryst.* **27,** 613–621.
30. Sheldrick, G. M. and Gould, R. O. (1995) Structure solution by iterative peaklist optimization and tangent expansion in space group P1. *Acta Cryst.* **B51,** 423–431.
31. Sheldrick, G. M., Hauptman, H.A., Weeks, C.M., Miller, R., and Uson, I. (2001) Ab initio phasing. In: *International Tables for Crystallography. Crystallography of Biological Macromolecules, Vol. F,* (Rossman, M. G. and Arnold, E., eds.), Kluwer Academic Publishers, Dordrecht, Germany, pp. 233–245.
32. Schneider, T. R. and Sheldrick, G. M. (2002) Substructure solution with SHELXD. *Acta Cryst.* **D58,** 1772–1779.
33. Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H., and Sieker, L. C. (1993) The application of direct methods and Patterson interpretation to high-resolution native protein data. *Acta Cryst.* **D49,** 18–23.
34. Terwilliger, T. C. and Berendzen, J. (1999) Automated MAD and MIR structure solution. *Acta Cryst.* **D55,** 849–861.
35. Smith, G. D. (2002) Matching selenium-atom peak positions with a different hand or origin. *J. Appl. Cryst.* **35,** 368–370.
36. Dall'Antonia, F., Baker, P. J., and Schneider, T. R. (2003) Optimization of selenium substructures as obtained from SHELXD. *Acta Cryst.* **D59,** 1987–1994.

37. De La Fortelle, E. and Bricogne, G. (1997) Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Meth. Enzym.* **276,** 472–494.
38. Brunger, A. T., Adams, P. D., Clore, G. M., et al. (1998) Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta Cryst.* **D54,** 905–921.
39. Collaborative Computational Project, N. (1994) The CCP4 suite: programs for protein crystallography. *Acta Cryst.* **D50,** 760–763.
40. Otwinowski, Z. (1991) Maximum likelihood refinement of heavy atom parameters. In: *Isomorphous Replacement and Anomalous Scattering, Proc. Daresbury Study Weekend,* SERC Daresbury Laboratory, Warrington, UK, pp. 80–85.
41. Furey, W. and Swaminathan, S. (1997) PHASES-95: a program package for processing and analyzing diffraction data from macromolecules. *Meth. Enzym.* **277,** 590–620.
42. Yao, J. X. (2002) ACORN in CCP4 and its applications. *Acta Cryst.* **D58,** 1941–1947.
43. Rice, L. M. and Brunger, A. T. (1999) Crystal structure of the vesicular transport protein sec17: implications for snap-mediated snare complex disassembly. *Mol. Cell* **4,** 85–96.
44. Weeks, C. M., Adams, P. D., Berendzen, J., et al. (2003) Automatic solution of heavy-atom substructures. *Meth. Enzym.* **374,** 37–83.
45. DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R., and Hauptman, H. A. (1994) Structure solution by minimal function phase refinement and Fourier filtering: theoretical basis. *Acta Cryst.* **A50,** 203–210.
46. Blessing, R. H. and Smith, G. D. (1999) Difference structure factor normalization for heavy-atom or anomalous-scattering substructure determinations. *J. Appl. Cryst.* **32,** 664–670.
47. Terwilliger, T. C. (2000) Maximum-likelihood density modification. *Acta Cryst.* **D56,** 965–972.
48. Cordell, S. C., Anderson, R. E., and Löwe, J. (2001) Crystal structure of the bacterial cell division inhibitor MinC. *Embo J.* **20,** 2454–2461.
49. Debreczeni, J. E., Bunkóczi, G., Ma, Q., Balser, H., and Sheldrick, G. M. (2003) In-house measurement of the sulfur anomalous signal and its use for phasing. *Acta Cryst.* **D59,** 688–696.
50. Ravelli, R. B., Schroeder Leiros, H. K., Pan, B., Caffrey, B., and McSweeney, S. (2003) Specific radiation damage can be used to solve macromolecular crystal structures. *Structure* **11,** 217–224.
51. Clemons, W. M. J., Brodersen, D. E., McCutcheon, J. P., et al. (2001) Crystal structure of the 30S ribosomal subunit from Thermus thermophilus: purification, crystallization and structure determination. *J. Mol. Biol.* **310,** 827–843.
52. Yeates, T. O. (1997) Detecting and overcoming crystal twinning. *Meth. Enzym.* **276,** 344–360.
53. Rudolph, M. G., Kelker, M. S., Schneider, T. R., et al. (2003) Use of multiple anomalous dispersion to phase highly merohedrally twinned crystals of interleukin-1b. *Acta Cryst.* **D59,** 290–298.

54. Schmidt, A., Gonzalez, A., Morris, R. J., Costabel, M., Alzari, P. M., and Lamzin, V. S. (2002) Advantages of high-resolution phasing: MAD to atomic resolution. *Acta Cryst.* **D58,** 1433–1441.
55. Podjarny, A., Schneider, T. R., Cachau, R. E., and Joachimiak, A. (2003) Structural information content at high resolution: MAD versus native. *Meth. Enzym.* **374,** 321–341.
56. Hensgens, C. M., Kroezinga, E. A., van Montfort, B. A., van der Laan, J. M., Sutherland, J. D., and Dijkstra, B. W. P. (2002) Purification, crystallization and preliminary X-ray diffraction of Cys103Ala acyl coenzyme A:isopenicillin N acyltransferase from Penicillium chrysogenum. *Acta Cryst.* **D58,** 716–718.
57. Dall'Antonia, F., Baker, P. J. and Schneider, T. R. (2003) Optimization of selenium substructures as obtained from SHELXD. *Acta Cryst.* **D59,** 1987–1994.
58. Chen, C. C., Zhang, H., Kim, A. D., et al. (2002) Degradation pathway of the phosphonate ciliatine: crystal structure of 2-aminoethylphosphonate transaminase. *Biochemistry* **41,** 13,162–13,169.

12

Automated Structure Solution With autoSHARP

Clemens Vonrhein, Eric Blanc, Pietro Roversi, and Gérard Bricogne

Summary

We present here the automated structure solution pipeline "autoSHARP." It is built around the heavy-atom refinement and phasing program SHARP, the density modification program SOLOMON, and the ARP/wARP package for automated model building and refinement (using REFMAC). It allows fully automated structure solution, from merged reflection data to an initial model, without any user intervention. We describe and discuss the preparation of the user input, the data flow through the pipeline, and the various results obtained throughout the procedure.

Key Words: SHARP; autoSHARP; automation; structure solution.

1. Introduction

Automation of crystal structure determination, from processed data to an initial macromolecular model, is desirable in two areas: in the high-throughput structural genomics efforts, and in making the power of today's most sophisticated methods accessible to the structural biologist without much crystallographic experience. This requires user-friendliness both in defining the experimental phasing protocol and in presenting and commenting the results, and also in the tracking of work flow. Here we present a system (autoSHARP) that tries to accommodate the novice user as well as the expert.

2. Materials

The automatic structure solution pipeline autoSHARP is a set of computer programs and scripts designed to run on a variety of modern, UNIX-like computers. The main autoSHARP tasks are written in Bourne shell *(1)* and Perl *(2)* with some helper applications used as system-specific binaries. Further details on requirements and installation instructions can be found at our website http://www.globalphasing.com/sharp/. The development of autoSHARP started

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

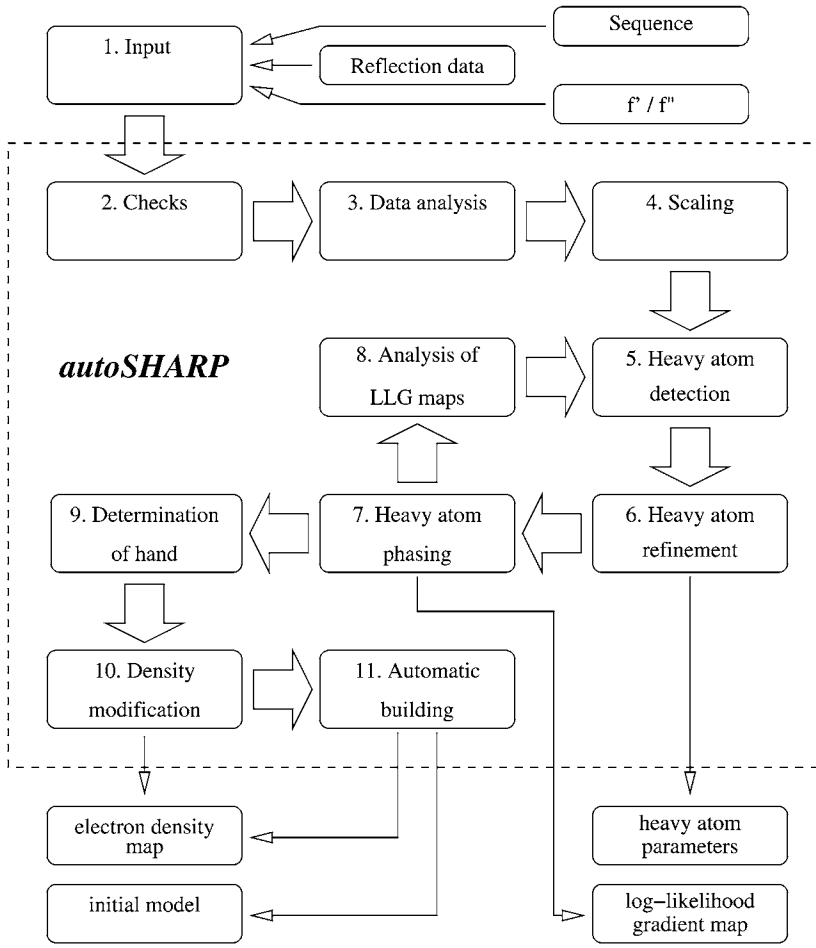


Fig. 1. Flowchart of autoSHARP procedure.

in 2000, with the first public release in January 2002 and an active user base with more than 1500 installations at the date of writing.

3. Methods

The methods described next outline the various steps for successfully running the automatic structure solution pipeline in autoSHARP. An on-line manual is also available at: http://www.globalphasing.com/sharp/. The steps include (1) preparation of the input data, (2) steps performed by the program, (3) analysis of the program output, and (4) presentation of results. The logic follows closely the structure solution process as shown in the flowchart of **Fig. 1.**

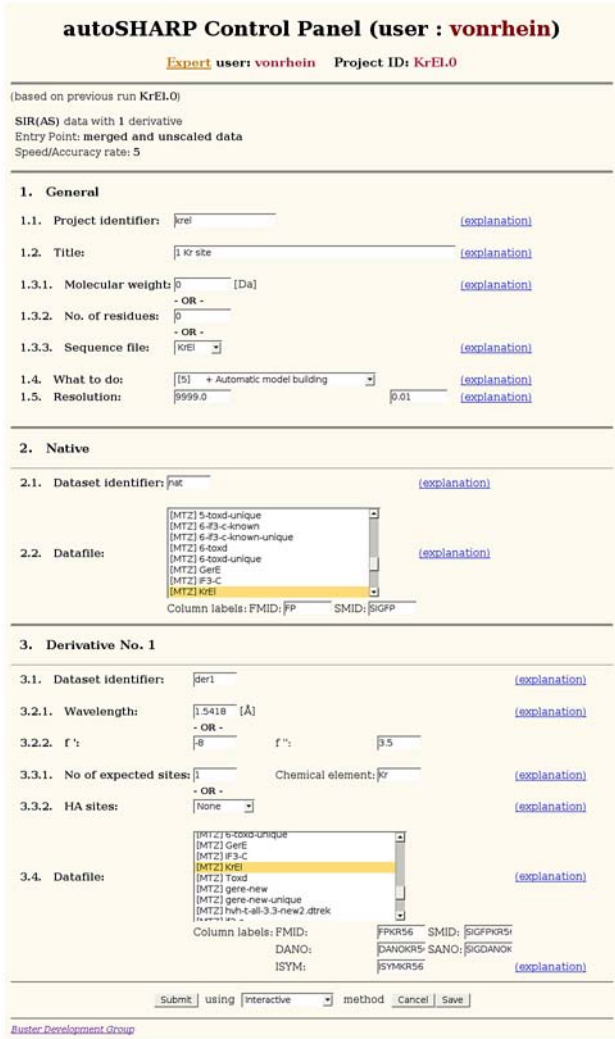


Fig. 2. User input form for SIR(AS) experiment.

3.1. Input

In order to start autoSHARP some mandatory and some optional information has to be provided for each wavelength in a SAD/MAD experiment, or for each derivative in a SIR(AS)/MIR(AS) experiment (*see* **Fig. 2**). Although autoSHARP tries to apply sensible defaults whenever it encounters a missing input value, the automatic structure solution process (and the decisions that

need to be taken within this process) will be only as good as the starting information given to the program.

3.1.1. Mandatory Information

This information needs to be supplied or given as part of the file header for reflection data.

3.1.1.1. Heavy-Atom Type

Usually, only one type of heavy atom is present in the crystal (e.g., selenium [Se] or mercury [Hg]). In some cases, however, a metal ion might be bound to the macromolecule (e.g., calcium [Ca^{2+}]) or a cofactor contains a heavy atom (e.g., a heme group with an iron [Fe] atom). Although autoSHARP can handle a list of different heavy atoms it is usually enough to define the heavy atom that will be most visible during the specific experiment (when collecting the peak wavelength of a Se-MAD experiment this will be the Se atoms—even if in final maps the sulfur atoms of cysteine residues might be visible through their weak contribution to anomalous scattering) (*see* **Note 1**).

3.1.1.2. Number of Expected Heavy Atoms

The exact number of heavy atoms bound to the macromolecule is usually not known *a priori*—unless the number of molecules in the asymmetric unit is known and the heavy atom is an intrinsic part of the macromolecule (selenomethionine [Se-Met], sulfurs of Met and/or Cys residues, metals of cofactors, and so on). In general, autoSHARP will dynamically adjust the number of heavy atoms actually used for phasing—up to a maximum limit given by the user. Therefore, overestimating the number of heavy atoms usually has no negative effect on the performance (*see* **Note 2**).

3.1.1.3. Space Group

This will be picked up automatically from the header of the reflection file. However, quite often data are processed in a space group without any of the possible screw axes (e.g., P222 instead of $P2_12_12_1$), in order to check for systematic absences along the various axes. The data files going into autoSHARP should have the correct space group in their header. If no data were collected along a potential screw axis, autoSHARP should probably be run with and without the screw-axis component. The ambiguity in handedness of a screw axis (4_1 vs 4_3) will be taken into account automatically and does not need to be checked.

3.1.1.4. Wavelength

In order to define the scattering properties of the heavy atom for a given experiment, the various wavelengths at which data were collected need to be

known. They will be used for calculating f′ and f″ values for all heavy atoms declared (*see* Note 3).

3.1.1.5. Reflection Data

The MTZ *(3)* and SCALEPACK *(4)* reflection file formats are supported. Other formats can easily be converted into one of these widely used ones (*see* **Note 4**).

3.1.2. Optional Input

Additionally, the following information is usually available, and giving it to autoSHARP will improve its efficiency.

3.1.2.1. Content of Asymmetric Unit

This can be given either as molecular weight, number of residues, or in the form of a sequence file. If the asymmetric unit can accommodate more than one copy of the macromolecule, but the exact number is not yet known, it is recommended to specify this sequence for a single molecule only. The use of a sequence file is preferred, because all other parameters can be calculated from this (*see* **Note 5**).

3.1.2.2. f′ and f″ Values

Although these can be calculated from the knowledge of the heavy atom's chemical type and the wavelength *(5)*, the actual value observed during the experiment can vary substantially from the theoretical value when collecting data close to the absorption edge of an element. This may be caused by the specific environment of the anomalous scatterer, and also by the spread of wavelengths around its mean value in the X-ray beam. Therefore, it is always advisable to perform a fluorescence scan on the same (or very similar) crystal as was used for data collection (*see* **Note 6**).

3.2. Checks

The first step within the autoSHARP procedure is to check the user input for syntax errors. Any mandatory input parameter that is missing will be reported and will result in the program halting after all other checks have been performed.

3.3. Data Analysis

During the next few steps, the program performs as much data analysis as possible.

3.3.1. Crystal-Independent Analysis

The autoSHARP scripts carry out the following tasks:

1. Calculation of molecular weight and the number of residues based on a sequence file (if given).

2. Calculation of f′ and f″ values (in case only the wavelength was given) using the program CROSSEC *(3)*.
3. Analysis of f′ and f″ values for a MAD experiment to automatically determine which of the datasets belongs to "peak," "inflection," "high-energy remote," or "low-energy remote."
4. In case of sulfur or Se-Met phasing, a comparison between the specified number of heavy atoms and the number of Cys/Met residues in the sequence file is done.

3.3.2. Crystal- and Dataset-Dependent Analysis

For each crystal and each dataset pertaining to that crystal, the following tasks are performed:

1. Extraction of cell parameters, space group name (and number) from the data files, and a test for consistency of these values between datasets.
2. Determination of resolution limits of each dataset, as well as overall and common resolution limits.
3. Estimation of overall temperature factor by using the least-squares line fit for a Wilson plot *(6)* (*see* **Note 7**).
4. Estimation of the most likely number of monomers per asymmetric unit, as well as the possible range for this number using the Matthews coefficient *(7)*.

3.3.3. Dataset-Dependent Analysis

For each dataset, the autoSHARP scripts carry out the following analysis:

1. Determination of overall and anomalous completeness per dataset (*see* **Note 8**).
2. Detection and removal of anomalous difference outliers. By analyzing the value of $|\Delta^{ano}|/F$, those reflections with unusually large values are reported. Reflection where this value is larger than 1.9 are also removed from the dataset. Unless a very strong anomalous signal is expected these nearly always point back to problems during data processing (*see* **Note 9**).
3. Calculation of self-rotation functions *(8)*, assuming a globular shape of the macromolecule and a radius based on the size of the monomer. These are plotted with the program POLARRFN *(3)* to help define possible noncrystallographic symmetry (NCS). This NCS information is at the moment not actively used within autoSHARP, but it will be used in future versions to impose constraints on the heavy-atom substructure or to perform real-space averaging during density modification.
4. Calculation and analysis of the native Patterson function *(9,10)*. This can bring to light the presence of purely translational NCS, an extreme case of which could lead to a mistake in the assignment of the correct space group where a centering operation has been missed.

3.4. Scaling

Apart from SAD experiments, other types of structure solution protocols (SIR[AS], MIR[AS], MAD) require more than one dataset. These need to be scaled relative to each other, which will give additional information such as:

1. Outlier analysis based on normalized structure factors. Reflections for which the normalized structure factor (E-value) in different datasets differs by more than five are reported. Furthermore, reflections for which two values differ by more than 10 are deleted from the dataset (*see* **Note 10**).
2. Outlier analysis of isomorphous/dispersive and anomalous differences. During the scaling of two datasets using either Kraut's method *(11)* as implemented in FHSCAL *(3)* or a simple anisotropic scaling model as implemented in SCALEIT *(3)*, reflections with unlikely large differences are detected and an appropriate cut-off value determined.
3. Cross-table of R-factor values between all dataset pairs within a common overall resolution range.
4. Table of correlation coefficients between Δ_{ano} values for all dataset pairs in case of a MAD experiment.

All of the previously listed statistics are used for determining a reasonable high-resolution limit for the heavy-atom detection step. Furthermore, in cases of significantly larger R-factors in the lowest resolution shell the low-resolution limit will also be restricted.

3.5. Heavy-Atom Detection

Several powerful programs for heavy-atom substructure detection are available—like SHELXD *(12)*, Shake-and-Bake (SnB) *(13)*, HySS *(14)*, SOLVE *(15)*, or CRUNCH2 *(16)*—the procedures for heavy-atom substructure solution currently implemented in autoSHARP use either SHELXD or RANTAN *(17)*. However, solutions from other programs can be input through a simple file, in either Protein Data Bank (PDB) *(18)* or in our internal fractional coordinates format.

The heavy-atom substructure for a given dataset (or a collection of datasets in the case of MAD) can be determined from a variety of data:

1. Anomalous differences measured for each dataset.
2. Isomorphous differences between a derivative and a native dataset.
3. Dispersive differences between different wavelengths of a MAD experiment, especially between inflection and remote wavelengths.

The strategy for finding a substructure solution is:

1. When using RANTAN: each type of difference is converted into normalized differences using the program ECALC *(3)* and is given to RANTAN, which then generates three sets of phases that are most consistent with the data. Using these phases and the input E-values, a real-space map can be calculated which should show peaks at the heavy-atom positions. For further checking, the list of peaks is analyzed for consistency with an origin-removed Patterson function (calculated using $[E^2-1]$ coefficients based on the same E-values as are input into RANTAN). Furthermore, any peaks on special positions or in unusual relation to each other (e.g., several peaks differing only in the coordinate along a polar axis) are removed. The resulting list is sorted by peak height and used for further analysis.

For MAD experiments, the optimized value of normalized anomalous scattering as calculated by the program REVISE *(19)* can be used in the same manner for the determination of the heavy-atom substructure.

Because any solution from the direct methods program RANTAN is automatically checked against the corresponding difference Patterson map (anomalous, isomorphous, or dispersive), the space group-specific Harker sections of this Patterson map are plotted as well. With experience it is very easy to judge the quality of the signal visually from these plots.

2. When using SHELXD: different types of protocols (MAD, SIR[AS], SAD) are used to generate a set of F_A-values using SHELXC *(20)*.

   To not only assess the correctness of a substructure solution, but also to allow for possible errors in the initial estimation of the number of heavy atoms, the top N peaks of this list are used in a stepwise manner (where the cut-off value on peak height is slowly lowered to include more and more lower scoring peaks in the analysis). These peaks are converted into atoms of the given heavy atom type, with a temperature factor based on the Wilson plot (determined in **Subheading 3.3.2.**) and an occupancy dependent on both the type of experiment and the peak height (in case of a Se- or S-MAD/SAD experiments all atoms will have the same occupancy). A correlation coefficient between the (E^2–1) values (for RANTAN) or the F_A-values (for SHELXD) of the observed data and the heavy-atom structure factor amplitudes calculated from this set of atoms is used as a score to judge the quality of the substructure solution. This correlation coefficient can either use all reflections or only a test set of reflections, which were not used for detecting the substructure (similar to the free R-value as introduced by **ref.** *21*).

   In case of a MIR(AS) experiment, the various lists of heavy-atom sites for each derivative cannot be used together directly, because each substructure may be defined with a different origin and enantiomorph. Therefore, the solution with the best correlation coefficient is used as a starting point (effectively starting as a SIR[AS] experiment).

3.6. Heavy-Atom Refinement

Once an initial set of heavy atoms is available, it is given to SHARP *(22)* for refinement. SHARP will first estimate the absolute scale (based on the analysis of asymmetric unit content done previously) so that the initial heavy-atom occupancies are on a meaningful scale. Also, an overall anisotropic temperature factor is estimated to accommodate any anisotropy that might be present in the data. If more than a single dataset is used, the scale and temperature factors relative to the first dataset (the first wavelength in a MAD experiment, or the native dataset in a SIR[AS] or MIR[AS] experiment) are estimated.

SHARP will then refine coordinates, occupancy, and temperature factors for the heavy-atom sites, together with scaling parameters and nonisomorphism parameters for each dataset. This refinement is done in a stepwise manner, first refining the most critical parameters, then adding more parameters until all are refined

together to convergence. Coefficients for a map of the final log-likelihood gradient (LLG) are calculated for further analysis of the current model (*see* **ref. *22***).

3.7. Heavy-Atom Phasing

Once the heavy-atom refinement in SHARP has converged, a set of phases and map coefficients is calculated. This also includes Hendrickson–Lattman coefficients ***(23)*** for subsequent phase combination in density modification programs. Furthermore, additional information about the two-dimensional phase probability for each reflection is output ***(24)***.

The main statistical descriptors regarding heavy-atom phasing are:

1. Mean phasing power in resolution shells for each derivative (separate for isomorphous/dispersive and anomalous differences).
2. Cullis R-factor ***(25)*** in resolution shells for each derivative (separate for isomorphous/dispersive and anomalous differences).
3. Mean figure-of-merit values within resolution shells (separate for centric and acentric reflections).

3.8. Analysis of LLG Maps

Once convergence has been reached during the refinement of a given heavy-atom model, SHARP produces maps for the calculated heavy-atom density and for the LLG corresponding to each dataset. The final heavy-atom parameters and the associated maps are then analyzed in order to:

1. Detect heavy-atom sites for which the refined parameters indicate a nonexistent site (i.e., very low occupancy, very large temperature factor, or no significant density in the calculated model map): these are deleted from the current model.
2. Detect new and additional heavy-atom sites, where the LLG maps have significant positive density within a minimum distance of already existing sites; these will be automatically added.
3. Detect cases where the anomalous LLG map shows significant positive or negative features directly at most of the current heavy atom positions: for these datasets the refinement of f″-values will be switched on.

The current heavy-atom model is then updated by deletion and/or addition of sites as well as update or refinement protocol, and another heavy-atom parameter refinement with SHARP is run (*see* **Note 1**). This is done automatically, but a detailed description of the decisions made is presented to the user, who has the option of overriding them through the graphical user interface.

3.9. Determination of Hand

Once no further update of the heavy-atom model seems necessary, a final set of phases is calculated in both hands. For this purpose, the original heavy-atom configuration is inverted (mostly around the origin accompanied by a switch to

the enantimorph space group—but for $I4_1$, $I4_122$, and $F4_132$ the inversion is around a different point without the need to change the space group). The two phase sets are then analyzed to determine which heavy-atom configuration (and space group) is correct.

The criterion for a correct set of phases is that the resulting electron density map has features that are consistent with a macromolecule. Because exactly the same criterion is used in density modification methods *(26)*, the latter can be used to decide on the correct hand. For that purpose, a single cycle of solvent flipping with SOLOMON *(27)* is performed for each phase set, and the resulting statistics are used to decide which is the correct hand. Two criteria can be used as a basis for this decision:

1. The value of the correlation coefficient between observed E^2 values and E^2 values based on the modified map (which should be higher for the correct hand).
2. The value of the ratio between the standard deviation in the solvent region (which should have a low value because this region contains mainly disordered solvent) and the standard deviation in the protein region (which should be high because of well-defined, positive features).

3.10. Density Modification

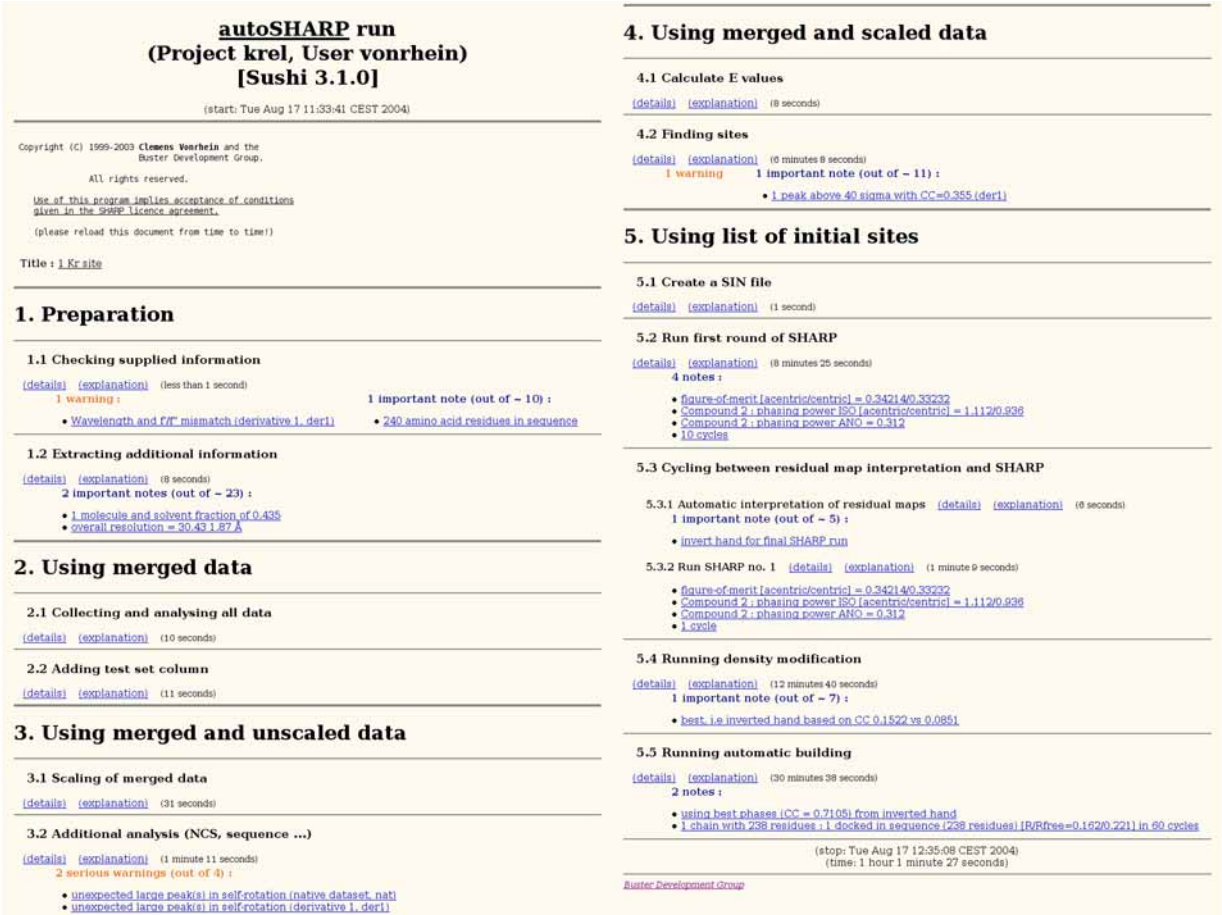
In order to have the best electron density map for building and analyzing the crystal structure, density modification procedures are used for phase improvement. Here, the solvent flipping method as implemented in SOLOMON *(27)* is applied—taking into account some specific features of the two-dimensional probability information coming from SHARP *(24)*. Because even at this stage the exact number of molecules per asymmetric unit might not be known precisely, a simple optimization of the solvent content (first stepwise and finally through a parabolic fit) is performed. The resulting map should show all molecules within the asymmetric unit, without the danger of "flattening" density for unexpected molecules or artificially including density of the disordered solvent in the protein region.

3.11. Automatic Building

The best density modified map is finally handed over to the ARP/wARP suite of programs for automatic model building *(28)*. The protocol used within this program suite takes into account the available experimental phase information from SHARP (in form of Hendrickson–Lattman coefficients) as well as the possible presence of heavy atoms in the macromolecule (in case of SAD or MAD experiments).

3.12. Results From autoSHARP

Results are produced by autoSHARP at various stages within the automated pipeline (*see* **Fig. 3**): if the data resolution is sufficient, a PDB file of a partially built model will often be produced.



autoSHARP run
(Project krel, User vonrhein)
[Sushi 3.1.0]

(start: Tue Aug 17 11:33:41 CEST 2004)

Copyright (C) 1999-2003 **Clemens Vonrhein** and the
Buster Development Group.

All rights reserved.

Use of this program implies acceptance of conditions
given in the SHARP licence agreement.

(please reload this document from time to time!)

Title : 1 Kr site

1. Preparation

1.1 Checking supplied information

(details) (explanation) (less than 1 second)

1 warning :

- Wavelength and f'/f" mismatch (derivative 1, der1)

1 important note (out of ~ 10) :

- 240 amino acid residues in sequence

1.2 Extracting additional information

(details) (explanation) (8 seconds)

2 important notes (out of ~ 23) :

- 1 molecule and solvent fraction of 0.435
- overall resolution = 30.43 1.87 Å

2. Using merged data

2.1 Collecting and analysing all data

(details) (explanation) (10 seconds)

2.2 Adding test set column

(details) (explanation) (11 seconds)

3. Using merged and unscaled data

3.1 Scaling of merged data

(details) (explanation) (31 seconds)

3.2 Additional analysis (NCS, sequence ...)

(details) (explanation) (1 minute 11 seconds)

2 serious warnings (out of 4) :

- unexpected large peak(s) in self-rotation (native dataset, nat)
- unexpected large peak(s) in self-rotation (derivative 1, der1)

4. Using merged and scaled data

4.1 Calculate E values

(details) (explanation) (8 seconds)

4.2 Finding sites

(details) (explanation) (6 minutes 8 seconds)

1 warning 1 important note (out of ~ 11) :

- 1 peak above 40 sigma with CC=0.355 (der1)

5. Using list of initial sites

5.1 Create a SIN file

(details) (explanation) (1 second)

5.2 Run first round of SHARP

(details) (explanation) (8 minutes 25 seconds)

4 notes :

- figure-of-merit [acentric/centric] = 0.34214/0.33232
- Compound 2 : phasing power ISO [acentric/centric] = 1.112/0.936
- Compound 2 : phasing power ANO = 0.312
- 10 cycles

5.3 Cycling between residual map interpretation and SHARP

5.3.1 Automatic interpretation of residual maps (details) (explanation) (6 seconds)

1 important note (out of ~ 5) :

- invert hand for final SHARP run

5.3.2 Run SHARP no. 1 (details) (explanation) (1 minute 9 seconds)

- figure-of-merit [acentric/centric] = 0.34214/0.33232
- Compound 2 : phasing power ISO [acentric/centric] = 1.112/0.936
- Compound 2 : phasing power ANO = 0.312
- 1 cycle

5.4 Running density modification

(details) (explanation) (12 minutes 40 seconds)

1 important note (out of ~ 7) :

- best, i.e inverted hand based on CC 0.1522 vs 0.0851

5.5 Running automatic building

(details) (explanation) (30 minutes 38 seconds)

2 notes :

- using best phases (CC = 0.7105) from inverted hand
- 1 chain with 238 residues : 1 docked in sequence (238 residues) [R/Rfree=0.162/0.221] in 60 cycles

(stop: Tue Aug 17 12:35:08 CEST 2004)
(time: 1 hour 1 minute 27 seconds)

Buster Development Group

Fig. 3. Main autoSHARP output file for SIR(AS) experiments.

The main categories of results are:

1. A refined heavy-atom model, comprising: (1) position, occupancy, and temperature factor of the heavy-atom sites, (2) scaling parameters (including anisotropic temperature factors), and (3) nonisomorphism parameters.
2. LLG or "residual" maps, which can be further analyzed to fine-tune the refinement parameters, e.g., typically to introduce anisotropic thermal parameters for certain sites.
3. Electron density maps from heavy-atom phasing, density modification, and automatic building.
4. PDB file of a partially built model.

There are separate graphical user interfaces available to: (1) change the parameterization of the heavy-atom refinement model, (2) analyze and view LLG maps, (3) view electron density maps, (4) change parameters for density modification, (5) perform automatic model building with ARP/wARP, and (6) detect NCS with GETAX *(29)*.

4. Notes

1. It can be very helpful to check the features of heavy-atom sites in the final LLG maps, especially the ones based on anomalous differences. Different heavy atoms should have different scattering properties (f″), which should be visible there. However, a wrongly assigned heavy atom type for a specific site might be accommodated for through an occupancy value that refines to very low or very high values. Furthermore, the heavy-atom positions in sulfur or Se-Met-phasing experiments can be used to help assign the correct sequence to the macromolecular chain during manual model building. Finally, during the final stages of structure refinement and model completion, when the solvent structure is being completed (either automatically or by hand), the difficult differentiation between a water atom or a bound metal ion can be helped by checking the anomalous residual (LLG) or Fourier maps at this position.
2. In cases where the actual content of the asymmetric unit is unknown and could vary substantially (within a sensible value for the solvent content resulting from a given number of monomers), it is recommended to give the asymmetric unit content for a single copy of the macromolecule, as well as the number of heavy atoms expected to be present for a monomer. autoSHARP will adjust the actual number of molecules during structure solution at two stages: initially during data analysis (to have a reasonable solvent content of around 50%) and during the density modification step (where the optimization of the solvent content should give a better estimate of the actual number of molecules in the asymmetric unit).
3. Accurate values for f′ and f″ are needed for a MAD experiment because only a single occupancy and temperature factor for each site (which is shared between the different datasets) is being refined. In other types of experiments (SAD, SIR[AS], or MIR[AS]) any error in these values will probably be compensated by the refinement of occupancy and temperature factor. However, if these parameters are far

from their true values, the starting occupancy at the first refinement cycle can be quite wrong (which is additionally influenced by the data not being on absolute scale, when the correct number of molecules is not known).

4. The reflection data used within autoSHARP should have been processed as thoughtfully as possible. In particular, the treatment of resolution limits has to be done carefully: including high-resolution data with hardly any signal can give difficulties during the heavy-atom detection because the direct methods program RANTAN uses normalized structure factors (E-values) so that no automatic down-weighting of weak and noisy data in the outer resolution shells occurs. Furthermore, errors in the handling of the area behind the beam stop during the integration of diffraction images can lead to large errors in low-resolution strong reflections, hence, to problems in the detection of outliers during internal scaling or merging of data. A substantial part of the statistics and analyses tries to deal with these problematic reflections (or resolution shells). Any warning message from autoSHARP about data quality usually points to problems during data integration and scaling/merging effects that autoSHARP can only try to diminish but is unable to correct.
5. In cases where the asymmetric unit can accommodate a large range of numbers of molecules, some additional information is usually available to help decide on an appropriate value. Crystals that diffract only to low resolution quite often have a large solvent content or a loose packing of few molecules. On the other hand, well-diffracting crystals can have very tight packing of several molecules. Moreover, some biochemical data might be available to suggest an oligomerization state in vivo—and the asymmetric unit (maybe in conjunction with special symmetry elements) should probably take this into account.
6. If no fluorescence scan was performed, then some "standard" values for the typical wavelengths at which to perform a MAD experiment (e.g., Se-Met) can be input. These can usually be obtained from the beam line scientist. Using a more uncommon heavy atom with more complicated absorption edges will require the measurement of a fluorescence scan which can be analyzed, e.g., with the program CHOOCH ***(30)***.
7. The overall temperature factor obtained from a Wilson plot is used as a starting value for the temperature factor of newly found heavy-atom sites (additional sites that are found through analysis of the LLG maps will have the mean temperature of all already existing heavy-atom sites within this dataset). Although heavy atoms introduced through soaking of the crystal (e.g., Hg or Pt soaks) usually have a much larger temperature factor, it seems a good starting point.
8. During data collection, the exposure time, crystal-to-detector distance, oscillation angle, and strategy can easily be adapted to give a nearly complete dataset. If there are time restraints during data collection it might be better to collect a complete, lower resolution dataset (by increasing detector-to-crystal distance, which will allows larger oscillation ranges and/or shorter exposure) than a high resolution but incomplete dataset. This is especially true when the missing data are not randomly distributed, but rather a wedge of data is missing. This not only can lead to streaks in the resulting electron density maps, but also makes the analysis of LLG

maps more difficult. The estimation of overall anisotropic temperature factors can give wrong results because this determination can be ill determined.

9. As with many of these statistics, any outlier or unlikely value for low-resolution reflections should be seen as suspicious. Usually, low-resolution reflections are strongest and should have been integrated most accurately (whereas at the outer diffraction limit reflections are weak and show greater noise). However, if some systematic errors were introduced during data processing, these strong reflections not only can have a wrong value, but their corresponding variance might be wrongly estimated—leading to a wrong weighting of these reflections during maximum-likelihood calculations as those performed within SHARP.
10. If reflections within a very narrow resolution range are listed here, it might be because an ice-ring hasn't been processed appropriately for one of the datasets. Also, a larger number of rejections at very low resolution can again show problems with the data processing and the treatment of the beamstop.

Acknowledgments

The authors wish to acknowledge partial financial support for this work from European Commission Grant no. QLRI-CT-2000-00398 within the AUTOSTRUCT project.

References

1. Bourne, S. R. (1978) Unix time-sharing system: The Unix shell. *Bell Sys. Tech. J.* **57,** 1971–1990.
2. Wall, L., Christiansen, T., and Orwant, J. (2000) *Programming Perl, 3rd ed.*, O'Reilly and Associates, Inc., Sebastopol, CA.
3. Collaborative Computational Project, Number 4 (1994) The CCP4 suite: programs for protein crystallography. *Acta Cryst.* **D50,** 760–763.
4. Otwinowski, Z. and Minor, W. (1997) *Methods in Enzymology, Vol. 276, Macromolecular Crystallography, Part A,* (Carter, C. W., Jr and Sweet, R. M., eds.), Academic Press, New York, NY, pp. 307–326.
5. Cromer, D. T. (1983) Calculation of anomalous scattering factors at arbitrary wavelengths. *J. Appl. Cryst.* **16,** 437–438.
6. Wilson, A. J. C. (1942) Determination of absolute from relative X-ray intensity. *Nature* **150,** 151–152.
7. Matthews, B. W. (1968) The solvent content of protein crystals. *J. Mol. Biol.* **33,** 491–497.
8. Rossmann, M. G. and Blow, D. M. (1962) The detection of sub-units within the crystallographic asymmetric unit. *Acta Cryst.* **15,** 24–31.
9. Patterson, A. L. (1934) A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.* **46,** 372–376.
10. Eagles, P. A. M., Johnson, L. N., Joynson, M. A., McMurray, C. H., and Gutfreund, H. (1969) Subunit structure of aldolase: chemical and crystallographic evidence. *J. Mol. Biol.* **45,** 533–544.

11. Kraut, J., Sieker, L. C., High, D. F., and Freer, S. T. (1962) Chymotrypsinogen: a three-dimensional Fourier syntheis at 5AA resolution. *Proc. Nat. Acad. Sci.USA* **48,** 1417–1424.
12. Schneider, T. R. and Sheldrick, G. M. (2002) Substructure solution with *SHELXD*. *Acta Cryst.* **D58,** 1772–1779.
13. Weeks, C. M. and Miller, R. (1999) The design and implementation of SnB v2.0. *J. Appl. Cryst.* **32,** 120–124.
14. Grosse-Kunstleve, R. W. and Adams, P. D. (2003). Substructure search procedures for macromolecular structures. *Acta Cryst.* **D59,** 1966–1973.
15. Terwilliger, T. C. (2003) Automated structure solution, density modification and model building. *Acta Cryst.* **D58,** 1937–1940.
16. Ness, S. R., de Graaff, R. A. G., Abrahams, J. P., and Pannu, N. S. (2004). Crank: new methods for automated macromolecular crystal structure solution. *Structure* **12,** 1753–1761.
17. Yao, J. (1981) On the application of phase relationships to complex structures. XVIII. *RANTAN*-random *MULTAN*. *Acta. Cryst.* **A37,** 642–644.
18. Westbrook, J. and Fitzgerald, P. M. (2003) The PDB format, mmCIF formats and other data formats. In: *Structural Bioinformatics* (Bourne, P. E. and Weissig, H., eds.), John Wiley and Sons, Inc., Hoboken, NJ, pp. 161–179.
19. Fan, H.-F., Woolfson, M. M., and Yao, J.-X. (1993) New techniques of applying multi-wavelength anomalous scattering data. *Proc. R. Soc. Lond. A* **442,** 13–32.
20. Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R., and Uson, I. (2001) Direct methods: *ab initio* phasing. In: *International Tables for Macromolecular Crystallography, Vol. F* (Rossmann, M. G., and Arnold, E., eds.), Kluwer Academic Publishers, Dordrecht, Germany, pp. 333–345.
21. Brünger, A. T. (1997) Free R value: cross-validation in crystallography. *Meth. Enzym.* **277,** 366–396.
22. De La Fortelle, E. and Bricogne, G. (1997) Maximum-likelihood heavy-atom parameter refinement for the multiple isomorphous replacement and multiwavelength anomalous diffraction methods, *Meth. Enzym.* **276,** 472–494.
23. Hendrickson, W. A. and Lattman, E. E. (1970) Representation of phase probability distributions for simplified combination of independent phase informaton. *Acta Cryst.* **B26,** 136–143.
24. Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M., and Paciorek, W. (2003) Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Cryst.* **D59,** 2023–2030.
25. Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G., and North, A. C. T. (1961) The structure of haemoglobin. VIII. A three-dimensional Fourier synthesis at 5.5 Å resolution: Determination of the phase angles. *Proc. Roy. Soc. A* **265,** 15–38.
26. Cowtan, K. D. and Zhang, K. Y. (1999) Density modification for macromolecular phase improvement. *Prog. Biophys. Mol. Biol.* **72,** 245–270.
27. Abrahams, J. P. and Leslie, A. G. W. (1996) Methods used in the structure determination of bovine mitochondrial F^1 ATPase. *Acta Cryst.* **D52,** 30–42.

28. Perrakis, A., Morris, R. J., and Lamzin, V. S. (1999) Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6,** 458–463.
29. Vonrhein, C. and Schulz, G. E. (1999). Locating proper non-crystallographic symmetry in low-resolution electron-density maps with the program GETAX. *Acta Cryst.* **D55,** 225–229.
30. Evans, G. and Pettifer, R. F. (2001) CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra. *J. Appl. Cryst.* **34,** 82–86.

13

Introduction to Macromolecular Refinement

Dale E. Tronrud

Summary

The process of refinement is such a large problem in function minimization that even the computers of today cannot perform the calculations to properly fit X-ray diffraction data. Each of the refinement packages currently under development reduces the difficulty of this problem by utilizing a unique combination of targets, assumptions, and optimization methods. This chapter summarizes the basic methods and underlying assumptions in the commonly used refinement packages. This information can guide the selection of a refinement package that is best suited for a particular refinement project.

Key Words: Crystallography; macromolecular; refinement; least-squares refinement; maximum likelihood.

1. Introduction

Refinement is the optimization of a function of a set of observations by changing the parameters of a model.

This is the definition of macromolecular refinement at its most basic level. To understand refinement we need to understand the definitions of its various parts. The four parts are "optimization," "a function," "observations," and "the parameters of a model."

Although formally different topics, these concepts are tightly connected. For example, one cannot choose an optimization method without considering the nature of the dependence of the function on the parameters and observations. Further, in some cases one's confidence in an observation is so great that the parameters are devised to make an inconsistent model impossible (e.g., the observations become constraints).

Each of these topics will be described in detail. An understanding of each topic and their implementation in current programs will enable the selection of the most appropriate program for a particular project.

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

2. Observations

The "observations" include everything known about the crystal prior to refinement. This set includes commonly noted observations such as unit cell dimensions, structure factor amplitudes, standardized stereochemistry, and experimentally determined phase information. Other types of knowledge about the crystal, which are usually not thought about in the same way, include the primary structure of the macromolecules and the mean electron density of the mother liquor.

For a particular observation to be used in refinement, it must be possible to gage the consistency of the model with this observation. Current refinement programs require that this measure be continuous. If a property is discrete, some mathematical elaboration must be created to transform the measure of the model's agreement into a continuous function.

As an example, consider chirality: the α-carbon of an amino acid is either in the D or L configuration. It cannot be 80% L and 20% D. Because the agreement of a model to this bit of knowledge is discrete, the derivative of the agreement function is not informative. To allow the correction of this sort of error most programs use some function in which the chirality is expressed as its sign (e.g., a chiral volume or improper dihedral angle). Because the additional information in these residual functions is simply the ideal bond lengths and angles, restraining chirality in this fashion causes geometrical restraints to be included in the refinement via two different routes. This duplication makes it difficult to assign proper weights to this information.

This problem is not encountered with most types of observations. Diffraction amplitudes, bond lengths, and angles calculated from the model can be varied by small changes in the model's parameters.

Each observation should be accompanied by an indication of its confidence. If the uncertainty in an observation follows a normal distribution, then one's confidence in the observation is indicated by the standard deviation ("σ") of that distribution. In more complicated situations, a complete description of the probability distribution will be required. Experimental phases are examples of this difficult class of observations as their uncertainties can be quite large and multiple maxima are possible.

2.1. Stereochemical Restraints

When a diffraction dataset is missing high-resolution reflections, the details of the molecule cannot be visualized. Fortunately, the molecule can be viewed as a set of bond lengths, bond angles, and torsion angles, instead of the conventional view of a set of atoms floating in space (*see* **Fig. 1**). The advantage derived from this geometrical view of a structure is that the values of the bond lengths and angles, and their standard uncertainties, can be predicted using high-resolution, small molecule structures ***(1)***.

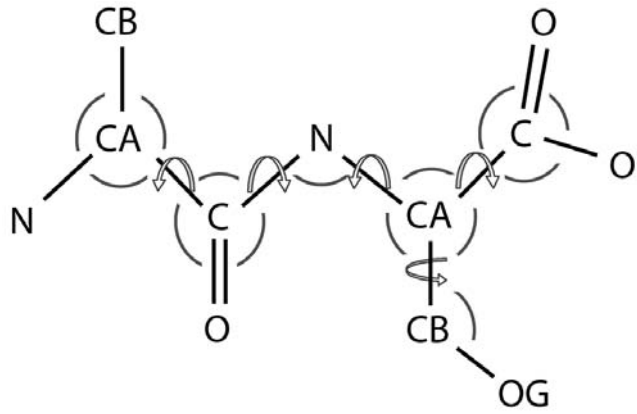


Fig. 1. Stereochemical restraints in a dipeptide. This figure shows the bonds, bond angles, and torsion angles for the dipeptide Ala-Ser. The dark lines indicate bonds, arcs bond angles, and arrows torsion angles. The values of the bond lengths and bond angles are, to the precision required for most macromolecular refinement problems, independent of the environment of the molecule and can be estimated reliably from small molecule crystal structures. The values of most torsion angles are influenced by their environment and, although small molecule structures can provide limits on their values, they cannot be uniquely determined without information specific to this crystal.

It is instructive to note that this example molecule contains 12 atoms and requires 36 degrees of freedom to define their positions (12 atoms with three coordinates for each atom). The molecule contains 11 bonds, 14 bond angles, and 5 torsion angles, which among themselves define 30 degrees of freedom. The unaccounted for degrees of freedom are the six parameters, which define the location and orientation of the entire dipeptide. This result is general; the sum of the number of bonds, the number of bond angles, the number of torsion angles, and six will always be three times the number of atoms.

Other stereochemical restraints, such as chiral volume and planarity, are redundant. For example, the statement that the carbonyl carbon and the atoms that bond to it form a planar group is equivalent to saying that the three bond angles around the carbonyl carbon sum to 360°. These types of restraints are added to refinement packages to compensate for their, incorrect, assumption that deviations from ideality for bond angles are independent of each other.

To be honest, the most interesting aspects of a molecule are the angles of rotation about its single bonds. If the φ and ψ angles of the backbone of the polypeptide chain and the χ angles of the side chains were known, most of the questions about the structure could be addressed. The scatter of bond angles and planarity seen in accurate structures is large enough, however, that one cannot constrain a model to "ideal" values without creating significant errors. For example, if a peptide ω angle (the angle of rotation about the C and N atoms in

the peptide bond) differs from the value which results in a planar peptide bond (as it can, *see* **ref.** *2* as one example of many) but is forced into a plane, the protein's backbone will be distorted over many residues to compensate for the error. Refinement with constrained bond lengths and angles was implemented in the early 1970s in Diamond's real space refinement program *(3)* but was eventually abandoned, in part, because of this problem.

Even though stereochemical constraints on bond lengths and angles do not work, this knowledge can still be applied as restraints. Constraints simplify the refinement problem by reducing the number of parameters. Restraints work instead by increasing the number of observations: a penalty is imposed for deviations from ideal stereochemistry just as a penalty is imposed for deviations from observed structure factor amplitudes, but deviations are allowed.

3. The Parameters of the Model

The general perception of the parameters of a molecular model is dominated by the Protein Data Bank (PDB; www.pdb.org) file format *(4)*. In this file a molecule is a collection of atoms, each defined by its location, an occupancy, and a "B" factor. (Each atom also has a type, but because these types are not continuously variable they are usually not considered parameters.)

The B factor provides an estimate of an atom's vibration about its central position. The usual form is to either define the B factor as isotropic, meaning that the atom vibrates equally in all directions and can be visualized as lying within a sphere, or to define an anisotropic B factor, which describes vibration of the atom within an ellipsoid centered at the atomic coordinate. Six parameters are required to define such an ellipsoid *(5)*. The B factor is isotropic when the off-diagonal elements of this matrix are equal to zero and the diagonal elements are all equal to each other. Therefore, only one number is required to define an isotropic B.

In the traditional formulation, each atom is defined by (1) three numbers that give its location in the unit cell, (2) one number for its occupancy, and (3) either one number for an isotropic B or six numbers for an anisotropic B. These numbers form the principal set of parameters for the model.

In a medium-sized protein there are approx 2500 atoms. With five parameters for each atom there would be 12,500 parameters and with 10 parameters per atom there would be 25,000 parameters. For such a protein, a diffraction dataset to 2 Å resolution would contain about 22,000 reflections. Because the mathematical relationship between the structure factors and the model is nonlinear, macromolecular refinement will not produce useful results unless there are many times more reflections in the dataset than parameters in the model. Clearly, the refinement of a model with anisotropic Bs at 2 Å resolution will be problematic, and one with isotropic Bs is borderline.

This difficulty is not restricted to molecules of a particular size. The larger the molecule the greater the number of parameters, but the unit cell will also increase in size, which increases the number of reflections at a given resolution. The ratio of observation to parameters essentially depends only on resolution for all sizes of molecules. There is some effect as a result of solvent content of the cell, with large solvent content cells resulting in relatively larger sets of reflections at the same resolution.

At such resolutions something must be done to simplify the parameterization of the model (or the number of observations should be increased). This can be done by imposing constraints on the parameters (i.e., forcing the model to be exactly consistent with the prior knowledge), or recasting the parameters into some form where the lack of compliance is impossible. The creation of program codes to implement these solutions can be very difficult. Some of the traditional solutions were devised because of limited time and limited computers and are not easily justified now.

The first parameter to go is the occupancy. Because the difference map feature, which results from an error in occupancy is very similar to that resulting from an error in a isotropic B factor, only quite high-resolution diffraction data can generate difference maps that have sufficient clarity to distinguish the two. Because the two parameters are confounded in this fashion, it would be advantageous to eliminate one of them. Fortunately, most of the atoms in the crystal are chemically bonded together and have the same occupancy, which is very close to 1.0. Applying this knowledge as a constraint allows the model to be refined with one fewer parameter per atom.

Although this simplification is quite appropriate for the atoms in the macromolecule, it is not for individual water molecules. It is quite likely that particular water molecules are not present with full occupancy, but the problem of discriminating between a low occupancy and a high B factor remains. The traditional solution is to again hold the occupancy fixed at 1.0. Although this solution is not credible, it does allow the atom to refine to flatten the difference map, and it lowers the R-value almost as much as refining both parameters would. Because of this constraint, the B factor must be redefined to be a combination of motion and occupancy. This generalization in interpretation of B factor is implicit in most macromolecular models, but is not clearly stated in the deposited models.

Unless the resolution of the diffraction data is very high, refinement of a structure containing anisotropic Bs results in a model that is physically unreasonable. To avoid this absurdity refinement is performed with isotropic Bs. This choice is not made because the motions of the atoms are actually believed to be isotropic, but simply to limit the number of parameters. The result is the paradox that the crystals that are most likely to have large anisotropic motions are modeled with isotropic Bs.

3.1. Rigid Body Parameterization

One common restructuring of the standard set of parameters is that performed in rigid body refinement. When there is an expectation that the model consists of a molecule whose structure is essentially known but whose location and orientation in the crystal is unknown, the parameters of the model are refactored. The new parameters consist of a set of atomic positions specified relative to an arbitrary coordinate system, and up to six parameters to specify how this coordinate system maps onto the crystal; up to three to describe a translation of the molecule and three to define a rotation. The traditional set of coordinates are calculated from this alternative factorization with the equation

$$\mathbf{x}_t = \mathbf{R}[q_1,q_2,q_3]\mathbf{x}_r + \mathbf{t}, \tag{1}$$

where x_t are the positions of the atoms in the traditional, crystallographic, coordinate system, $R[\theta_1,\theta_2,\theta_3]$ is the rotation matrix that rotates the molecule into the correct orientation, and t is the translation required to place the properly orientated molecule into the unit cell.

In principle, all of these parameters could be refined at the same time. Refinement is usually performed separately for each parameter class, however, because of their differing properties. The values of the orientation and location parameters are defined by diffraction data of quite low resolution, and the radius of convergence of the optimization can be increased by ignoring the high-resolution data. In addition, in those cases where rigid body refinement is used, one usually knows the internal structure of the molecule quite well, whereas the location and orientation are more of a mystery.

For this reason, molecular replacement can be considered a special case of macromolecular refinement. Because the internal structure of the molecule is known with reasonable certainty, one creates a model parameterized as the rigid body model previously described. Then one "refines" the orientation and location parameters. Because this is a small number of parameters and no good estimate for starting values exists, one uses search methods to locate an approximate solution, and gradient descent optimization to fine-tune to orientation parameters.

The principal drawback of the rigid body parameterization is that macromolecules are not rigid bodies. If the external forces of crystal packing differ between the crystal where the model originated and the crystal where the model is being placed then the molecule will be deformed. Optimizing the rigid body parameters alone cannot result in the final model for the molecule.

3.2. Noncrystallographic Symmetry Constrained Parameterization

When the asymmetric unit of a crystal contains multiple copies of the same type of molecule, and the diffraction data is not of sufficient quantity or quali-

ty to define the differences between the copies, it is useful to constrain the non-crystallographic symmetry (NCS) to perfection. In such a refinement the parameterization of the model is very similar to that of rigid body refinement. There is a single set of atomic parameters (positions, B factors, and occupancies [usually constrained equal to unity]) for each type of molecule, and an orientation and location (six parameters) for each copy.

As with rigid body refinement, the orientation and location parameters are refined separately from the internal structure parameters. First, the orientation and location parameters are refined at low resolution (typically 4 Å resolution) while the atomic parameters are held fixed. Then the atomic parameters are refined against all the data while the external parameters are held fixed.

The constrained NCS parameterization has the same shortcoming as rigid body parameters. Each copy of the macromolecule experiences a different set of external forces resulting from their differing crystal contacts and it is expected that they will respond by deforming in differing ways. The constraint that their internal structures be identical precludes the model from reflecting these differences. If the diffraction data are of sufficient resolution to indicate that the copies differ, but are not high enough to allow refinement of unconstrained parameters (without explicit consideration of NCS), then the model will develop spurious differences between the copies ***(6)***.

Relaxing the constraints and implementing NCS restraints is the usual solution chosen to overcome this problem. Most implementations of NCS restraints continue to assume that the molecules are related by a rigid body rotation and translation except for the random, uncorrelated displacements of individual atoms. If two molecules differ by an overall bending the NCS restraints will impede the models from matching that shape.

The program SHELXL ***(7)*** contains an option for restraining NCS by suggesting that the torsion angles of the related molecules be similar, instead of the positions of the atoms being similar after rotation and translation. By removing the rigid body assumption from its NCS restraints, this program allows for deformations that are suppressed by other programs.

3.3. Torsion Angle Parameterization

The substitution of atomic coordinates with torsion angles dramatically reduces the total number of parameters (*see* **Fig. 1**). This decrease is advantageous when the resolution of the diffraction data is quite low (lower than 3 Å). At these resolutions there are many fewer reflections to define the values of parameters in the traditional model. Even with the addition of bond length and angle information as restraints, these models tend to get stuck in local minima or overfit the data.

Increasing the weight on the stereochemical restraints to compensate for the lack of diffraction data does not work well because the size of the correlation

in the position of the related atoms increases greatly and optimization methods, which ignore this effect, become ineffective.

Simulated annealing also has difficulty accommodating high weights on bond lengths and angles *(8)*. When the "force constant" of a bond is large the bond's vibrational frequency increases. The highest frequency motion determines the size of the time step required in the slow-cooling molecular dynamics calculation, so increasing the weight on stereochemistry greatly increases the amount of time taken by the slow-cooling calculation.

The programs commonly used to refine models at these low resolutions (XPLOR *[9]* and CNS *[10]*) use simulated annealing and gradient descent methods of optimization. Optimization methods that use the off-diagonal elements of the normal matrix are not used in these circumstances because their radii of convergence are not large enough to correct the errors that typically are found in low-resolution models.

One solution to the problem of large stereochemistry weights is to choose a parameterization of the model where the bond lengths and angles simply cannot change. If the parameters of the model are the angles of rotation about the single bonds the number of parameters drops considerably and there is no need for a stereochemical weight (it is effectively infinite). There are, on average, about five torsion angles and about eight atoms per amino acid. Changing from an atomic model to a torsion angle model will replace 24 positional parameters with 5 angular parameters. This, nearly fivefold, reduction in parameters greatly improves the observation-to-parameter ratio in addition to improving the power of simulated annealing and gradient descent optimization.

The nature of torsion angle parameters makes the implementation of their refinement much more difficult that the other parameters described here. When working with atomic positions, for example, one can estimate the shifts to be applied by considering the improvement in the residual function by moving each atom in turn, holding the other atoms fixed in space. This form of calculation cannot be performed with torsion angle parameters. If the first main chain torsion angle is varied the majority of the molecule is moved out of density and any amount of shift is rejected. The variation of a torsion angle can only lead to improvement if other torsion angles are simultaneously varied in compensatory fashion. The most flexible solution to this problem, to date, is described in **ref.** *8*.

3.4. TLS B Factor Parameterization

Probably the most significant, inappropriate constraint applied generally to protein models is the isotropic B factor. It is quite certain that atoms in crystals that diffract to resolutions lower than 2 Å move anisotropically, and yet they are routinely modeled as isotropic. Although the excuse for this choice is the un-

deniable need to reduce the number of parameters in the model, this clearly is not a choice likely to improve the fit of the model to the data.

Schomaker and Trueblood *(11)* described a parameterization that allows the description of anisotropic motion with many fewer parameters than an independent anisotropic B factor for each atom. This parameterization is called TLS, for translation, libration, and screw. In this system the motion of a group of atoms is described by three matrices, one for a purely translational vibration of the group, a second for libration (or wobbling) of the group about a fixed point, and a third for a translation and libration, which occurs in concert. The explicit assumption of TLS B factors is that the group of atoms moves as a rigid unit. More complicated motions can be modeled by nesting several TLS groups within a larger group, creating a tree-like data structure.

TLS B factors are difficult to implement as parameters in a refinement program. The programs RESTRAIN *(12)* and, more recently, REFMAC *(13,14)* include the option of refining TLS B factors.

In the TLS formalism, 20 parameters are used to describe the motion of the entire group of atoms. Because the anisotropic B of one atom requires six parameters, any TLS group composed of more than three atoms results in a decrease in the total number of parameters. Of course a large number of small TLS groups will not reduce the parameter count very much, and will only be refinable with higher resolution data than a TLS model containing large groups. Then again a TLS model composed of large groups might not be able to mimic the set of anisotropic B factors required to fit the data.

In the absence of a related structure refined with anisotropic Bs at atomic resolution it is difficult to objectively define rigid groups larger than side chains with aromatic rings.

4. The Function

In crystallographic refinement three functions are commonly used. They are the empirical energy function, the least-squares residual, and maximum likelihood.

4.1. Empirical Energy

The idea that the best model of a protein would be the one with the lowest energy has been used since the early 1970s (*see* **ref.** *15* as an example). To a person with a background in biochemistry, such a measure is quite intuitive. The program will give the difference between two conformations or two models in kcal/mol, which is a familiar unit.

There are two principal problems with this function as a refinement residual. The first problem is that it has been impossible, so far, to devise an empirical energy function that is accurate enough to reproduce experimental results. If the function

is not reliable the models generated using it cannot be trusted either. The second problem is that there is no statistical theory underlying this function. None of the vast array of mathematical tools developed in other fields can be applied to an analysis of neither the quality of the model nor the nature of its remaining errors.

Although the refinement packages XPLOR *(9)* and CNS *(10)* use the language of energy in their operation, the actual function used is closer to one of the other two functions. It is important to remember that these programs are not even attempting to calculate "energies" that relate to binding energies and stability.

4.2. Least Squares

Least squares is the simplest statistical method used in macromolecular refinement. Like empirical energy its history in macromolecular structure determination extends back to the 1970s *(16)* and continues to be used today.

The least-squares residual function is

$$f(\mathbf{p}) = \sum_{i}^{\text{all data}} \frac{(Q_o(i) - Q_c(i,\mathbf{p}))^2}{\sigma_o(i)^2}, \tag{2}$$

where $Q_o(i)$ and $\sigma_o(i)$ are the value and standard deviation for observation number i. $Q_c(i, \mathbf{p})$ is the model's prediction for observation *i* using the set of model parameters **p**. The larger the difference between the observation and the model's prediction, the worst the model. The more precisely we know an observation, the more important that observation becomes in the overall sum. One varies the parameters of the model to find a set that gives the lowest sum of deviants.

The values of the parameters found by minimizing this function are those that have the smallest individual standard deviation, or the smallest probable error *(17)*. This statement is only true, however, if the assumptions of the method are correct. The assumptions of least squares are that the errors in the observations obey a normal distribution with completely known ("observed") variances, and that, given perfect observations and the best parameters, the model would predict the observations perfectly.

In recent years it has been shown *(18–20)* that these assumptions are incorrect in many refinement problems. The simplest example occurs when the model is incomplete, say missing a domain. With an imperfect model of this type it is impossible for any set of parameters to reproduce all the observations. The refinement function must account for the unknown contribution of the unmodeled part of the molecule, and least squares cannot do that.

4.3. Maximum Likelihood

To construct a refinement function that is not limited by the assumptions of least squares, one must generalize the method. Such a generalization is called maximum likelihood. Currently maximum likelihood options are available in

the programs CNS *(10)*, REFMAC *(13)*, and BUSTER/TNT *(21,22)*. These programs are listed in order of increasing sophistication of their implementation of maximum likelihood.

Maximum likelihood is a generalized statistical framework for estimating the parameters of a model based on observations *(23,24)*. It differs from least squares in that it can accommodate observations with uncertainties of arbitrary character and model parameters whose values are expected to have such uncertainties as well.

Although the maximum likelihood method is completely general, macromolecular refinement is such a difficult problem that no computer can perform a likelihood refinement in complete generality. The authors of each computer program must make particular assumptions about the nature of the uncertainties in the observations and the parameters of the final model in order to produce a program that will produce a result in a reasonable amount of time.

Although least squares is rather simple and usually implemented similarly in all programs, maximum likelihood depends critically on a detailed model of how errors are distributed and the consequences of these errors. Each implementation of maximum likelihood makes it own set of assumptions, and one may work better than another may in any particular problem.

4.3.1. Overview of Bayesian Inference

Maximum likelihood itself is an approximation of the general Bayesian inference procedure *(24)*. Bayesian inference is a means of combining all information known about a problem in a completely general fashion.

When using it, one starts by calculating, for every combination of values of the parameters of the model, how probable that set of parameters is when all of the information known prior to the current experiment is considered. In crystallographic refinement this information would include basic properties (e.g., that anisotropic B factors must be positive definite, and isotropic B factors must be positive), stereochemical information (e.g., atom CA of a particular residue is approx 1.52 Å from atom C and its isotropic B factor is approx 4 $Å^2$ smaller), and various conventions (e.g., that at least one atom of each molecule should lie in the conventional asymmetric unit of the unit cell). This probability distribution is named the prior distribution.

The second probability distribution is called the likelihood distribution. This distribution contains, for every combination of values for the parameters of the model, the probability that the experiment would have turned out as it did, assuming that set of values was correct. If the predicted outcome of the experiment for a particular set of values differs from the actual experimental results by much more than the expected uncertainty in both the measurements and the ability of the model to predict, then the probability is quite low.

Any set of values is only worth considering if it has high probabilities in both distributions. Therefore, the two distributions are multiplied to generate a new probability distribution, called the posterior probability, which includes all of the information about the values of the parameters. If the posterior distribution contains a single, well-defined peak, that peak is the solution. The width of the peak would indicate how precisely these values are known. If there are multiple peaks of about the same height or if there is a single peak that is diffuse, then the experiment has not produced information sufficient to distinguish between the various possible sets. In this case one can study the posterior probability distribution to help design the next experiment.

Unfortunately, calculating the prior and likelihood distributions for all combinations of values for the parameters of a macromolecular model is well beyond the capability of current computers.

As described here the posterior probability is not normalized. To normalize it one must divide it by the probability of the experimental data given what was known about such data prior to the experiment. In the case of diffraction data this information would include Wilson statistics *(25,26)* and the non-negativity of structure factor amplitudes. Because we have one set of experimental data this normalization factor is simply one number, and can be ignored without affecting the shape of the posterior probability distribution.

4.3.2. The Maximum Likelihood Approximation

The maximum likelihood method depends on the assumption that the likelihood distribution has a single peak, whose location is approximately known. This assumption allows one to ignore nearly all of the volume of the distribution and concentrate on the small region near the starting model. The procedure for finding the values for the parameters that result in the greatest likelihood reduces to a function optimization operation very similar in structure to that used by the least-squares refinement programs of the past. To increase this similarity the negative logarithm of the likelihood function is minimized in place of maximizing the likelihood itself.

The basic maximum likelihood residual is

$$f(\mathbf{p}) = \sum_{i}^{\text{all data}} \frac{(Q_o(i) - < Q_c(i,\mathbf{p}) >)^2}{(\sigma_o(i)^2 + \sigma_c(i,\mathbf{p})^2}, \tag{3}$$

where the symbols are very similar to those of **Eq. 2**. In this case, however, the quantity subtracted from $Q_o(i)$ is not simply the equivalent quantity calculated from the parameters of the model, but the expectation value of this quantity calculated from all the plausible models similar to **p**. $\sigma_c(i, \mathbf{p})$ is the width of the distribution of values for $Q_c(i, \mathbf{p})$ over the plausible values for **p**.

For diffraction data the "quantities" are the structure factor amplitudes. The expectation value of the amplitude of a structure factor ($<|F_{calc}|>$) calculated from a structural model, which itself contains uncertainties, is calculated by integrating over all values for the phase as in **Fig. 2C**. The mathematics of this integral are difficult and beyond the scope of this chapter. The calculation of ($<|F_{calc}|>$) is discussed in **refs.** ***27*** and ***13***.

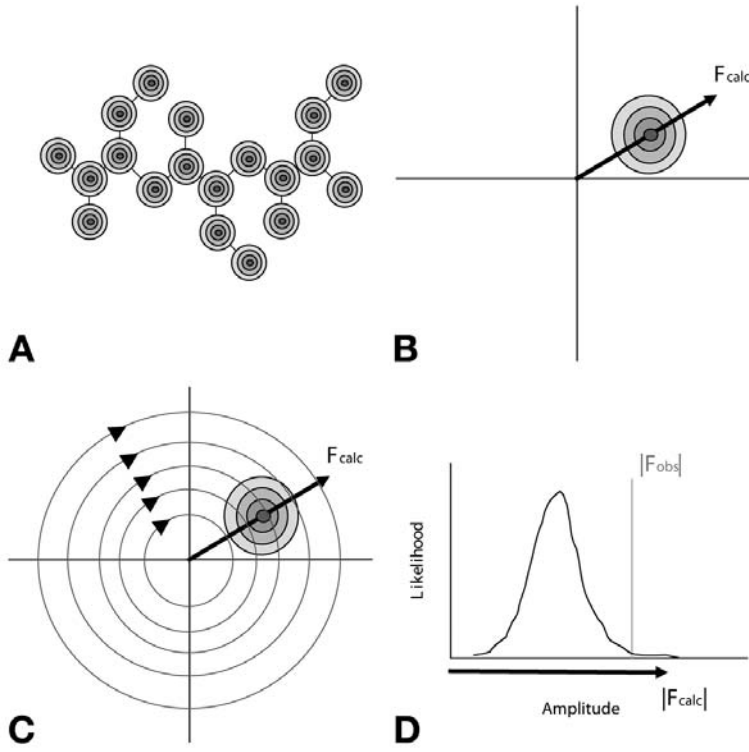
The maximum likelihood method also depends on the assumption that the prior probability distribution contains no information. This assumption is certainly not valid in macromolecular refinement where there is a wealth of information about macromolecules. Somehow maximum likelihood must be modified to preserve this knowledge. The authors of the current refinement programs overcome this problem by including the stereochemical information in the likelihood calculation as though they were results of the "experiment" in essentially the same way it is done in least-squares programs.

Perhaps a simpler way of viewing this solution is to call the procedure "maximum posterior probability" and optimize the product of the likelihood and prior distributions by varying the values of the parameters in the neighborhood of a starting model.

4.3.3. Comparing Maximum Likelihood and Least Squares

Figure 3 shows the mathematical world that crystallographic least-squares refinement inhabits. There are two key features of least squares that are important when a comparison with maximum likelihood is made: (1) the identification of the measurement of the observation as the only source of uncertainty, and (2) the absence of any consideration of the uncertainty of phase of the reflection. **Figures 2** and **4** show probability distributions used in maximum likelihood equivalent to **Fig. 3**.

A fundamental difference between the least-square worldview and that of maximum likelihood is that least squares presumes that small random changes in the values of the parameters will cause small random changes in the predicted observations. Although atomic positions are recorded to three places beyond the decimal point in a PDB file, this degree of precision was never intended to be taken seriously. Usually somewhere in the paper a statement similar to "the coordinates in this model are accurate to 0.15 Å" is made. When calculating structure factors to be compared with the observed structure factor amplitudes, the structure factor of the particular model listed in the deposition is not the value desired. Instead, the central (or best) structure factor of the population of structures that exist within the error bounds quoted by the author is needed. When there is a linear relationship between the parameters of the model and the observations this distinction is not a problem. The center of the distribution of parameter values transforms to the center of the distribution of observations.



When the relationship is not linear this simple result is no longer valid. One must be careful to calculate the correct expectation value for the predicted observation with consideration of the uncertainties of the model. This complication was anticipated by Srinivansan and Parthasarathy *(28)* and Read *(29)* but was not incorporated into refinement programs until the 1990s.

The mathematical relation that transforms a coordinate model of a macromolecule into structure factors is shown in **Fig. 2**. The uncertainty in the positions and B factors of the model causes the expectation value of the structure factor to have a smaller amplitude than the raw calculated structure factor, but the same phase. The greater the uncertainty, the smaller the amplitude of the expectation value, with the limit of complete uncertainty being an amplitude of zero. As expected when the uncertainty of the values of the parameters increases the uncertainty of the prediction of the structure factor also increases.

Figure 4 shows the Argand diagram for the case where one also has atoms in the crystal that have not been placed in the model. If one has no knowledge of the location of these atoms then the vector F_{part} has an amplitude of zero

Fig. 2. Probability distributions for one reflection in the maximum likelihood worldview. **(A)** The maximum likelihood method begins with the assumption that the current structural model itself contains errors. The first panel of this figure represents the probability distributions of the atoms in the model. Instead of a single location, as the least-squares method assumes, there is a cloud of locations each atom could occupy. Although not required by maximum likelihood, the computer programs available today assume that the distribution of positions are normal and have equal standard deviations. (The value of which is defined to be that value which optimizes the fit of the model to the test set of diffraction data *[27,30]*.) **(B)** The distribution of structures shown in **A** results in a distribution of values for the complex structure factors calculated from that model. An example distribution is shown in the second panel. The value of the structure factor calculated from the most probable model is labeled F_{calc}. The nonlinear relationship between real and reciprocal space causes this value not to be the most probable value for the structure factor distribution. As shown by Read *(29)* the most probable value has the same phase as F_{calc} but has an amplitude which is only a fraction of that of F_{calc}. This fraction, conventionally named D, is equal to unity when the model is infinitely precise and zero when the model is infinitely uncertain. The width of the distribution, named σ_{calc} also arises from the coordinate uncertainty and is large when D is small and zero when D is unity. The recognition that the structure factor calculated from the most probable model is not the most probable value for the structure factor is the key difference between least squares and the current implementations of maximum likelihood. **(C)** In refinement without experimental-phase information, the probability distribution of the calculated value of the structure factor must be converted to a probability distribution of the amplitude of this structure factor. This transformation is accomplished by mathematically integrating the two-dimensional distribution over all phase angles at each amplitude. The third panel represents this integral by a series of concentric circles. **(D)** The fourth panel shows the probability distribution for the amplitude of the structure factor. The bold arrow below the horizontal axis represents the amplitude of F_{calc}; calculated from the most probable model. As expected, the most probable amplitude is smaller than $|F_{calc}|$. With this distribution the likelihood of any value for $|F_{obs}|$ can be evaluated, but more importantly one can calculate how to modify the model to increase the likelihood of $|F_{obs}|$. In this example, the likelihood of $|F_{obs}|$ is improved by either increasing $|F_{calc}|$ or increasing the precision of the model. This action is the opposite of the action implied by the least-squares analysis of **Fig. 3**.

and the phase of the center of the distribution is the same as that calculated from the structural model (as was the case in **Fig. 2**). If, however, one has a vague idea where the unbuilt atoms lie, their contribution (F_{part}) will have a nonzero amplitude and the center of the probability distribution for this reflection will have a phase different than that calculated from the current model. The ability to alter the probability distribution by adding this additional information reduces the bias of the distribution toward the model

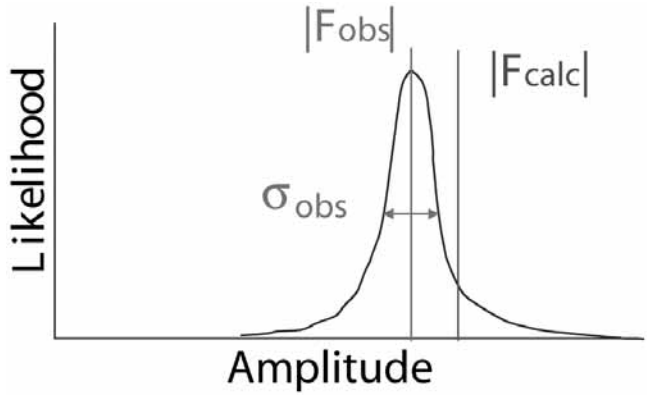
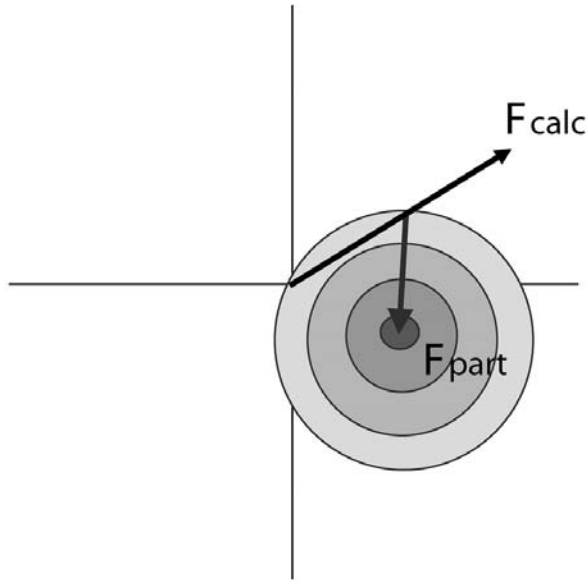


Fig. 3. Probability distribution for one reflection in the least-square worldview. In least squares it is assumed that the observed and calculated structure factors have exactly the same phase, so the only error to consider is in the magnitude of the observation. The true value of $|F_{obs}|$ is assumed to be represented by a one-dimensional Gaussian centered at its measured value and with a spread related to its estimated standard uncertainty, σ_{obs}. The calculated amplitude is assumed to have no spread at all. In this example, the parameters of the model should be modified to cause $|F_{calc}|$ to become smaller.



already built. Such models can only be refined with BUSTER/TNT ***(31)*** at this time.

5. The Optimization Method

Function minimization methods fall on a continuum (*see* **Fig. 5**). The distinguishing characteristic is the amount of information about the function that must be explicitly calculated and supplied to the algorithm. All methods require the ability to calculate the value of the function given a particular set of values for the parameters of the model. Where the methods differ is that some require only the function values (simulated annealing is such a method: it uses the gradient of the function only incidentally in generating new sets of parameters), whereas others require the gradient of the function as well. The latter class of methods is called gradient descent methods.

The method of minimization that uses the gradient and all of the second derivative (i.e., curvature) information is called the "full-matrix" method. The full-matrix method is quite powerful but the requirements of memory and computations for its implementation have been beyond current computer technology, except for small molecules and smaller proteins. Also, for reasons to be discussed, this algorithm can only be used when the model is very close to the minimum—closer than most "completely" refined protein models. For proteins, it has only been applied to cases where the molecule is small (<2000 atoms), which diffract to high resolution and have previously been exhaustively refined with gradient descent methods.

The distance from the minimum at which a particular method breakdown is called the "radius of convergence." It is clear from **Fig. 5** that the full-matrix method is much more restrictive than other gradient descent methods, and the gradient descent methods are more restrictive than simulated annealing.

Fig. 4. Probability distribution for maximum likelihood in the presence of unbuilt structure. This figure shows the probability distribution in the complex plane for the case where, in addition to the modeled parts of the crystal, there is a component present in the crystal for which an explicit model has not been built. This distribution is an elaboration of that shown in **Fig. 2B**. That distribution is convoluted with the probability distribution of the structure factor calculated from the envelope where the additional atoms are believed to lie and weighted by the number of atoms in this substructure (which can be represented as a distribution centered on the vector F_{part}). The resulting distribution has a center that is offset by F_{part} and a width that is inflated relative to that of **Fig. 3B** by the additional uncertainty inherent to the unbuilt model.

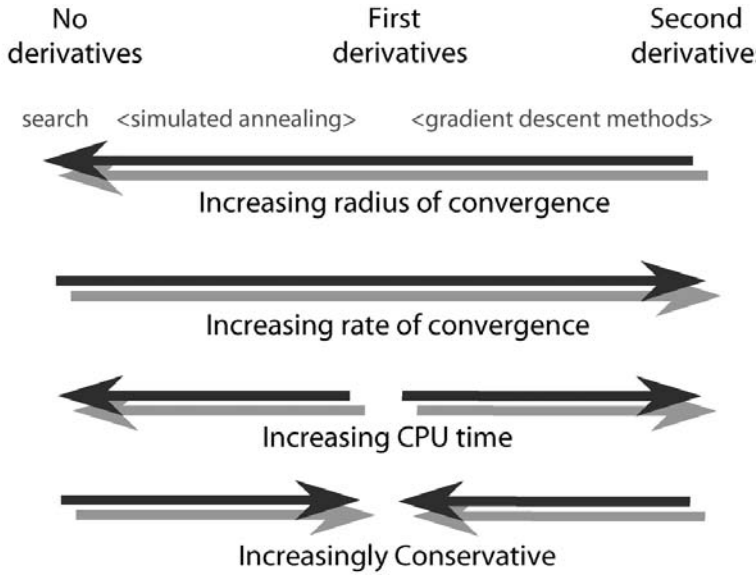


Fig. 5. Properties of various optimization methods. The principal properties of optimization methods considered here are the "radius of convergence," "rate of convergence," "CPU time," and "conservativity." The radius of convergence is a measure of the accuracy required of the starting model. The rate of convergence is the number of iterations of the method required to reach an optimum solution. The CPU time represents the amount of time required to reach the optimum. The conservativity is a measure of the tendency of a method of optimization to preserve the values of parameters when changes would not affect the fit of the model to the data.

The locations of several optimization methods on these continuums are indicated by the placement of their names. The search method uses no derivatives and is located furthest to the left. The simulated annealing method occupies a range of positions, which is controlled by the temperature of the slow-cooling protocol. Steepest descent uses only first derivatives, although conjugate gradient, preconditioned conjugate gradient, and the full-matrix methods use progressively more second derivatives.

Basically the less information about the function calculated at a particular point, the larger the radius of convergence will be.

5.1. Search Methods

Of the many methods of minimizing functions, the simplest methods to describe are the search methods. Pure search methods are not used in macromolecular refinement because of the huge amount of computer time that they would require, but are routinely used in molecular replacement. To determine

the best orientation of the trial model in a crystal one simply calculates the fit of the model to the observations for an exhaustive set of trials. Once the entire set of calculations has been completed the best one is simple to identify.

The common motif of search methods is that they each have some means of selecting which combination of parameters to test and they simply keep track of the best one found so far. One can systematically sample all combinations, or randomly pick values for the parameters. If the function being minimized has some property that restricts the possible solutions, this information can be used to guide the search (such as packing restrictions in molecular replacement).

The more combinations tested the greater the chance that the absolute best solution will be stumbled upon, and the greater the precision of the answer. It is rare for a search method to find the best parameters exactly. Usually the answers from a search method are used as the starting point for a gradient descent minimization, which will fine tune the result.

5.1.1. Simulated Annealing

Simulated annealing ***(32,33)*** is a search method. A random set of models are compared with the observations. Because it is known that the correct model must have good bond lengths and angles, the random model generator is chosen to ensure that all its output has reasonable geometry. The random generator used is a molecular dynamics simulation program. Because the parameters of the model have "momentum" they can "move" through flat regions in the function, and even over small ridges and into different local minima. The "annealing" part of the method is to start with high "velocities" ("temperature"), to allow the model great freedom, and slowly reduce the momentum until eventually the model is trapped in a minimum that is hoped to be the global minimum.

The explanation of simulated annealing involves a lot of quotes. These words (e.g., momentum and temperature) are analogies and should not be taken too seriously.

The principal advantage of simulated annealing is that it is not limited by local minima and, thus, can correct errors that are quite large. This can save time by reducing the effort required for manual rebuilding of the model.

The principal disadvantage is the large amount of computer time required. Because so much time is required to complete a proper slow-cool protocol, the protocols used in crystallographic refinement are abbreviated versions of what is recommended in the wider literature. Because of this compromise the model can get trapped with poor conformations. It also becomes possible that some regions of the model that were correct at the start will be degraded by the process. To reduce the chance of this occurring the slow-cool runs should be

very slow and the starting temperature should be lowered when the starting model is better (e.g., when the starting model is derived from the crystal structure of a molecule with very similar sequence and/or the addition of a relatively small adduct).

5.2. Gradient Descent Methods

Gradient descent methods of optimization use calculus to determine how to shift the parameters in a "down hill" direction. Because the gradient of a function (its first derivatives listed as a vector) is defined as the direction of most rapid increase of the function, one can decrease the function's value by moving the parameters in the opposite direction. This particular method is called steepest descent.

The principal drawback of steepest descent is that the gradient of a function does not indicate how far the parameters must be shifted to reach their optimal values. Although each cycle of steepest descent results in a better set of parameters, a great many cycles might be required to reach the best set.

The only way to overcome this limitation is to make assumptions about the shape of the function in the neighborhood of its optimum value and use higher order derivatives. The most common choice is to assume that the function is shaped like a parabola, which means that derivatives higher than second order are equal to zero.

The optimization method that assumes the function is a parabola and uses all of the second derivatives is called the full-matrix method. The matrix referred to in the name is called the normal matrix and contains all of the second derivatives of the function. This matrix contains a great many derivatives in typical refinement problems and causes this method to demand a great deal of computation to calculate the parameter shifts.

On the continuum displayed in **Fig. 5** the steepest descent method falls on the left end of the "gradient descent methods" range and the full-matrix method is on the right end. Between these extremes lie a great many methods that disregard various subsets of the second derivatives in the normal matrix. These variations include conjugate gradient ***(34)***, preconditioned conjugate gradient ***(35,36)***, and the sparse matrix approximation ***(16)***; listed in order of left to right in **Fig. 5**.

6. Conclusion

Table 1 lists a summary of the properties of the refinement programs discussed in this chapter. The field of macromolecular refinement is blessed with a variety of programs that can be used to improve our structural models. With a firm understanding of the differences between them one should be able to choose the program that best fits the needs of any project.

Table 1
Properties of a Selection of Refinement Programs[a]

Program	Parameters	Function	Method
BUSTER/TNT	xyzb	ML, MLϕ, ML?	PCG
CNS	xyzb, torsion	EE, LS, ML, MLϕ	SA, CG
REFMAC	xyzb, TLS, aniso	LS, ML, MLϕ	Sparse, FM
SHELXL	xyzb, aniso, free	LS	Sparse, FM
TNT	xyzb	LS	PCG
XPLOR	xyzb, torsion	EE, LS, ML, MLϕ	SA, CG

[a]This table lists a summary of the properties of six commonly used refinement programs. The meanings of the various codes are:

Parameters xyzb, position, isotropic B factor, and occupancy; aniso, anisotropic B factor; TLS, group TLS B factors used to generate approximate anisotropic B factors; torsion, only allow variation of angles of rotation about single bonds; free, generalized parameters that can be used to model ambiguity in twinning, chirality, or static conformation.

Function EE, empirical energy; LS, least squares; ML, maximum likelihood using amplitude data; MLϕ, maximum likelihood using experimentally measured phases; ML?, maximum likelihood using envelopes of known composition but unknown structure.

Method SA, simulated annealing; CG, Powell variant conjugate gradient; PCG, preconditioned conjugate gradient; sparse, sparse matrix approximation to the normal matrix; FM, full matrix calculated for normal matrix.

Acknowledgments

This work was supported in part by National Institutes of Health grant GM20066 to B. W. Matthews.

This chapter is based on the article "Introduction to macromolecular refinement," by D. E. Tronrud, published in *Acta Cryst.* (2004). **D60,** 2156–2168, copyright IUCr (2004). Permission has been granted to reproduce parts of this article in this chapter.

References

1. Allen, F. H. (2002) The Cambridge structural database: a quarter of a million crystal structures and rising. *Acta Cryst.* **B58,** 380–388.
2. König, V., Vértesy, L., and Schneider, T. R. (2003) Structure of the a-amylase inhibitor tendamistat at 0.93Å. *Acta Cryst.* **D59,** 1737–1743.
3. Diamond, R. (1971) A real-space refinement procedure for proteins. *Acta Cryst.* **A27,** 436–452.
4. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., et al. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 535–542.
5. Stout, G. H. and Jensen, G. H. (1989) *X-Ray Structure Determination: A Practical Guide, Second ed.* John Wiley and Sons, New York, NY.

6. Kleywegt, G. J. and Jones, T. A. (1995) Where freedom is given, liberties are taken. *Structure* **3,** 535–540.
7. Sheldrick, G. M. and Schneider, T. R. (1997) SHELXL: high-resolution refinement. In: *Macromolecular Crystallography, Part B, Vol. 277,* (Sweet, R. M. and Carter, Jr, C. W., eds.), Academic Press, New York, NY, pp. 319–343.
8. Rice, L. M. and Brünger, A. (1994) Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. *Protein: Structure, Function, and Genetics* **19,** 277–290.
9. Brünger, A. T., Kuriyan, K., and Karplus, M. (1987) Crystallographic R factor refinement by molecular dynamics. *Science* **235,** 458–460.
10. Brünger, A. T., Adams, P. D., Clore, G. M., et al. (1998) Crystallography and NMR system: a new software system for macromolecular structure determination. *Acta Cryst.* **D54,** 905–921.
11. Schomaker, V. and Trueblood, K. N. (1968) On the rigid-body motion of molecules in crystals. *Acta Cryst.* **B24,** 63–76.
12. Haneef, I., Moss, D. S., Stanford, M. J., and Borkakoti, N. (1985) Restrained structure-factor least-squares refinement of protein structures using a vector processing computer. *Acta Cryst.* **A41,** 426–433.
13. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst.* **D53,** 240–255.
14. Winn, M. D., Isupov, M. N., and Murshudov, G. N. (2001) Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Cryst.* **D57,** 122–133.
15. Levitt, M. (1974) Energy refinement of hen egg-white lysozyme. *J. Mol. Biol.* **82,** 393–420.
16. Konnert, J. H. (1976) A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units. *Acta Cryst.* **A32,** 614–617.
17. Mandel, J. (1984) *The Statistical Analysis of Experimental Data.* Dover Publications, New York, NY.
18. Bricogne, G. (1988) A Bayesian statistical theory of the phase problem. I. A multichannel maximum-entropy formalism for constructing generalized joint probability distributions of structure factors. *Acta Cryst.* **A44,** 517–545.
19. Read, R. J. (1990) Structure-factor probabilities for related structures. *Acta Cryst.* **A46,** 900–912.
20. Bricogne, G. (1993) Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. *Acta Cryst.* **A49,** 37–60.
21. Bricogne, G. and Irwin, J. J. (1996) Maximum-likelihood structure refinement: theory and implementation within BUSTER+TNT. In: *Macromolecular Refinement: Proceedings of the CCP4 Study Weekend, January 1996,* (Dodson, E., Moore, M., Ralph, A., and Bailey, S., eds.), Daresbury Laboratory, Warrington, UK, pp. 85–92.
22. Tronrud, D. E., Ten Eyck, L. F., and Matthews, B. W. (1987) An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Cryst.* **A43,** 489–501.

23. Bricogne, G. (1997) Bayesian statistical viewpoint on structure determination: basic concepts and examples. In: *Macromolecular Crystallography, Part A, Vol. 276,* (Sweet, R. M. and Carter, Jr., C. W., eds.), Academic Press, New York, NY, pp. 361–423.
24. Sivia, D. S. (1996) *Data Analysis: A Bayesian Tutorial.* Oxford University Press, Oxford, UK.
25. Wilson, A. J. C. (1942) Determination of absolute from relative X-ray intensity data. *Nature* **150,** 151–152.
26. Wilson, A. J. C. (1949) The probability distribution of X-ray intensities. *Acta Cryst.* **2,** 318–321.
27. Pannu, N. S. and Read, R. J. (1996) Improved structure refinement through maximum likelihood. *Acta Cryst.* **A52,** 659–669.
28. Srinivansan, R. and Parthasarathy, S. (176) *Some Statistical Applications in X-ray Crystallography,* Pergamon Press, Oxford, UK.
29. Read, R. J. (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Cryst.* **A42,** 140–149.
30. Brünger, A. T. (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355,** 472–475.
31. Roversi, P., Blanc, E., Vonrhein, C., Evans, G., and Bricogne, G. (2000) Modelling prior distributions of atoms for macromolecular refinement and completion *Acta Cryst.* **D56,** 1316–1323.
32. Krikpatrick, Jr., S., Gelatt, C. D., and Vecchi, M. P. (1983) Optimization by simulated annealing. *Science* **220,** 671–680.
33. Otten, R. H. J. M. and van Ginneken, L. P. P. P. (1989) *The Annealing Algorithm,* Kluwer Academic Publishers, Boston, MA.
34. Fletcher, R. and Reeves, C. (1964) Function minimization by conjugate gradients. *Computer Journal* **7,** 81–84.
35. Axelsson, O. and Barker, V. A. (1984) Quadratic functionals on finite-dimensional vector spaces. In: *Finite Element Solution of Boundary Value Problems,* Academic Press, Inc., Orlando, FL, pp. 1–63.
36. Tronrud, D. E. (1992) Conjugate-direction minimization: an improved method for the refinement of macromolecules. *Acta Cryst.* **A48,** 912–916.

14

Quality Control and Validation

Gerard J. Kleywegt

Summary

This chapter discusses two important aspects of the structure determination process that are related to the accuracy and reliability of the crystallographic model under investigation. Quality control is defined as the analysis of an intermediate model to identify aspects of it that are unusual in some sense and that could, therefore, be a result of errors in the model building or refinement process. Any such errors need to be fixed, if at all possible, prior to analysis and publication of the model. Validation is the process of assessing the reliability of the final model (or certain aspects of it, e.g., the active site residues) that is about to be analyzed, published, deposited, and possibly used in follow-up studies.

Key Words: Electron density; model bias; model building; protein structure; quality control; Ramachandran plot; refinement; structure; validation; X-ray crystallography.

1. Introduction

Before a model of a macromolecule can be analyzed in light of its biological function, one needs to have a thorough understanding of its reliability, both in terms of the overall structure and of its details *(1–4)*. The results of such an assessment should be reflected in the subsequent analysis. For instance, if a 3.5-Å structure is described, it will in general be inappropriate to discuss hydrogen bonds, let alone to list their lengths with a precision of 0.01 Å. To assist users of the model (biologists, enzymologists, medicinal chemists, and others), the final published model should be validated, i.e., its validity (reliability, accuracy, completeness) should be assessed and expressed quantitatively. A related task is quality control of intermediate models. This entails analysis of the model to identify any features that are unusual in some sense (e.g., in terms of geometry, temperature factors, or fit to the density). Whereas validation is usually concerned with overall statistics (i.e., pertaining to the entire

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

model), quality control is typically performed at the level of individual residues and small molecules (ligands, ions, and so on). Nevertheless, the tools used are often the same. For instance, whereas the percentage of outlier residues in the Ramachandran plot conveys information about the global quality of a protein model, each of the individual outlier residues will be the object of scrutiny in a quality control exercise.

Quality indicators are statistics that are calculated from the model, the data, or both, and that provide information about the quality of the model or the data, or of the fit of the model to the data. Assessment of the quality of the data is really a part of the data-processing stage (*see* Chapter 5) and will not be considered further in this chapter. Quality indicators can be classified in various ways. For instance, calculation of some quality indicators requires only the model as input (e.g., deviations from ideal geometry), whereas others require both the model and the data (e.g., the R-value). Some indicators measure global properties (e.g., the free R-value), whereas others are local (e.g., side chain conformations). Finally, some indicators pertain to properties of the model that are "orthogonal" to the properties that have been refined (e.g., the unrestrained torsion angles ϕ and ψ), whereas others reflect aspects of the model that have been used (explicitly or implicitly) in the refinement process (e.g., bond lengths and angles). The former kind of quality indicators are sometimes called "strong" because they provide independent support for the reliability of the model, whereas the latter are "weak" and only verify that the refinement program has done its work properly ***(3)***. Obviously, strong quality criteria are much more informative for validation purposes than weak ones (although outliers of weak criteria should be scrutinized carefully). For quality control purposes, the objects of interest are the individual residues and small molecules (ligands, solvent molecules, salt ions, cofactors, inhibitors, and so on ***[5]***), and hence *strong* and *local* quality indicators are the most useful. For validation of the final model, on the other hand, *strong* and *global* indicators are the most useful. However, many local indicators also provide information about the global quality of the model e.g., through their averages, variances, Z-scores, or fraction outliers.

The purpose of quality control is to identify residues or small molecules that are unusual in some sense. It is important to realize that unusual model aspects (outliers) can be one of two things: either a genuine *feature* of the *structure*, or an *error* in the *model*. In many cases, the distinction can only be made when the experimental data (in the form of an electron density map) is available, and this is precisely the task of the crystallographer in the quality control process. Any errors need to be fixed using the tools in the modeling program, a process that is known as rebuilding, and that usually proceeds concurrently with the quality control operation.

Table 1
Links to Websites From Which the Materials in Subheading 2. Can be Obtained

Resource	URL
CCP4	http://www.ccp4.ac.uk/main.html
O	http://xray.bmc.uu.se/~alwyn/o_related.html
Protein Data Bank	http://www.pdb.org/
Uppsala Software Factory (OOPS2, MAPMAN, MOLEMAN2)	http://xray.bmc.uu.se/usf
WHAT_CHECK	http://www.cmbi.kun.nl/gv/whatcheck/
WHAT IF	http://www.cmbi.kun.nl/gv/whatif/

In this chapter, the choice of software tools and procedures is heavily, but inevitably, biased by local customs and personal experience and preferences. However, alternatives to the programs and protocols are provided in the Notes section. The chapter covers quality control and validation from the point of view of a practicing crystallographer who is refining or about to analyze and publish a structure. The emphasis is on quality control and validation of protein models, although many of the tools are equally applicable to nucleic acids and other molecules. The theory behind individual quality criteria (Ramachandran analysis, real-space fit scores, rotamer fits, and so on) is not discussed here. The reader is instead referred to an extensive compendium of quality criteria *(3)* and the references to the primary literature provided therein.

One aspect of validation that is beyond the scope of this chapter is validation by nonexperts of models retrieved from public databases, using publicly available (and often web-based) tools. This topic has, however, been discussed elsewhere recently *(6)*. In addition, there is a web-based tutorial available on this subject (http://xray.bmc.uu.se/gerard/embo2001/modval). Other useful sources include **refs.** *2*,*7–12*.

2. Materials

For a list of related websites, *see* **Table 1**. These, and all other links in this chapter, have also been collected in the following webpage: http://xray.bmc.uu.se/gerard/supmat/qualcont. This webpage also provides suitable example files for readers who wish to test the methods described here. For alternatives to some of the programs mentioned next, *see* the cited notes.

1. Atomic model. This is usually the product of the previous refinement cycle, but can also be obtained from elsewhere, e.g., from colleagues or the Protein Data Bank (PDB) *(13*,*14)*.

2. Experimental structure factor data (amplitudes or intensities, experimental sigmas, test-set flags), either the result of one's own data collection and processing, or obtained elsewhere.
3. Collaborative Computational Project Number 4 (CCP4) phase and map calculation software (*see* **Note 1**) *(15)*.
4. WHAT IF (or WHAT_CHECK) validation software (*see* **Note 2**) *(16)*.
5. MAPMAN software (*see* **Note 3**) *(17)*.
6. OOPS2 software (*see* **Note 4**) *(18)*.
7. O validation and model building software (*see* **Note 5**) *(19)*.
8. Computer(s) to run the various software programs (*see* **Note 6**).

3. Methods

3.1. Electron Density Maps

A model and two or more electron density maps are the basic inputs to the validation (and model-building) software, and both are usually obtained as output from a cycle of refinement (*see* Chapter 13). If the refinement program REFMAC *(20)* is used through the CCP4i interface *(21)*, it can be instructed to produce two σ_A-weighted maps *(22)*, one with coefficients ($2mF_o - DF_c$, α_{calc}) and another with coefficients ($mF_o - DF_c$, α_{calc}). The former shows positive density features for properly placed atoms that are already in the model and for strong features that have not yet been included (e.g., missing loops, ligands, waters). The latter provides information about features that have not yet been included but should be, about features that have been included but should not be, and about atoms that need adjusting. As yet unmodeled features will show up with positive density values, whereas features that ought to be removed will show up as negative density. Finally, a combination of positive and negative density near an atom indicates that the atom should be moved in the direction of the positive density (but *see* **Note 7**).

An issue that tends to be confusing is the selection of the appropriate contour level for the various maps. The so-called σ-level of a map is simply the root-mean-square (RMS) value of the electron density in the entire unit cell (or asymmetric unit). Maps of the type ($2mF_o - DF_c$, α_{calc}) are conventionally contoured at a level of "1 σ." In order to identify atoms that contain many electrons (sulfur, phosphorous, metals, chloride, and so on) it is also useful to contour this map at a higher level (e.g., 4 σ) in a different color. However, there are indications that it may actually be a good idea to contour at different levels in different parts of the cell, depending on the local average temperature factor of the model.

It is unfortunate that many crystallographers (and, indeed, referees) fail to recognize that the σ level of ($mF_o - DF_c$, α_{calc}) maps is a meaningless quantity. In the beginning of a refinement, when the model is incomplete and full of (random and other) errors, there is so much signal in such maps that a "2 σ" feature

could well be highly significant. However, toward the end of the refinement process, when the model is largely complete and correct, this type of map will be essentially flat (indeed, flatness of the $mF_o - DF_c$ difference map is a criterion for assessing the convergence of a refinement) and contain only random noise. In this extreme, even a "4 σ" feature need not necessarily be significant. The notion that each and every "3 σ" feature in an $mF_o - DF_c$ difference map must be significant is a fallacy, and engenders a deluge of false water molecules (i.e., noise peaks misinterpreted as water). The best way to prevent this is to avoid the notion of σ levels altogether, at least for $mF_o - DF_c$ difference maps. Instead, if all maps are calculated on an absolute scale (i.e., in terms of electrons per cubic Ångström), the ($mF_o - DF_c$, α_{calc}) map could be contoured at, for instance, one or two times the σ level of the corresponding ($2mF_o - DF_c$, α_{calc}) map (both positive and negative levels should be contoured). Any features visible at such levels would most likely be significant at any stage of the refinement procedure *(23)*.

Apart from σ_A-weighted maps, many other types of maps can (and should) be used when appropriate. For instance, in the presence of noncrystallographic symmetry or multiple, nonisomorphous crystal forms, averaged maps *(24)* can reveal features that may not be visible in the density of any of the individual subunits, or help resolve ambiguities in the tracing. If there is doubt about the tracing of a molecule, or when a poor molecular replacement model is rebuilt and one suspects model bias, various kinds of omit maps (regular, simulated annealing, or complete) can be used *(25–27)*. If an experimental map (MAD, MIR, SIRAS, and others) is available, it should be consulted and checked at later stages, as this map is guaranteed to be unbiased by the model that is being constructed. Similarly, an anomalous difference map may be useful to pinpoint or verify the location of anomalous scatterers in the model (e.g., sulfur or selenium atoms).

3.2. Validation With WHAT IF

Validating a model with WHAT IF (or WHAT_CHECK; *see* **Note 2**) *(16,28–30)* is simple: all that is needed is to start up the program and to type:

check ful model.pdb x y

where "*model.pdb*" needs to be replaced by the name of the PDB file that contains the current model. The program produces a number of files, but the most important one will be called "*pdbout.txt*." This file contains an extensive report (in English) that should be checked carefully (*see* the guide for crystallographers at http://xray.bmc.uu.se/usf/whatif.html). However, the file can also be input to the program OOPS2 (*see* **Subheading 3.4.**), and this program will extract the information about individual residues that have been flagged by WHAT IF.

3.3. Validation With O

Assuming that the program O (v9.0.0 or newer) is used both for calculating several residue-based quality indicators *(31,19)* and for rebuilding *(32,33)*, a number of things need to be done (*see* **Note 8**).

1. Start up the program (without any arguments) and answer all questions by hitting the return-key until the graphics window appears. A new, empty O database has now been created. Save this database to a file, e.g.:

   save model.o

2. Read in the current model, e.g., with the following command:

   pdb_read model.pdb mod y y

   This will read in the file "*model.pdb*," name this model "*mod*" inside O, strip any hydrogen atoms and save any X-PLOR or CNS segment identifiers.
3. Select the molecule, define its cell constants and space group, and draw an all-atom model of it:

   mol mod
   sym_setup mod; p212121
   zone; end

   and subsequently center the molecule on your screen using the mouse.
4. Define the maps that are to be used during rebuilding as well as their contouring levels (*see* **Note 9**). In this example, a ($2mF_o - DF_c$, α_{calc}) map is used that is in a file called "*2mfo_dfc.map*" that will be called "*2fofc*" inside O, and for which the symmetry operators for space group $P2_12_12_1$ should be used (*see* **Note 10**). Subsequently, the contour drawing radius for this map is set to 20 Å, the line type to solid, and two contour levels are defined: 1 σ to be drawn in blue and 4 σ in cyan. In a similar fashion the ($mF_o - DF_c$, α_{calc}) map is defined (but *see* **Note 11**):

   fm_file 2mfo_dfc.map 2fofc p212121
   fm_set 2fofc 20.0 solid 2 1.0 blue 4.0 cyan
   fm_file mfo_dfc.map fofc p212121
   fm_set fofc 20.0 solid 2 3.5 green -3.5 red

5. Calculate the pep-flip value for every amino acid residue (a measure of how unusual the residue's peptide plane orientation is), and save the resulting O data block in a text file (e.g., "*pepflip.o*") for use with OOPS2 (*see* **Note 12**). Assuming that the first and the last amino acid residues are called "A2" and "C135," respectively, the following commands need to be issued:

   pep_flip mod a2 c135
   write mod_residue_pepflip pepflip.o ;

6. Calculate the rotamer side chain fit score for every amino acid residue (a measure of how unusual the side chain conformation of the residue is), and save the resulting O data block in a text file (e.g., "*rsc.o*") for use with OOPS2:

refi_init mod
refi_gen mod ;
rsc mod a2 c135
write mod_residue_rsc rsc.o ;

7. Calculate real-space fit values (a measure of how well each residue fits its surrounding density) for all residues, ligands, water molecules, etc. (for alternatives, *see* **Note 13**). This requires, besides the model, a ($2mF_o - DF_c$, α_{calc}) map, a definition for every residue type of the atoms that should be included in the calculations (*see* **Note 14**), and values for some parameters (*see* **Note 15**). You can choose to calculate the real-space fit value as an R-value ("*rfact*" in the first command) or as a correlation coefficient (in that case, type "*cc*" instead of "*rfact*"):

   rs_fit mod a2 c135 2fofc all rfact
   write mod_residue_rsfit rsfit.o ;
   stop

3.4. Validation and Macro Generation With OOPS2

OOPS2 (the successor of the slightly more cumbersome program OOPS *[18]*) is a program that has multiple purposes in the quality control process (*see also* its on-line manual at http://xray.bmc.uu.se/usf/oops2_man.html). First, given a model, it can carry out a few quality checks of its own (e.g., Ramachandran plot analysis). Second, it can read and parse the results of validation tests carried out with other programs (most notably, WHAT IF, WHAT_CHECK, and O). Most important, however, is the fact that its output facilitates a systematic (and educational) approach to the quality control and rebuilding process. The output consists of a residue-by-residue critique of the model (both in a text file that can be edited during the subsequent manual model rebuilding session, and in an HTML file that can be posted on a website) and a set of macros for the program O. These macros are small files, one for each residue, that contain instructions for O to center on a residue, possibly draw the local electron density, and to print a list of aspects of the residue that are unusual (e.g., an unusual bond angle, unsatisfied hydrogen-bond donor, close contact, or outlier in the Ramachandran plot).

1. Before OOPS2 can be run, a subdirectory called "*oops*" needs to be created (if it does not exist yet), e.g., using the Unix command:

   mkdir oops

2. Now start up the program and answer the questions: do you want OOPS2 to print statistics and histograms, do you want it to generate some plot files, what is the name of your molecule in O, and what is the name of your model's PDB file? OOPS2 will read your model, print some information about it and then present you with a menu of options (*see* **Table 2**). By issuing one of these commands, you include the corresponding quality check in the validation process. By issuing the

Table 2
Menu of Options Presented by OOPS2

POSSIBLE COMMANDS :

PEP = include pep-flip	RSC = include rotamer
RAM = include Ramachandran	WIF = include WHAT IF diagnostics
RSR = include real-space R (all atoms)	CCA = include real-space CC (all)
CCM = include real-space CC (main)	CCS = include real-space CC (side)
BFA = include B-factors	OCC = include occupancies
MSK = include mask fit	DIS = include disulfides
CAC = include CA chirality	PLA = include peptide planarity
PRE = include previous model	
GO = get going !	QUI = quit without doing anything

same command again, the check will be excluded again. Most commands require some additional input, although the default values are usually appropriate. In the following, the most frequently used commands will be discussed briefly (more details can be found in the on-line manual at http://xray.bmc.uu.se/usf/oops2_man.html).

3. PEP: the program will prompt you for the name of the file that contains the O data block with the pep-flip values (e.g., "*pepflip.o*"). If you opted for statistics and histograms to be listed, you will be presented with information about the distribution of the pep-flip values. Finally, you are asked to supply a cut-off value—any residue whose pep-flip value exceeds the cut-off will be considered an outlier, and will, thus, be brought to your attention during the OOPS2-guided inspection or rebuilding of the model. For most models, a cut-off value of 2.5 Å is suitable. OOPS2 will list the number and percentage of outliers (the latter number could be included in a table with validation statistics). Once the protein model does not change too much any longer, the pep-flip check can be omitted, although it should be included again in the final rebuilding round.
4. RSC: provide the name of the data block with rotamer side chain fit values (e.g., "*rsc.o*") and a cut-off value (typically, 1.0 Å). As with the pep-flip check, once the protein model is essentially finished, this check can be omitted, although it should be carried out again in the final rebuilding round.
5. RAM: this command carries out a Ramachandran analysis using the definition of Kleywegt and Jones to assign outliers *(33)*. Optionally, OOPS2 can also flag proline residues with unusual ϕ torsion angles, as well as any nonglycine residues with positive ϕ values. Furthermore, residues that are in a left-handed helical conformation can be flagged.
6. WIF: provide the name of the file created by WHAT IF (normally, this file will be called "*pdbout.txt*"). OOPS2 will read this file and parse the results of more than two dozen of the WHAT IF quality checks.
7. RSR: the real-space R-values can be extracted either from a data block file created by O, or from a file created by MAPMAN (*see* **Note 13**). Select the appropriate

option and provide the corresponding filename. You also need to supply a cut-off value—all residues whose real-space R-value exceeds this cut-off will be flagged by OOPS2. This cut-off value is usually chosen as one or two standard deviations above the average real-space R-value.

8. CCA: this command can be used if you calculated real-space correlation coefficients instead of R-values. The CCM and CCS commands can be used if you also performed the calculations separately for main chain and side chain atoms.
9. BFA: provide a lower and upper threshold for temperature factors (and a separate upper threshold for water molecules if you like). Residues that contain at least one atom with a temperature factor below the lower, or above the upper threshold will be flagged by OOPS2.
10. OCC: similar to the BFA command, in that any residues that contain atoms with unusual occupancies can be flagged by OOPS2. Because some or all atoms in residues with alternative conformations will have nonunit occupancies, using a lower cut-off value of, say, 0.99 is a simple way to get OOPS2 to flag all such residues and, hence, bring them to your attention during the rebuilding session (*see* **Note 16**).
11. PRE: the current model can be compared to (any) previous model. Residues that have undergone important changes (or that have newly been inserted or mutated) will be flagged by OOPS2. Changes in position, temperature factor, occupancy, as well as main chain and side chain torsion angles can all be taken into consideration.
12. When all the appropriate checks have been included, the GO command needs to be issued. Any user-defined criteria can be included at this stage (not normally used). You are also asked to provide a line of O commands that you want to execute for every residue that has been flagged by OOPS2 (*see* **Note 17**). Finally, you are asked three more questions about the O macros that OOPS2 will produce, but the default answers are almost always appropriate.

    The program will now consider every residue in turn and check if it was flagged (e.g., by WHAT IF, or by the real-space fit check, and so on). If so, a message will be printed and an O macro will be generated in the "*oops*" subdirectory (the macro has the same name as the residue). An example of such a macro is shown in **Table 3**. Some of the other files that are created at this stage are:

    a. "*oops.omac*": this is the main O macro that you need to execute when you start up O again and want to go on a journey along all residues that were flagged by OOPS2.
    b. "*mod_oops.html*": this contains a residue-by-residue critique of the model in HTML format.
    c. "*mod_rebuild.notes*": a similar critique, but as a simple text file (*see* **Table 4**). You may find it useful to edit this file as you rebuild your model so as to keep track of the changes you make to the model, any observations you make about the structure, and so on.

3.5. Inspection of Unusual Model Aspects

Inspection of the unusual aspects of the current model is a simple matter. First, start up O again, and provide the name of your previously created O data-

Table 3
Example of an O Macro Generated by OOPS2[a]

```
centre_atom MOD C126 CA
@foreach
print .....
print Residue ASP C126 [Loop or turn        ]
message OOPS - Residue ASP C126 [Loop or turn]
symbol oops_irc    114
print Bad RSC = 1.26
print Bad Phi-Psi = 91.55 -46.75
print NOTE - non-Gly positive PHI = 91.55
print Too high temperature factor = 32.86
print Unusual backbone torsions : 114 ASP ( 126 ) C Poor phi/psi
print Bumps : 113 ASP ( 125 ) C O — 114 ASP ( 126 ) C CB 0.108 2.692
print Unusual backbone conformations : 114 ASP ( 126 ) C 0
print .....
print Hit or type "@oops/c130" for next baddy
menu @oops/c130 on
menu @oops/c126 off
```

[a]This is a macro for one particular residue (in this case "*C126*"). The macro instructs O to center on this residue and to execute a line of user-defined O commands (in this case, the user has chosen to execute a macro: "@foreach"). Subsequently, the macro prints a lot of information about aspects of this residue that are unusual. Finally, it puts a new command on the O user menu, activating this command will take the user to the next flagged residue (in this case, residue "*C130*").

base file (e.g., "*model.o*"). Second, execute the main O macro created by OOPS2:

@oops.omac

This macro prints a list of all the quality checks that were included, and then executes the macro it created for the first flagged residue. For this residue, it will execute the user-defined commands (e.g., to draw the maps), and print a list of criteria that are violated or unusual for this residue. It is then up to you, the crystallographer, to inspect the model and the density and to decide if action is warranted, and if so, what kind of action. This is the process of model rebuilding, which falls outside the scope of this chapter. Unfortunately, even though model rebuilding is a very important part of the structure determination process, there is precious little literature addressing this issue ***(33)***.

Once inspection and possibly rebuilding of a residue is finished, the macro to proceed to the next flagged residue can be executed by clicking the appropriate command on the O user menu (this will be called something like

Table 4
Example of Some of the Contents of the Notebook File Produced by OOPS2

```
Created by OOPS2 V. 021121/1.2.5 at Tue Jul 1 23:41:13 2003 for gerard
Molecule MOD
OOPS has checked:
Pep-flip values; cutoff =   2.5
RS R-factor (all atoms); cutoff =    0.150 ; WATERs =    0.150
RSC values; cutoff =   1.
Too low temperature factors; cutoff =   5.
Too high temperature factors; cutoff =   30.000 ; WATERs =   40.000
Too low occupancies; cutoff =   0.99000001
Too high occupancies; cutoff =   1.
Phi-Psi angle combinations (Ramachandran)
WHAT IF diagnostics

OOPS - ASN A2 [Loop or turn ]
Bad RS R-factor (all atoms) = 0.216
Too high temperature factor = 32.04
H/N/Q side chain flips : 1 ASN ( 2 ) A
COMMENTS/ACTION —>

OOPS - PHE A3 [Loop or turn ]
Bad RSC = 1.37
Bumps : 2 PHE ( 3 ) A N — 44 GLN ( 45 ) A NE2 0.273 2.727
COMMENTS/ACTION —>

OOPS - ALA A4 [Loop or turn ]
Unusual backbone torsions : 3 ALA ( 4 ) A omega poor
Unusual backbone conformations : 3 ALA ( 4 ) A 0
Unsatisfied H-bond donors : 3 ALA ( 4 ) A N
COMMENTS/ACTION —>
```

"@oops/a3", with "*a3*" being the name of the residue in question in this example).

3.6. Validation Statistics

Validation of the final model is essentially an exercise in collecting quality-related statistics from a wide variety of sources. Here, a number of statistics and their usefulness for validation purposes are discussed. For an extensive discussion (including references to the primary literature) of quality indicators, *see* **ref.** *3*.

To assess the quality of the fit of the model to the data, both the conventional and the free R-value *(34,35)* should be reported. In addition, some description of the real-space fit should be included, either qualitatively (e.g., "poor or no density was observed for residue 12 to 21 as well as the side chains of residues 47, 98, and 132") or, preferably, quantitatively. The average, standard deviation, and extremes of the real-space R-value or correlation coefficient can be reported in a table. However, plots of either quantity as a function of residue number are far superior in conveying the information. Such plots should be provided to users of the model as well as referees, even if they are not included in the actual manuscript.

To assess the quality of the model *per se*, results of strong global quality checks need to be reported. For proteins, these include the quality of the Ramachandran plot (including a reference to the definition that was used to derive the score, e.g., **refs.** *30,36–38*) and some measure of the "regularity" of the fold, such as a profile score *(39,40)* or the average DACA score *(28)*. In addition, the percentage of residues with unusual pep-flips or side chain conformations can be reported. The WHAT IF report file ("*pdbout.txt*") contains a number of useful (strong and global) statistics under the heading "*structure Z-scores.*" If noncrystallographic symmetry is present, many statistics can be used to express the degree of similarity of the models *(41)* (pertaining to positions, torsion angles, and temperature factors) and even of their densities *(42)*.

Other statistics that are often reported, but that do not necessarily convey much information about the correctness of a model, include RMS deviations from ideal values of geometric quantities (bond lengths, bond angles, and so on *[43–45]*), average temperature factors, number of atoms or refined parameters, coordinate error estimates, and so on.

Finally, it should go without saying that if a model leads to a scientific publication, both the model and the experimental data should be deposited in the public structural database, the Protein Data Bank *(23,46)*.

4. Notes

1. Other programs that can be used to calculate electron density maps include X-PLOR *(47)*, CNS *(48)*, TNT *(49)*, XtalView/Xfit *(50)*, and SHELX *(51)*.
2. Many other programs exist that produce validation information, but we find that WHAT IF is the most comprehensive of these (for an annotated listing of its quality checks, *see* http://www.cmbi.kun.nl/gv/pdbreport/checkhelp/). A subset of WHAT IF that only contains the validation functionality is available free of charge under the name WHAT_CHECK, and this program can in principle be used instead of WHAT IF. Alternative programs include PROCHECK *(36)*, DDQ *(52)*, MolProbity *(38)*, Verify3D *(39,40)*, ERRAT *(53,40)*, MOLEMAN2 *(54)*, and NUCHECK (specifically for nucleic acid models) *(55)*.

3. MAPMAN is used to convert electron density maps between different formats (e.g., to convert CCP4 or CNS maps into O format). It can also be used to calculate residue-based real-space fit statistics (*see* **Note 13**).
4. OOPS2 (and its predecessor OOPS) generate and use residue-based quality information to generate a set of macros for O. When executed, these macros will take the crystallographer on a journey along all residues that are unusual in some sense or other.
5. Alternatives for O as a model building program include various derivatives of FRODO *(56)*, XtalView/Xfit *(50)*, and Quanta *(57)*. Note, however, that OOPS2 only works in conjunction with O.
6. Nowadays, almost all crystallographic software can be run on comparatively inexpensive personal computers running Unix-like operating systems, such as Linux and Mac OS X, equipped with graphics cards to enable the use of interactive graphics programs such as O.
7. Positive and negative peaks in such maps may also have other causes. In general, positive density indicates that the model contains too little scattering matter locally. This could also be caused by the temperature factors being too high, or even incorrect assignment of an atom type (e.g., oxygen instead of sulfur, or sodium instead of potassium).
8. For more information about the syntax, usage, or purpose of specific O commands, please consult "A-to-Z of O" (http://xray.bmc.uu.se/alwyn/A-Z_of_O/A-Z_frameset.html).
9. In fact, you will probably want to execute these commands every time you start up O with your current database. A simple way to accomplish this is to use a text editor to create a small file in your directory which you call "*on_startup*" and which contains these commands (and any others you want to execute automatically when you start up O).
10. It is important to keep in mind if the map that you use comprises an asymmetric unit or unit cell, or whether it contains some other part of the cell (e.g., cut out around the molecule). The former type of map is strongly preferred in O. In that case, you can provide the name of the space group, and O will use the appropriate symmetry operators to determine the density values outside the part of space that is covered by the map. Moreover, the σ level of the map will be that of the unit cell. If you use maps that contain an arbitrary part of space, you should not provide the space group name, because map expansion may fail. In addition, the σ level of such a map is not equal to that of the unit cell and hence contour levels need to be adjusted accordingly. For instance, if the RMS density level in the unit cell is 0.36 e/Å^3, and that in a map carved out around the molecule is 0.30 e/Å^3, then the latter must be contoured at a level of 0.36/0.30 = 1.2 "σ" in order to portray the density at the unit cell RMS level.
11. The contour levels that O uses are expressed in terms of the RMS density values in the corresponding map file. In order to draw the ($mF_o - DF_c$, α_{calc}) map at a level that is equivalent to the σ level of the ($2mF_o - DF_c$, α_{calc}) map, we need to divide the absolute σ level of the latter by that of the former. For instance, if the σ level

of the ($2mF_o - DF_c$, α_{calc}) map is 0.344 e/Å^3 and that of the ($mF_o - DF_c$, α_{calc}) map is 0.097 e/Å^3, then the proper contour level for the latter in O is 0.344/0.097, which is approx 3.5 σ units. Note that the RMS (or σ) level of a map is printed by O when you open the map with the "*fm_file*" command (look for a line that says "Min, max, sigma").

12. The steps involving the calculation of pep-flip, rotamer side chain fit, and real-space fit values can also be carried out with the help of an O macro (ftp://xray.bmc.uu.se/pub/gerard/omac/pre_oops.omac). Note that each of these commands automatically creates (or overwrites) a data block in O's database. The name of the data block will be the name of the molecule ("*mod*," in the example), followed by the string "*_residue_*" and finally the name of the property (e.g., "*pepflip*").
13. Real-space fit calculations can also be carried out with the program MAPMAN (use its RS_fit command) and can then still be used with OOPS2 (*see* the on-line MAPMAN manual at http://xray.bmc.uu.se/usf/mapman_man.html). Note that this calculation requires two maps as input, namely a ($2mF_o - DF_c$, α_{calc}) and an (F_c, α_{calc}) map, and that both maps must cover the model. Other programs that calculate real-space fit values are CNS and SFCHECK ***(58)***.
14. The atoms that are to be included in the real-space fit calculations are defined in the major dictionary file for O, "*stereo_chem.odb*." If there are any entities in your model that are not yet in the dictionary file, they must be added to it. However, in future versions of O this will not be necessary any longer.
15. In order to calculate real-space fit values, O needs to compute a map based on the atomic model alone and compare that to the external map that you provide (*see* **refs.** ***59*** and ***60*** and references cited therein). There are two parameters in these formulas that may need adjusting, called A0 and C. The default values for these parameters (0.9 and 1.04, respectively) tend to work well if the resolution is around 1.8 Å. At different resolutions, these parameters can be optimized as follows:
    a. Find a residue that fits the density very well and one that fits very poorly (by eye or using the default values of A0 and C and calculating the real-space correlation coefficient).
    b. A0 can usually be kept constant at 0.9.
    c. Vary the value of C between 0.5 and 1.2 in steps of 0.05 and calculate the real-space correlation coefficient for the zones around the good and the bad residues (include two residues at both sides, e.g., if "*a69*" is your good residue, do "*rs_fit mod a67 a71*").
    d. The best value for C is that which gives the largest difference between the real-space fit values for the good and the bad residues.

    Once you have obtained proper values for the parameters A0 and C (for instance 0.9 and 0.8), you can set them to these values either by using the "*rsr_setup*" command in O or, quicker, by typing:

    db_set_dat .rsr_real 8 8 0.9
    db_set_dat .rsr_real 7 7 0.8

16. To check the labeling, positional degeneracy, and summed occupancies of atoms that occur in multiple conformations, you can use the program MOLEMAN2 (use the "*PDb SAnity*" command). *See* the on-line manual for more details (http://xray.bmc.uu.se/usf/moleman2_man.html).
17. It is usually easiest to execute an O macro instead of typing all desired commands here. This macro is simply a text file that you create and that contains all those commands (e.g., to draw the maps, to draw a sphere of residues, to save your database, to generate symmetry-related molecules, and so on). In that case your line of O commands is simply an "@"-sign followed by the name of your macro file (e.g., "@foreach.omac"). This macro file could contain commands such as:

    fm_draw fofc
    fm_draw 2fofc
    sym_sph mod ; 10
    bell
    save

Acknowledgments

The author would like to thank Emma Jakobsson for assistance with some refinement and map calculations, Sara Nystedt for acting as guinea-pig and trying out the methods described in this chapter, and Alwyn Jones, Emma Jakobsson, and Sara Nystedt for useful comments on the chapter.

The author is a Royal Swedish Academy of Sciences (KVA) Research Fellow, supported through a grant from the Knut and Alice Wallenberg Foundation. He is supported by KVA, Uppsala University, the Linnaeus Centre for Bioinformatics, and the Swedish Structural Biology Network (SBNet).

References

1. Brändén, C. I. and Jones, T. A. (1990) Between objectivity and subjectivity. *Nature* **343,** 687–689.
2. Kleywegt, G. J. and Jones, T. A. (1995) Where freedom is given, liberties are taken. *Structure* **3,** 535–540.
3. Kleywegt, G. J. (2000) Validation of protein crystal structures. *Acta Crystallogr.* **D56,** 249–265.
4. Davis, A. M., Teague, S. J., and Kleywegt, G. J. (2003) Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem. Int. Ed.* **42,** 2718–2736.
5. Kleywegt, G. J., Henrick, K., Dodson, E. J., and van Aalten, D. M. F. (2003) Pound-wise but penny-foolish: How well do micromolecules fare in macromolecular refinement? *Structure* **11,** 1051–1059.
6. Kleywegt, G. J. and Hansson, H. (2005) Retrieval and validation of structural information. In: *Structural Genomics and High Throughput Structural Biology* (Sundström, M., Norin, M., and Edwards, A., eds.), Taylor and Francis, Boca Raton, FL, pp. 185–222.

7. MacArthur, M. W., Laskowski, R. A., and Thornton, J. M. (1994) Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy. *Curr. Opin. Struct. Biol.* **4,** 731–737.
8. Zou, J. Y. and Mowbray, S. L. (1994) An evaluation of the use of databases in protein structure refinement. *Acta Crystallogr.* **D50,** 237–249.
9. Kleywegt, G. J. and Jones, T. A. (1995) Braille for pugilists. In: *Making the Most of Your Model,* (Hunter, W. N., Thornton, J. M., and Bailey, S., eds.), SERC Daresbury Laboratory, Warrington, UK, pp. 11–24.
10. EU 3-D Validation Network (1998) Who checks the checkers? Four validation tools applied to eight atomic resolution structures. *J. Mol. Biol.* **276,** 417–436.
11. Laskowski, R. A., MacArthur, M. W., and Thornton, J. M. (1998) Validation of protein models derived from experiment. *Curr. Opin. Struct. Biol.* **8,** 631–639.
12. Laskowski, R. A. (2003) Structural quality assurance. In: *Structural Bioinformatics,* (Bourne, P. E. and Weissig, H., eds.), Wiley-Liss, Hoboken, NJ, pp. 273–303.
13. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., et al. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 535–542.
14. Berman, H. M., Westbrook, J., Feng, Z., et al. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242.
15. Collaborative Computational Project, Nr. 4 (1994) The *CCP4* suite: programs for protein crystallography. *Acta Crystallogr.* **D50,** 760–763.
16. Hooft, R. W. W., Vriend, G., Sander, C. and Abola, E. E. (1996) Errors in protein structures. *Nature* **381,** 272–272.
17. Kleywegt, G. J. and Jones, T. A. (1996) xdlMAPMAN and xdlDATAMAN - programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. *Acta Crystallogr.* **D52,** 826–828.
18. Kleywegt, G. J. and Jones, T. A. (1996) Efficient rebuilding of protein structures. *Acta Crystallogr.* **D52,** 829–832.
19. Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjeldgaard, M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr.* **A47,** 110–119.
20. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr.* **D53,** 240–255.
21. Potterton, E., Briggs, P., Turkenburg, M., and Dodson, E. (2003) A graphical user interface to the CCP4 program suite. *Acta Crystallogr.* **D59,** 1131–1137.
22. Read, R. J. (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr.* **A42,** 140–149.
23. Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wählby, A., and Jones, T. A. (2004) The Uppsala electron-density server. *Acta Crystallogr.* **D60,** 2240–2249.
24. Kleywegt, G. J. and Read, R. J. (1997) Not your average density. *Structure* **5,** 1557–1569.
25. Bhat, T. N. (1988) Calculation of an OMIT map. *J. Appl. Crystallogr.* **21,** 279–281.
26. Hodel, A., Kim, S. H., and Brünger, A. T. (1992) Model bias in macromolecular crystal structures. *Acta Crystallogr.* **A48,** 851–858.

27. Vellieux, F. M. D. and Dijkstra, B. W. (1997) Computation of Bhat's OMIT map with different coefficients. *J. Appl. Crystallogr.* **30,** 396–399.
28. Vriend, G. and Sander, C. (1993) Quality control of protein models: directional atomic contact analysis. *J. Appl. Crystallogr.* **26,** 47–60.
29. Hooft, R. W. W., Sander, C., and Vriend, G. (1996) Verification of protein structures: side-chain planarity. *J. Appl. Crystallogr.* **29,** 714–716.
30. Hooft, R. W. W., Sander, C. and Vriend, G. (1997) Objectively judging the quality of a protein structure from a Ramachandran plot. *Comput. Applic. Biosci.* **13,** 425–430.
31. Kleywegt, G. J. and Jones, T. A. (1998) Databases in protein crystallography. *Acta Crystallogr.* **D54,** 1119–1131.
32. Jones, T. A. and Kjeldgaard, M. (1997) Electron density map interpretation. *Methods Enzymol.* **277,** 173–208.
33. Kleywegt, G. J. and Jones, T. A. (1997) Model-building and refinement practice. *Methods Enzymol.* **277,** 208–230.
34. Brünger, A. T. (1992) Free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355,** 472–475.
35. Kleywegt, G. J. and Brünger, A. T. (1996) Checking your imagination: applications of the free R value. *Structure* **4,** 897–904.
36. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26,** 283–291.
37. Kleywegt, G. J. and Jones, T. A. (1996) Phi/Psi-chology: Ramachandran revisited. *Structure* **4,** 1395–1400.
38. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, et al. (2003) Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins Struct. Funct. Genet.* **50,** 437–450.
39. Eisenberg, D., Lüthy, R., and Bowie, J. U. (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol.* **277,** 396–404.
40. Dym, O., Eisenberg, D., and Yeates, T. O. (2001) Detection of errors in protein models. In: *International Tables for Crystallography. Volume F. Crystallography of Biological Macromolecules* (Rossmann, M.G. and Arnold, E., eds.), Kluwer, Dordrecht, The Netherlands, pp. 520–525.
41. Kleywegt, G. J. (1996) Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallogr.* **D52,** 842–857.
42. Kleywegt, G. J. (1999) Experimental assessment of differences between related protein crystal structures. *Acta Crystallogr.* **D55,** 1878–1884.
43. Engh, R. A. and Huber, R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr.* **A47,** 392–400.
44. Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T., and Berman, H. M. (1996) New parameters for the refinement of nucleic acid-containing structures. *Acta Crystallogr.* **D52,** 57–64.
45. Engh, R. A. and Huber, R. (2001) Structure quality and target parameters. in *International Tables for Crystallography. Volume F. Crystallography of Biological Macromolecules*, (Rossmann, M. G., and Arnold, E., eds.), Kluwer, Dordrecht, The Netherlands, pp. 382–392.

46. Jones, T. A., Kleywegt, G. J. and Brünger, A. T. (1996) Storing diffraction data. *Nature* **381,** 18–19.
47. Brünger, A. T., Kuriyan, J., and Karplus, M. (1987) Crystallographic *R* factor refinement by molecular dynamics. *Science* **235,** 458–460.
48. Brünger, A. T., Adams, P. D., Clore, G. M., et al. (1998) Crystallography and NMR System: a new software suite for macromolecular structure determination. *Acta Crystallogr.* **D54,** 905–921.
49. Tronrud, D. E. (1997) The TNT refinement package. *Methods Enzymol.* **277,** 306–319.
50. McRee, D. E. (1999) XtalView/Xfit - a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125,** 156–165.
51. Sheldrick, G. M. and Schneider, T. R. (1997) SHELXL: high-resolution refinement. *Methods Enzymol.* **277,** 319–344.
52. Van den Akker, F. and Hol, W. G. J. (1999) Difference density quality (DDQ): a method to assess the global and local correctness of macromolecular crystal structures. *Acta Crystallogr.* **D55,** 206–218.
53. Colovos, C. and Yeates, T. O. (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Prot. Sci.* **2,** 1511–1519.
54. Kleywegt, G. J., Zou, J. Y., Kjeldgaard, M., and Jones, T. A. (2002) Around O. *International Tables for Crystallography, Volume F, Crystallography of Biological Macromolecules* (Rossmann, M. G. and Arnold, E., eds.), Kluwer, Dordrecht, The Netherlands, pp. 353–367.
55. Das, U., Chen, S., Fuxreiter, M., et al. (2001) Checking nucleic acid crystal structures. *Acta Crystallogr.* **D57,** 813–828.
56. Jones, T. A. (1978) A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.* **11,** 268–272.
57. Oldfield, T. J. (2001) A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent. *Acta Crystallogr.* **D57,** 82–94.
58. Vaguine, A. A., Richelle, J., and Wodak, S. J. (1999) *SFCHECK*: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr.* **D55,** 191–205.
59. Zou, J. Y. and Jones, T. A. (1996) Towards the automatic interpretation of macromolecular electron-density maps: qualitative and quantitative matching of protein sequence to map. *Acta Crystallogr.* **D52,** 833–841.
60. Jones, T. A. and Liljas, L. (1984) Crystallographic refinement of macromolecules having non-crystallographic symmetry. *Acta Crystallogr.* **A40,** 50–57.

15

Crystallographic Software

A Sustainable Resource for the Community

Stephen J. Everse and Sylvie Doublié

Summary

Virtually all software is constantly changing and evolving (and crystallographic software is no exception), which makes it nearly impossible to write a chapter that will remain current. In this chapter, we introduce CRYSTAL, a website (http://crystal.uvm.edu) where a comprehensive list of available crystallographic packages for each step of macromolecular structure determination will be maintained. Additionally, we provide links to books, journals, and a number of educational sites on crystallography. For each program/site included at CRYSTAL, a detailed description will include: the most current version, the authors, and appropriate operating systems to host the program. Links to download the programs, available tutorials, and references to cite when using the program/site are also provided.

Key Words: Crystallographic software; crystal structure; macromolecular crystallography.

1. Introduction

Solving a crystal structure generally takes much less time than even a decade ago. It is not unusual for researchers to collect data at the synchrotron, solve the phase problem, and return to their home institution with a refined structure. For example, nearly a dozen crystal structures were solved on-site by students during the 2 d of data collection at RapiData 2005 (http://www.px.nsls.bnl.gov/RapiData2005/) at the National Synchrotron Light Source. This accelerated pace is also reflected in the exponential growth of deposited structures in the Protein Data Bank *(1,2)* (PDB; http://www.pdb.org). Interestingly, the structures deposited by the structural genomics groups now account for 20% of the weekly PDB releases *(3)*. Several technological advances are responsible for the formidable growth the PDB is experiencing, including the availability of relatively cheap computational resources, the increase in the number of beamlines available for

From: *Methods in Molecular Biology, vol. 364: Macromolecular Crystallography Protocols: Volume 2: Structure Determination* Edited by: S. Doublié © Humana Press Inc., Totowa, NJ

macromolecular crystallography, and constant improvement in both the intensity of the beam and the user friendliness of the beamlines *(4,5)*. In fact, some beamlines are now equipped with crystal-mounting robots and expert help around the clock *(6)*. Another component is the fact that crystallographic software is widely available, generally free for academic labs, and frequently able to run on a variety of operating systems.

A chapter that attempts to describe the newest and most popular programs available to solve macromolecular structures is doomed to be outdated before publication. Therefore, we opted to generate a database of software that can be regularly updated. Our site, CRYSTAL, is available at http://crystal.uvm.edu. In addition to cataloging the software necessary to solve, evaluate, or generate figures, we will provide links to popular publications (e.g., **refs.** *7* and *8*) and to educational resources dedicated to understanding crystallography, such as X-ray 101 (http://ruppweb.dyndns.org/Xray/101index.html) (*see* **Note 1**). Not only do these sites provide nice images for teaching, many are arranged as self-paced tutorials appropriate for new students. As UNIX continues to be the most popular OS platform for crystallographic software, we also link to several sites for teaching the fundamentals required to become proficient with UNIX. For small molecule structure determination please refer to Sincris (http://journals.iucr.org/sincris-top/logiciel/), which maintains a list for both macromolecular and small molecule software packages.

2. Details

The format of our homepage (http://crystal.uvm.edu) provides menus on the left panel to allow easy access to all aspects of macromolecular crystallography (**Fig. 1**). We have cataloged our information within the following major headings: Software, Publications, Organizations, Synchrotrons, and Educational. Each major heading is subdivided into several categories to make searching more convenient. On selecting a category a set of links and a short description of each site/software is presented. When appropriate, this page will also refer users to a specific chapter in the book. On selecting a site/software link, a page with details will be provided (**Fig. 2**) including: a longer description, current version, operating systems supported, links to documentation (main homepage, tutorial, documentation and software websites), and up to three recommended citations (*see* **Note 2**).

The central panel of the homepage allows one to perform Google® searches (worldwide or within our site) as well as an update of the latest news from scientific publications and societies including Nature Structural and Molecular Biology, Structure, IUCr journals, AAAS, and so on using RSS (rich/site summary) newsfeeds. Most of these sites update their headlines once per day, thus providing instantaneous access to the newest information. Planned upgrades

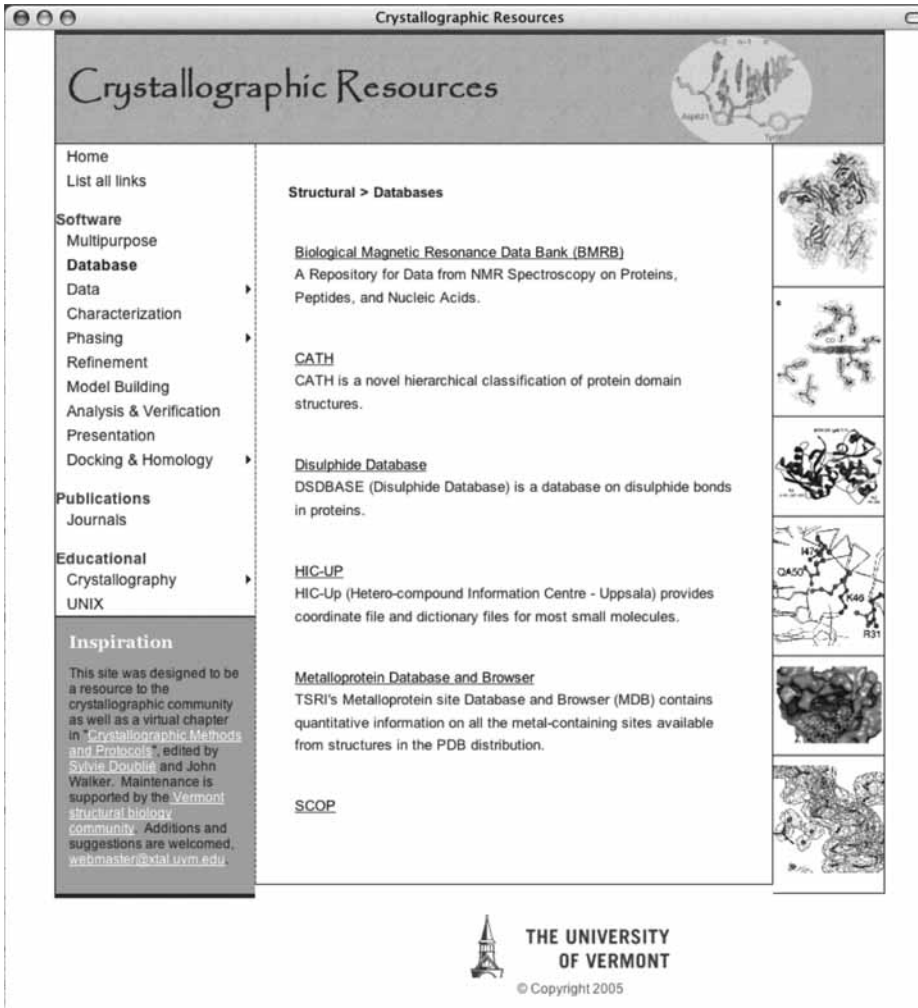


Fig. 1. Crystallographic resources at http://crystal.uvm.edu. Selections from the menu (left) place short descriptions of the resources into the center panel. Resource titles link to a details page (*see* **Fig. 2**).

include a threaded newsreader to allow contact with common crystallographic news groups (ccp4bb, cnsbb, o-info, and so on) without filling your inbox (*see* **Note 3**).

E-mail addresses are provided to allow visitors to inform us about programs we have missed, errors we may have introduced, and any other comments or suggestions helpful to the crystallographic community. Planned upgrades in this area include the ability for visitors to add and update software.

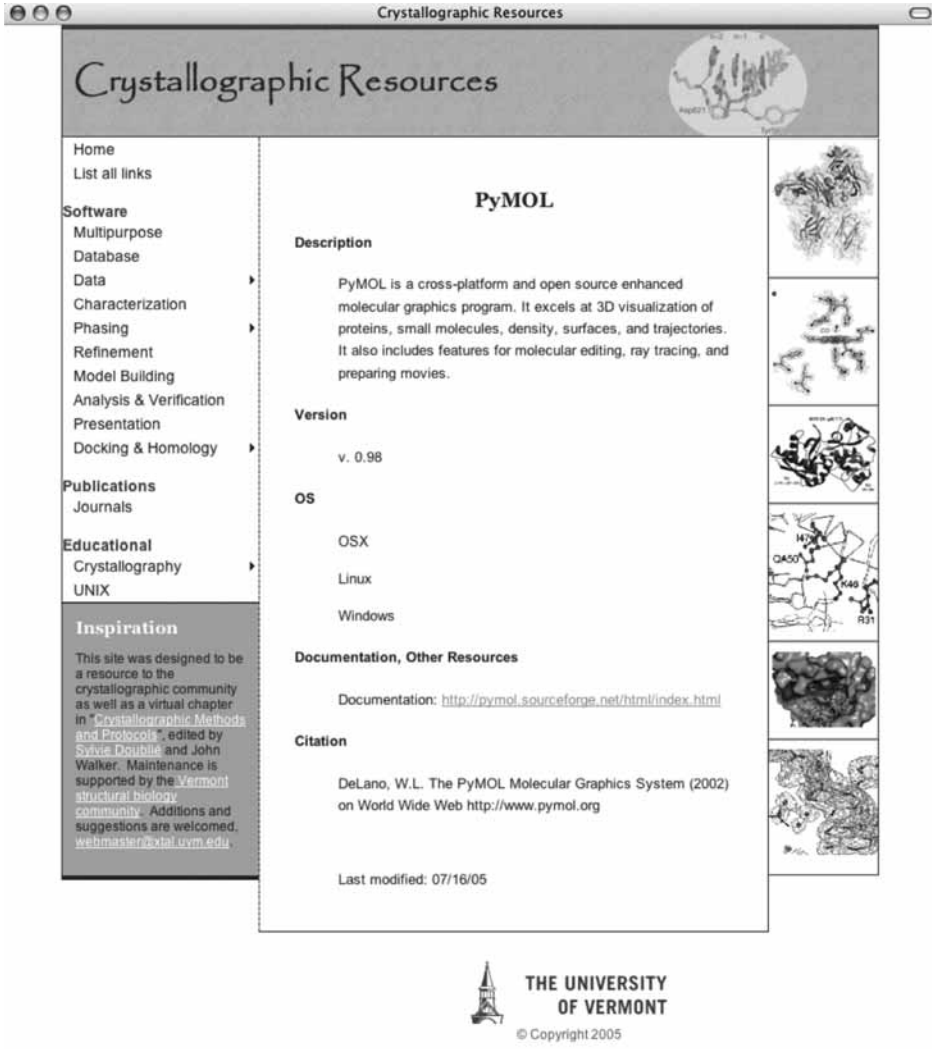


Fig. 2. Resource details. Each resource at the site links to a page providing such information as version number, operating system requirements, available documentation, tutorial sites, and recommended citations.

3. Conclusions

Our goal in building this database is to provide a resource for the crystallographic community. In populating the database we were excited to see how many new programs have been added in just the last few years. Some of these programs were the results of just a few people, whereas others were the result

of collaborations across countries and continents. We hope you find this site useful and, if not, communicate to us how we can make it better.

4. Notes

1. Several companies have started providing general crystallographic information on their websites. Hampton Research (Aliso Viejo, CA) deserves a special mention for their technical support page (http://www.hamptonresearch.com/support/) and their informative newsletter (CrystalNews) that provide details for the beginner as well as the advanced crystallographer on how to grow and improve crystals.
2. There was a time when to be a crystallographer required one to be a computer programmer. In today's crystallographic world, programming is no longer a required skill (although it is certainly handy). It is crucial, however, to have an understanding of a UNIX-like operating system to be able to compile and/or install software. To help people overcome the need to understand compiling options, William G. Scott (University of California, Santa Cruz, CA) has built an excellent site (http://xanana.ucsc.edu/xtal/) for Mac OS X users. Not only does he provide step-by-step instructions for installing many programs listed on our site using fink®, he provides instructions on what to change to compile much of the software. Also in addition, RedHat and Debian LINUX users can APT (Advanced Packaging Tool) or YUM (Yellow dog Updater Modified) many crystallographic packages at Morten Kjeldgaard's (Aarhus University, Aarhus, Denmark) RPM (RPM Package Manager) repository (http://apt.bioxray.dk/).
3. Mailing lists (ccp4bb, cnsbb, o-info, pymol, sharp, solve, and so on) continue to be an important tool for disseminating information, as well as a resource for confirming software bugs and learning about specific packages. We believe these are extremely valuable resources and should be read before posting because often the answer to a question is already available. All the lists previously mentioned are active and frequently you will find posts from the software authors themselves. Many of these lists can be subscribed to at: http://asda.bio.bnl.gov/cgi-bin/bb/bb_info.pl.

Acknowledgments

Support from National Institutes of Health grant GM62239 and the Human Frontier Science Program Organization (SD) is gratefully acknowledged. The structural biology initiative at the University of Vermont is supported by a DOE EPSCoR grant (USDOE DE-FG02-00ER45828) to Susan Wallace. Finally, we would like to take the opportunity here to thank all of the software authors for their hard work and dedication, and for making our structure determinations a lot easier.

References

1. Berman, H. M., Westbrook, J., Feng, Z., et al. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242.

2. Dutta, S. and Berman, H. M. (2005) Large macromolecular complexes in the Protein Data Bank: a status report. *Structure (Camb)* **13,** 381–388.
3. Laskowski, R. A., Watson, J. D., and Thornton, J. M. (2005) Protein function prediction using local 3D templates. *J. Mol. Biol.* **351,** 614–626.
4. Abola, E., Kuhn, P., Earnest, T., and Stevens, R. C. (2000) Automation of X-ray crystallography. *Nat Struct Biol* **7,** 973–977.
5. Jiang, J. and Sweet, R. M. (2004) Protein Data Bank depositions from synchrotron sources. *J. Synchrotron Radiation* **11,** 319–327.
6. Arzt, S., Beteva, A., Cipriani, F., et al. (2005) Automation of macromolecular crystallography beamlines. *Prog. Biophys. Mol. Biol.* **89,** 124–152.
7. Blundell, T. L. and Johnson, L. N. (1976) *Protein Crystallography.* Academic Press, New York, NY.
8. Ding, J. and Arnold, E. (2001) Macromolecular crystallography programs. In: *International Tables for Crystallography, Vol. F,* (Rossmann, M. G. and Arnold, E., eds.), Kluwer Academic Publishers, Dordrecht, Germany, pp. 685–694.

Index