INTERNATIONAL TABLES FOR CRYSTALLOGRAPHY

Volume F

CRYSTALLOGRAPHY OF BIOLOGICAL MACROMOLECULES

Edited by

E. ARNOLD, D. M. HIMMEL AND M. G. ROSSMANN

Second Edition

Published for

THE INTERNATIONAL UNION OF CRYSTALLOGRAPHY

by

John Wiley & Sons, Ltd

2011

Dedicated to
DAVID C. PHILLIPS

The editors wish to express their special thanks to David Phillips (Lord Phillips of Ellesmere) and Professor Louise Johnson for contributing an exceptional chapter on the structure determination of hen egg-white lysozyme. Although Chapter 25.1 describes the first structural investigations of an enzyme, the procedures used are still as fresh and important today as they were 45 years ago, and this chapter is strongly recommended to students of both crystallography and enzymology. Completion of this chapter was David's last scientific accomplishment only a few weeks before his death. This volume of *International Tables for Crystallography* is dedicated to the memory of David C. Phillips in recognition of his pivotal contributions to the foundations of the crystallography of biological macromolecules.

Advisors and Advisory Board

Contributing authors

P. D. Adams: The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [18.2, 18.6, 18.11]

P. V. Afonine: Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [18.11]

F. H. Allen: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [22.5]

W. B. Arendall: Department of Biochemistry, Duke University, Durham, NC 27710, USA. [21.6]

U. W. Arndt: Laboratory of Molecular Biology, Medical Research Council, Hills Road, Cambridge CB2 2QH, England.† [6.1]

E. Arnold: Biomolecular Crystallography Laboratory, Center for Advanced Biotechnology and Medicine & Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854-5638, USA. [1.1, 1.4, 13.4]

P. Auffinger: Institut de Biologie Moléculaire et Cellulaire, Centre National de la Recherche Scientifique, Unité Propre de Recherche 9002, Université de Strasbourg, 15 Rue René Descartes, F-67084 Strasbourg, France. [23.6]

E. N. Baker: School of Biological Sciences, University of Auckland, Private Bag 92-109, Auckland, New Zealand. [22.3]

T. S. Baker: Departments of Chemistry & Biochemistry and Molecular Biology, University of California, San Diego, 9500 Gilman Drive MC-0378, La Jolla, CA 92093-0378, USA. [19.6]

W. Baxter: Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, Albany, NY 12201-0509, USA. [19.8]

C. Bebeacua: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England. [19.9]

C. G. van Beek: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA.‡ [11.5]

J. A. Bell: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

J. Berendzen: Biophysics Group, Mail Stop D454, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [14.3]

H. M. Berman: Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087, USA. [21.2, 24.1, 24.2]

C. C. F. Blake: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.§ [25.1]

D. M. Blow: Biophysics Group, Blackett Laboratory, Imperial College of Science, Technology & Medicine, London SW7 2BW, England.† [13.1]

T. L. Blundell: Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge, CB2 1GA, England. [12.1]

R. Bolotovsky: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA.¶ [11.5]

D. Borek: UT Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75390-9038, USA. [11.4]

G. Bricogne: Laboratory of Molecular Biology, Medical Research Council, Cambridge CB2 2QH, England. [16.2]

A. T. Brunger: The Howard Hughes Medical Institute, and Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences, and Stanford Synchrotron Radiation Laboratory, Stanford Universty, 1201 Welch Road, MSLS P210, Stanford, CA 94305-5489, USA. [18.2, 18.6]

† Deceased.

‡ Present address: RJ Lee Instruments, 515 Pleasant Valley Road, Trafford, PA 15085, USA.

§ Present address: Kent House, 19 The Warren, Cromer, Norfolk NR27 0AR, England.

¶ Present address: Philips Analytical Inc., 12 Michigan Drive, Natick, MA 01760, USA.

G. Bunkóczi: Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge CB2 0XY, England. [18.11]

Y. Cao: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

H. L. Carrell: The Institute for Cancer Research, The Fox Chase Cancer Center, Philadelphia, PA 19111, USA. [5.1]

D. Carvin: Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Field, London WC2A 3PX, England. [12.1]

R. Chandrasekaran: Whistler Center for Carbohydrate Research, Purdue University, West Lafayette, IN 47907, USA. [19.5]

M. S. Chapman: Department of Chemistry & Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306-4380, USA. [22.2]

L. Chen: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

V. B. Chen: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.5, 21.6]

W. Chiu: Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030, USA. [19.2]

K. H. Choi: Department of Biochemistry and Molecular Biology, 6.614C Basic Science, The University of Texas Medical Branch, University Blvd, Galveston, TX 77555-0647, USA. [4.4]

J. C. Cole: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [22.5]

M. L. Connolly: 1259 El Camino Real 184, Menlo Park, CA 94025, USA. [22.2]

C. Cork: Structural Proteomics Development Group, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [9.2]

K. Cowtan: Department of Chemistry, University of York, York YO10 5DD, UK. [15.1, 15.3, 17.1]

D. W. J. Cruickshank: Chemistry Department, UMIST, Manchester M60 1QD, England.† [18.5]

J. de la Cruz: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

M. Cymborowski: Department of Molecular Physiology and Biological Physics, University of Virginia, 1300 Jefferson Park Avenue, Charlottesville, VA 22908, USA. [11.4]

V. M. Dadarlat: Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, Indiana 47907-1333, USA. [20.2]

U. Das: Unité de Conformation de Macromolécules Biologiques, Université Libre de Bruxelles, avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium. [21.2]

Z. Dauter: NCI Frederick & Argonne National Laboratory, Building 202, Argonne, IL 60439, USA. [9.1, 18.4]

I. W. Davis: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA.†† [18.11]

T. Day: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

W. L. DeLano: Graduate Group in Biophysics, Box 0448, University of California, San Francisco, CA 94143, USA. [18.6]

Z. S. Derewenda: Department of Molecular Physiology and Biological Physics, University of Virginia School of Medicine, Charlottesville, VA 22908-0736, USA. [4.3]

† Deceased.

†† Present address: GrassRoots Biotechnology, 598 Airport Boulevard, Morrisville, NC 27560, USA.

CONTRIBUTING AUTHORS

R. E. Dickerson: Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095-1570, USA. [23.4]

J. Drenth: Laboratory of Biophysical Chemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands. [2.1]

S. Dutta: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

O. Dym: UCLA–DOE Laboratory of Structural Biology and Molecular Medicine, UCLA, Box 951570, Los Angeles, CA 90095-1570, USA. [21.3]

T. Earnest: Structural Proteomics Development Group, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [9.2]

N. Echols: Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [18.11]

E. F. Eikenberry: Swiss Light Source, Paul Scherrer Institut, 5232 Villigen PSI, Switzerland. [7.1, 7.2]

H. M. Einspahr: PO Box 6483, Lawrenceville, NJ 08648-0483, USA. [2.2]

D. Eisenberg: UCLA–DOE Laboratory of Structural Biology and Molecular Medicine, Department of Chemistry & Biochemistry, Molecular Biology Institute and Department of Biological Chemistry, UCLA, Los Angeles, CA 90095-1570, USA. [21.3]

P. Emsley: Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK. [17.1]

D. M. Engelman: Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [19.4]

R. A. Engh: Pharmaceutical Research, Roche Diagnostics GmbH, Max Planck Institut für Biochemie, 82152 Martinsried, Germany. [18.3]

J. A. Ernst: Department of Protein Chemistry, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. [3.2]

R. Farid: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

Z. Feng: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

R. H. Fenn: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.‡ [25.1]

J. Frank: Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street, NY 10032, USA. [19.8]

D. T. Gallagher: Center for Advanced Research in Biotechnology of the Maryland Biotechnology Institute and National Institute of Standards and Technology, 9600 Gudelsky Dr., Rockville, MD 20850, USA. [24.3]

E. Gallicchio: BioMaPS Institute for Quantitative Biology, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. [18.10]

E. F. Garman: Department of Biochemistry, South Parks Road, Oxford, OX1 3QU, England. [10.3]

M. Gerstein: Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA. [22.1]

R. Giegé: Machineries Traductionnelles, ARN, UPR 9002, IBMC du CNRS, 15 rue René Descartes, Strasbourg, 67084, France. [4.1]

C. J. Gilmore: Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK. [16.1]

J. P. Glusker: The Institute for Cancer Research, The Fox Chase Cancer Center, Philadelphia, PA 19111, USA. [5.1]

T. Grant: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England. [19.9]

P. Gros: Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. [18.6]

R. W. Grosse-Kunstleve: The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [18.6, 18.11]

S. M. Gruner: Department of Physics, 162 Clark Hall, Cornell University, Ithaca, NY 14853-2501, USA. [7.1, 7.2]

J. R. Gunn: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

W. F. van Gunsteren: Laboratory of Physical Chemistry, ETH-Zentrum, 8092 Zürich, Switzerland. [20.1]

H. A. Hauptman: Hauptman–Woodward Medical Research Institute, Inc., 7000 Ellicott Street, Buffalo, NY 14203-1102, USA. [16.1]

J. J. Headd: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.6]

M. van Heel: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England. [19.9]

J. R. Helliwell: Department of Chemistry, University of Manchester, M13 9PL, England. [8.1]

R. Henderson: Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England. [19.6]

W. A. Hendrickson: Department of Biochemistry, College of Physicians & Surgeons of Columbia University, 630 West 168th Street, New York, NY 10032, USA. [14.2]

K. Henrick: Protein Data Bank in Europe (PDBe), EMBL–EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [24.1]

D. M. Himmel: Biomolecular Crystallography Laboratory, Center for Advanced Biotechnology and Medicine & Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854-5638, USA. [1.1, 1.4]

P. S. Ho: Department of Biochemistry and Molecular Biology, 1870 Campus Delivery, Colorado State University, Fort Collins, CO 80523-1870, USA. [23.6]

A. E. Hodel: Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA.† [23.3]

W. G. J. Hol: Biomolecular Structure Center, Department of Biological Structure, Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195-7742, USA. [1.3]

L. Holm: EMBL–EBI, Cambridge CB10 1SD, England. [23.2]

H. Hope: Department of Chemistry, University of California, Davis, One Shields Ave, Davis, CA 95616-5295, USA. [10.1]

R. Huber: Max-Planck-Institut für Biochemie, 82152 Martinsried, Germany. [12.2, 18.3]

S. H. Hughes: National Cancer Institute, Frederick Cancer R&D Center, Frederick, MD 21702-1201, USA. [3.1]

L.-W. Hung: Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [18.11]

R. M. Immormino: Department of Biochemistry, Duke University, Durham, NC 27710, USA. [21.6]

S. A. Islam: Institute of Cancer Research, 44 Lincoln's Inn Fields, London WC2A 3PX, England. [12.1]

J.-S. Jiang: Protein Data Bank, Biology Department, Brookhaven National Laboratory, Upton, NY 11973-5000, USA. [18.6]

J. E. Johnson: Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, California 92037, USA. [19.3]

L. N. Johnson: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.§ [25.1]

W. Kabsch: Max-Planck-Institut für Medizinische Forschung, Abteilung Biophysik, Jahnstrasse 29, 69120 Heidelberg, Germany. [11.3, 11.6]

G. J. Kapral: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.6]

D. A. Keedy: Department of Biochemistry, Duke University, Durham, NC 27710, USA. [21.6]

G. Kleywegt: Protein Data Bank in Europe (PDBe), EMBL–EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [21.1, 24.1]

R. Knott: Small Angle Scattering Facility, Australian Nuclear Science & Technology Organisation, Physics Division, PMB 1 Menai NSW 2234, Australia. [6.2]

D. F. Koenig: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.¶ [25.1]

A. A. Kossiakoff: Department of Biochemistry and Molecular Biology, CLSC 161A, University of Chicago, Chicago, IL 60637, USA. [19.1]

‡ Present address: 2 Second Avenue, Denvilles, Havant, Hampshire P09 2QP, England.

† Deceased.

§ Present address: Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford OX1 2QU, England.

¶ Present address unknown.

C. M. Koth: Department of Structural Biology, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. [3.2]

V. S. Lamzin: European Molecular Biology Laboratory (EMBL), Hamburg Outstation, c/o DESY, Notkestr. 85, 22603 Hamburg, Germany. [18.8]

R. A. Laskowski: Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, England. [21.4]

A. Leith: Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, Albany, NY 12201-0509, USA. [19.8]

A. G. W. Leslie: MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England. [11.2]

R. Levy: BioMaPS Institute for Quantitative Biology, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. [18.10]

C. Linnemayr: Division of Internal Medicine, Inflammation Research, University Hospital Zürich, Switzerland. [19.9]

B. Lohkamp: Department of Medical Biochemistry and Biophysics, Karolinska Institute, SE-171 77 Stockholm, Sweden. [17.1]

B. Lorber: UPR 9002, IBMC–CNRS, 15 rue René Descartes Cedex, Strasbourg, 67084, France. [4.1]

S. Ludtke: National Center for Macromolecular Imaging, Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [19.10]

V. Y. Lunin: Laboratory of Macromolecular Crystallography, Institute of Mathematical Problems of Biology of the Russian Academy of Sciences, Pushchino, Moscow Region 142290, Russian Federation. [16.3]

M. W. MacArthur: Biochemistry and Molecular Biology Department, University College London, Gower Street, London WC1E 6BT, England. [21.4]

A. J. McCoy: Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge CB2 0XY, England. [18.11]

A. McPherson: Department of Molecular Biology and Biochemistry, University of California, 560 Steinhaus, Irvine, CA 92697-3900, USA. [4.1]

P. Main: Department of Physics, University of York, York YO1 5DD, England. [15.1, 15.3]

G. A. Mair: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.‡ [25.1]

J. Markley: BioMagResBank, University of Wisconsin–Madison, Madison, Wisconsin 53706, USA. [24.1]

B. W. Matthews: Institute of Molecular Biology, Howard Hughes Medical Institute and Department of Physics, University of Oregon, Eugene, OR 97403, USA. [14.1]

C. Mattos: Department of Molecular and Structural Biochemistry, North Carolina State University, 128 Polk Hall, Raleigh, NC 02795, USA. [23.5]

H. Michel: Max-Planck-Institut für Biophysik, Heinrich-Hoffmann-Strasse 7, D-60528 Frankfurt/Main, Germany. [4.2]

R. Miller: Hauptman–Woodward Medical Research Institute, Inc., 7000 Ellicott Street, Buffalo, NY 14203-1102, USA. [16.1]

W. Minor: Department of Molecular Physiology and Biological Physics, University of Virginia, 1300 Jefferson Park Avenue, Charlottesville, VA 22908, USA. [11.4]

K. Moffat: Department of Biochemistry and Molecular Biology, The Center for Advanced Radiation Sources, and The Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois 60637, USA. [8.2]

P. B. Moore: Departments of Chemistry and Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [19.4]

N. W. Moriarty: Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [18.11]

L. W. Murray: Department of Biochemistry, Duke University, Durham, NC 27710, USA. [21.6]

G. N. Murshudov: York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5YW, UK. [18.4]

H. Nakamura: PDBj, Institute for Protein Research, Osaka University, Japan. [24.1]

J. Navaza: Laboratoire de Génétique des Virus, CNRS-GIF, 1. Avenue de la Terrasse, 91198 Gif-sur-Yvette, France. [13.2]

‡ Present address unknown.

A. C. T. North: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.§ [25.1]

R. Oeffner: Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge CB2 0XY, England. [18.11]

J. W. H. Oldham: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.† [25.1]

A. J. Olson: The Scripps Research Institute, La Jolla, CA 92037, USA. [17.2]

C. Orengo: Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England. [23.1]

Z. Otwinowski: UT Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75390-9038, USA. [11.4]

N. S. Pannu: Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. [18.6]

S. Parkin: Department of Chemistry, University of Kentucky, Lexington, Kentucky, USA. [10.1]

A. Perrakis: European Molecular Biology Laboratory (EMBL), Grenoble Outstation, c/o ILL, Avenue des Martyrs, BP 156, 38042 Grenoble Cedex 9, France. [18.8]

D. C. Phillips: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.† [25.1]

A. Podjarny: Structural Biology, IGBMC, BP 163 Cedex, Illkirch, 67404, France. [16.3]

R. J. Poljak: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.¶ [25.1]

J. Pontius: Unité de Conformation de Macromolécules Biologiques, Université Libre de Bruxelles, avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium. [21.2]

R. Portugal: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England. [19.9]

C. B. Post: Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, Indiana 47907-1333, USA. [20.2]

F. A. Quiocho: Howard Hughes Medical Institute and Department of Biochemistry, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [23.3]

R. J. Read: Department of Haematology, University of Cambridge, CIMR, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 0XY, England. [15.2, 18.6, 18.11]

L. M. Rice: Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [18.2, 18.6]

F. M. Richards: Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA. [22.1]

D. C. Richardson: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.5, 21.6]

J. S. Richardson: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.5, 21.6]

J. Richelle: Unité de Conformation de Macromolécules Biologiques, Université Libre de Bruxelles, avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium. [21.2]

D. Ringe: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, 415 South St, Waltham, MA 02254, USA. [23.5]

D. W. Rodgers: Department of Biochemistry, Chandler Medical Center, University of Kentucky, 800 Rose Street, Lexington, KY 40536-0298, USA. [10.2]

A. Rohou: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England.†† [19.9]

M. G. Rossmann: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA. [1.1, 1.2, 1.4, 11.1, 11.5, 13.4]

C. Sander: MIT Center for Genome Research, One Kendall Square, Cambridge, MA 02139, USA. [23.2]

§ Present address: Prospect House, 27 Breary Lane, Bramhope, Leeds LS16 9AD, England.

† Deceased.

¶ Present address: CARB, 9600 Gudelsky Drive, Rockville, MD 20850, USA.

†† Present address: Rosenstiel Center, Brandeis University, 415 South Street, Waltham, MA 02454-9110, USA.

V. R. Sarma: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.‡ [25.1]

C. Sauter: Institut de Biologie Moléculaire et Cellulaire (IBMC), Centre National pour la Recherche Scientifique (CNRS), 15 rue René Descartes, Strasbourg, F-67084, France. [4.1]

M. R. Sawaya: Institute for Genomics and Proteomics, UCLA–DOE, 611 Young Drive East, Los Angeles, CA 90095, USA. [18.12]

M. Schatz: Image Science Software GmbH, Gillweg 3, D-14193 Berlin, Germany. [19.9]

R. Schmidt: Image Science Software GmbH, Gillweg 3, D-14193 Berlin, Germany. [19.9]

B. Schneider: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA.§ [24.2]

B. P. Schoenborn: Life Sciences Division M888, University of California, Los Alamos National Laboratory, Los Alamos, NM 8745, USA. [6.2]

D. Shapiro: Advanced Light Source, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, MS 2-400, Berkeley, CA 94720, USA. [9.3]

K. A. Sharp: E. R. Johnson Research Foundation, Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA 19104-6059, USA. [22.4]

G. M. Sheldrick: Lehrstuhl für Strukturchemie, Georg-August-Universität Göttingen, Tammannstrasse 4, D-37077 Göttingen, Germany. [16.1, 18.9]

T. Simonson: Laboratoire de Biologie Structurale (CNRS), IGBMC, 1 rue Laurent Fries, 67404 Illkirch (CU de Strasbourg), France. [18.6]

J. L. Smith: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA. [14.2]

J. C. H. Spence: Department of Physics, Arizona State University, Tempe, Arizona, 85287, USA. [1.4]

M. J. E. Sternberg: Institute of Cancer Research, 44 Lincoln's Inn Fields, London WC2A 3PX, England. [12.1]

A. M. Stock: Center for Advanced Biotechnology and Medicine, Howard Hughes Medical Institute and University of Medicine and Dentistry of New Jersey – Robert Wood Johnson Medical School, 679 Hoes Lane, Piscataway, NJ 08854-5627, USA. [3.1]

U. Stocker: Laboratory of Physical Chemistry, ETH-Zentrum, 8092 Zürich, Switzerland. [20.1]

G. Stubbs: Department of Molecular Biology, Vanderbilt University, Nashville, TN 37235, USA. [19.5]

M. T. Stubbs: Institut für Pharmazeutische Chemie der Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany. [12.2]

S. Sun: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA. [1.4]

M. W. Tate: Department of Physics, 162 Clark Hall, Cornell University, Ithaca, NY 14853-2501, USA. [7.1, 7.2]

L. F. Ten Eyck: San Diego Supercomputer Center 0505, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA. [18.1, 18.7]

A. Teplyakov: University of Maryland Biotechnology Institute, Rockville, MD 20850, USA.¶ [13.5]

T. C. Terwilliger: Bioscience Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [14.3, 18.11]

J. Thornton: Biochemistry and Molecular Biology Department, University College London, Gower Street, London WC1E 6BT, England, and Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, England. [21.4, 23.1]

L. Tong: Department of Biological Sciences, Columbia University, New York, NY 10027, USA. [13.3]

D. E. Tronrud: Howard Hughes Medical Institute, Institute of Molecular Biology, 1229 University of Oregon, Eugene, OR 97403-1229, USA. [18.7]

Y. Tsai: Department of Chemistry and Biochemistry, University of California, Los Angeles, PO Box 951569, Los Angeles, CA 90095-1569, USA. [11.7]

H. Tsuruta: SSRL/SLAC & Department of Chemistry, Stanford University, PO Box 4349, MS69, Stanford, California 94309-0210, USA. [19.3]

M. Tung: Center for Advanced Research in Biotechnology of the Maryland Biotechnology Institute and National Institute of Standards and Technology, 9600 Gudelsky Dr., Rockville, MD 20850, USA. [24.3]

A. G. Urzhumtsev: Faculty of Sciences, University of Nancy 1, Vandoeuvre-lès-Nancy, 54506, France. [16.3]

I. Usón: Institució Catalana de Recerca i Estudis Avançats at IBMB–CSIC, Barcelona Science Park, Baldiri Reixach 15, 08028 Barcelona, Spain. [16.1]

A. Vagin: Structural Biology Laboratory, University of York, York YO10 5YW, England. [13.5, 21.2]

M. J. Vallejos: Department of Biochemistry and Molecular Biology, 1870 Campus Delivery, Colorado State University, Fort Collins, CO 80523-1870, USA. [23.6]

M. L. Verdonk: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [22.5]

C. L. M. J. Verlinde: Biomolecular Structure Center, Department of Biological Structure, Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195-7742, USA. [1.3]

C. A. Vernon: Department of Chemistry, University College London, Gower Street, London WC1E 6BT, England.† [25.1]

K. D. Watenpaugh: Structural, Analytical and Medicinal Chemistry, Pharmacia & Upjohn, Inc., Kalamazoo, MI 49001-0119, USA. [18.1]

C. M. Weeks: Hauptman–Woodward Medical Research Institute, Inc., 7000 Ellicott Street, Buffalo, NY 14203-1102, USA. [16.1]

M. S. Weiss: Helmholtz-Zentrum Berlin für Materialien und Energie, Macromolecular Crystallography (HZB-MX), Albert-Einstein-Str. 15, D-12489 Berlin, Germany. [2.2]

E. M. Westbrook: Molecular Biology Consortium, Argonne, Illinois 60439, USA. [5.2]

J. Westbrook: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

K. S. Wilson: York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5YW, England. [9.1, 18.4, 18.8]

S. J. Wodak: Unité de Conformation de Macromolécules Biologiques, Université Libre de Bruxelles, avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium, and EMBL–EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England. [21.2]

K. Wüthrich: Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule-Hönggerberg, CH-8093 Zürich, Switzerland. [19.7]

H. Yang: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

D. G. Yansura: Department of Antibody Engineering, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. [3.2]

T. O. Yeates: Department of Chemistry and Biochemistry, University of California, Los Angeles, PO Box 951569, Los Angeles, CA 90095-1569, USA. [11.7, 18.12, 21.3]

J. Young: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

C. Zardecki: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

K. Y. J. Zhang: Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N., Seattle, WA 90109, USA. [15.1, 15.3]

Z. Zhou: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

P. H. Zwart: Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [18.11]

‡ Present address: Department of Biochemistry, State University of New York at Stonybrook, Stonybrook, NY 11794-5215, USA.

§ Present address: Institute of Biotechnology, Academy of Science, Czech Republic.

¶ Present address: Centocor R&D Inc., Radnor, PA 19087, USA.

† Deceased.

Contents

CONTENTS

Preface

E. Arnold and M. G. Rossmann

International Tables for Crystallography, Volume F, *Crystallography of Biological Macromolecules*, was commissioned by the International Union of Crystallography (IUCr) in recognition of the extraordinary contributions that knowledge of macromolecular structure has made, and will make, to the analysis of biological systems, from enzyme catalysis to the workings of a whole cell. The volume covers all stages of a crystallographic analysis from the preparation of samples using the techniques of molecular biology and biochemistry, to crystallization, diffraction-data collection, phase determination, structure validation and structure analysis. Although the book is written for experienced scientists, it is recognized that the modern structural biologist is more likely to be a biologist interested in structure than a classical crystallographer interested in biology. Thus, there are chapters on the fundamentals, history and current perspectives of macromolecular crystallography, as well as on the availability of useful programs and databases including the Protein Data Bank. Each chapter has been written by an internationally recognized expert.

Macromolecular crystallography is undergoing a revolution. Just as crystallography became central to the study of chemistry, macromolecular crystallography has become a core science in biology. Macromolecular crystallography has shaped our view of biological molecular structure, and is providing a broader understanding of biological ultrastructure and the molecular interactions in living systems. As reflected by the exponential increase in entries in the Protein Data Bank over the past decade, there has been an explosion in the number of macromolecular structures determined, the majority by X-ray crystallography. Knowledge of the sequences of entire genomes, from bacteria to human, has sparked a structural genomics effort that aims to determine 10 000 new macromolecular structures in the next decade. Crystallography is expected to yield the largest share of this new crop of structures. The field of macromolecular crystallography is still evolving rapidly, and capturing its essence in a single volume is a challenge. Therefore, the volume emphasizes durable knowledge, but also contains articles on somewhat more volatile topics.

This project had its inception when Ted Baker (at that time President of the IUCr) approached one of us (MGR) about writing a book on macromolecular crystallography for the IUCr. Not only were there already some excellent books that covered most aspects of the subject, but the breadth of the subject was now so vast that no single person could possibly be an expert in all relevant topics. After further exchanges of e-mails, MGR realized that the officers of the IUCr were tacitly assuming that he would be willing to carry out the advice he had given so freely. He then asked his former post-doc and coauthor of an earlier article on molecular replacement in Volume B of *International Tables*, Eddy Arnold, to help him get out of a tight corner. After some serious deliberations of his own, Eddy agreed to be co-editor.

Together we fleshed out an outline that was broader than MGR's original plan, which had focused largely on crystallographic theory and technique. We felt that it would be valuable to briefly cover related techniques beyond X-ray diffraction, as well as to give an overview of the current field of structural biology. Although basic crystallography is also presented in the other volumes of *International Tables*, chapters describing fundamental crystallographic principles and practices have been included in an attempt to make the volume as coherent and self-contained as possible. We established an advisory board, developed a list of required chapters and obtained promises of participation from potential authors. In a departure from the style of previous volumes of *International Tables*, which have fewer articles and authors, we sought contributions for nearly 100 articles from an even larger number of contributing authors. The members of the advisory board reviewed the proposed outline of chapters and authors. We were pleasantly surprised when so many experts generously agreed to write articles for this volume, and delighted that the vast majority fulfilled their promises.

Significant events punctuating the process were the 1996 and 1999 IUCr congresses. At the 1996 IUCr Congress in Seattle, we convened a meeting with many of the authors. There we described the overall project design and received valuable suggestions. At that time, we hoped that the volume could be completed by 1999. At the 1999 IUCr Congress in Glasgow, we reviewed the detailed contents of the volume at an open meeting on the volumes of *International Tables* under development. By that time, we had received most of the articles and typesetting began in late 1999. The complexities of handling a large number of articles from so many authors led to delays at a number of stages. Ultimately, the completion date became mid-2001.

We are especially grateful to the staff at the IUCr and at our own institutions for their dedicated help in bringing this project to fruition. At the IUCr, we thank Nicola Ashcroft for an outstanding job on overall production of the volume, and for her patient correspondence and attention to detail. We also thank Peter Strickland, Sue King, Theo Hahn, Uri Shmueli, Mike Dacombe and Ted Baker for their help in coordinating the project. At Purdue University, we thank Cheryl Towell and Sharon Wilder for constant assistance, and Fay Chen for editorial suggestions. At the Center for Advanced Biotechnology and Medicine and Rutgers University, we thank Susan Mazzocchi and Barbara Shaver for their help in handling correspondence and galley proofs from the authors.

We are also especially indebted to the authors for their generous contributions and for documenting relevant expertise. We also thank the advisors and the members of the advisory board for their help. We are saddened to note that Paul Sigler, a member of the advisory board, passed away during the project. Paul was a towering figure who, with his medical background, recognized the role structure plays in providing insights into fundamental chemical and biological processes.

Preface to the second edition

E. Arnold, D. M. Himmel and M. G. Rossmann

Ten years after the appearance of the first edition of *International Tables for Crystallography*, Volume F, *Crystallography of Biological Macromolecules*, we are pleased to present the second edition. Preparation of the second edition has reflected the continuing evolution of macromolecular crystallography. Eddy Arnold and Michael Rossmann were glad to have Daniel Himmel, an accomplished crystallographer with diverse knowledge, join this effort. The three of us brainstormed about how to update this volume most effectively, and decided to seek new articles in key areas of rapid growth and ask authors from the first edition if they wished to revise their chapters. We were delighted when so many internationally recognized experts again made generous contributions that brought the volume up to date.

In the past decade the field of macromolecular crystallography has spawned an ever-increasing array of spectacular biological structures. The amazing conquests have included the translational machinery, the ribosome, in various functional states; the transcription machinery, multisubunit cellular RNA polymerases; a multitude of membrane proteins; and many more viruses. The marriage of crystallography and cryo-electron microscopy, including electron tomography, enables a multiscale view of biological structure including macromolecular inventories of cells in three dimensions. Multiprotein complexes are more routinely studied, and small-angle X-ray scattering has grown in prominence as a complementary technique to single-crystal X-ray diffraction. Structural genomics projects have been knocking out representative structures from a wide variety of proteomes from all levels of life. Structure-based drug-design prospects and the general problem of understanding protein–ligand interactions have benefitted from the development of techniques such as fragment screening, where thousands of small-molecule fragments are systematically soaked into crystals to interrogate potential binding sites – a central tool in the emerging field of chemical genetics. Software packages continue to evolve to enable rapid and reliable structure solution, visualization and analysis. X-ray lasers are now a reality and perhaps in the future many complex structures will be solved and imaged using data sets measured from tiny microcrystals that diffract their heart out before being destroyed by the pulsed high-brilliance beams. When thinking about the next ten years, we can expect more of the unexpected.

We wish to thank Nicola Ashcroft for handling the majority of correspondence with authors, and for maximizing the efficiency and quality of the overall process, thereby keeping the production schedule under control. We also thank our many colleagues and friends whose advice was valuable in designing the plan for the second edition.

SAMPLE PAGES

PART 1. INTRODUCTION

Chapter 1.1. Overview

E. Arnold, D. M. Himmel and M. G. Rossmann

Volume F is devoted to the crystallography of large biological molecules, complementing the existing volumes of *International Tables for Crystallography*. A background history of the subject is followed by a concise introduction to the basic theory of X-ray diffraction and other requirements for the practice of crystallography. Basic crystallographic theory is presented in greater depth in other volumes of *International Tables*. The information in the latter portions of Volume F is specifically related to macromolecular structure.

Chapter 1.2 presents a brief history of the field of macromolecular crystallography. This is followed by an article (Chapter 1.3) that describes many of the connections of crystallography with medicine and gives a look into the future possibilities of structure-based design of drugs, vaccines and other agents. Chapter 1.4 provides some personal perspectives on what the future may hold for crystallography and the other physical sciences, as well as how budding new technologies and breakthroughs in these fields may impact both the sciences and the world at large.

Chapter 2.1 introduces diffraction theory and fundamentals of crystallography, including concepts of real and reciprocal space, unit-cell geometry and symmetry. It is shown how scattering from electron density and atoms leads to the formulation of structure factors. The phase problem is introduced, as well as the basic theory behind some of the more common methods for its solution. As the field of macromolecular crystallography has grown and matured, a variety of metrics have evolved for monitoring the quality of data collection, data processing and structure determination. In some cases, disagreement has arisen over the best way to calculate some of these metrics or which ones are most useful for monitoring the progress of the crystallographic experiment. Chapter 2.2 provides a comprehensive compilation of the most commonly used indicators of quality, along with a definition for each. The application of some of these indicators to validate the refined crystal structure is addressed further in Part 21 (below).

Molecular biology has had a major impact in terms of accelerating progress in structural biology and remains a rapidly developing area. Chapter 3.1 is a primer on modern molecular-biology techniques for producing materials for crystallographic studies. Since large amounts of highly purified materials are required, emphasis is placed on approaches for efficiently and economically yielding samples of biological macromolecules suitable for crystallization. Chapter 3.2 extends this overview to methods for the expression and purification of membrane proteins for structural studies. These introductions to protein expression are complemented by Chapter 4.3, which describes the rapidly developing area of molecular-engineering approaches for obtaining high-quality crystals of biological macromolecules.

The basic theory and practice of macromolecular crystallization are described in Chapters 4.1 and 4.2. This, too, is a rapidly evolving area, with continual advances in theory and practice. It is remarkable to consider the macromolecules that have been crystallized. We expect macromolecular engineering, covered in Chapter 4.3, to play a central role in coaxing more macromolecules to form crystals suitable for structure determination in the future. These areas will be complemented by high-throughput crystallographic approaches, covered in Chapter 4.4. These include the expression of proteins in various organisms for their structure determination. The material in Part 4 is complemented by Part 5, which summarizes traditional properties of and methods for handling macromolecular crystals, as well as how to measure crystal density.

Part 6 gives a brief introduction to the theory and practice of generating X-rays and neutrons for diffraction experiments. Chapter 6.1 describes the theory of X-ray production from both conventional and synchrotron X-ray sources, as well as methods for defining the energy spectrum and geometry of X-ray beams. Numerous articles in the other volumes of *International Tables* go into greater depths. Chapter 6.2 describes the generation of neutron beams.

Part 7 describes common methods for detecting X-rays, with a focus on detection devices that are currently most frequently used, including storage phosphor image plate and charge-coupled device (CCD) detectors. This has been another rapidly developing area, particularly in the past two decades. A further article describing X-ray detector theory and practice is *International Tables* Volume C, Chapter 7.1.

Synchrotron-radiation sources have played a prominent role in advancing the frontiers of macromolecular structure determination in terms of size, quality and throughput. The extremely high intensity, tunable wavelength characteristics and pulsed time structure of synchrotron beams have enabled many novel experiments. Some of the unique characteristics of synchrotron radiation are being harnessed to help solve the phase problem using anomalous-scattering measurements, *e.g.* in multiwavelength anomalous diffraction (MAD) experiments (see Chapter 14.2). The quality of synchrotron-radiation facilities for macromolecular studies has been increasing rapidly, partly in response to the perceived value of the structures being determined. Many synchrotron beamlines have been designed to meet the needs of macromolecular experiments. Chapter 8.1 surveys many of the roles that synchrotron radiation plays in modern macromolecular structure determination. Chapter 8.2 summarizes applications of the age-old Laue crystallography technique, which has seen a revival in the study of macromolecular crystal structures using portions of the white spectrum of synchrotron X-radiation. Chapter 4.2 of *International Tables* Volume C is also a useful reference for understanding synchrotron radiation.

Chapter 9.1 summarizes many aspects of data collection from single crystals using monochromatic X-ray beams. Common camera-geometry and coordinate-system-definition schemes are given. Because most macromolecular data collection is carried out using the oscillation (or rotation) method, strategies related to this technique are emphasized. A variety of articles in Volume C of *International Tables* serve as additional references. With the

advent of modern synchrotron-radiation facilities that provide high-intensity beams and CCD detectors with rapid data readout, data collection has often become so rapid that the rate-limiting step in collecting a data set could very well be the time it takes the experimenter to mount the crystal manually, lock the hutch and begin to take X-ray exposures. The introduction of robotic crystal loading, covered in Chapter 9.2, makes it possible to mount and screen many more crystals, collecting and saving data from them, in the same amount of time, often with far less risk to the crystals. Chapter 9.3 summarizes another technology, coherent X-ray diffraction microscopy (CXDM), which is made possible by the brightness of third-generation synchrotron X-ray sources. Using this method, diffractive methods can be applied to image certain non-crystalline single particles to as high as 11 nm resolution. Single-particle phase retrieval theory is briefly presented, followed by the application of CXDM to the imaging of a yeast cell, a discussion of the ramifications of radiation damage and, finally, the use of stereoscopic viewing to access three-dimensional information.

The use of cryogenic cooling of macromolecular crystals for data collection ('cryocrystallography') has become the most frequently used method of crystal handling for data collection. Part 10 summarizes the theory and practice of cryocrystallography. Among its advantages are enhanced crystal lifetime and improved resolution. Most current experiments in cryocrystallography use liquid-nitrogen-cooled gas streams, though some attempts have been made to use liquid-helium-cooled gas streams. Although cryogenic temperatures substantially protect macromolecular crystals against radiation damage from the generation of free radicals, the problem of radiation damage has not been entirely eliminated. Methods continue to be explored to reduce the rate of radiation damage, such as the use of free-radical scavengers in cryoprotective solutions, or to correct reflections data sets for X-ray radiation dose. Methods have also been developed for utilizing radiation damage for phasing. Just a decade ago, it was still widely believed that many macromolecular crystals could not be studied successfully using cryocrystallography, or that the practice would be troublesome or would lead to inferior results. Now, crystallographers routinely screen for suitable cryoprotective conditions for data collection even in initial experiments, and often crystal diffraction quality is no longer assessed except using cryogenic cooling. However, some crystals have resisted attempts to cool them successfully to cryogenic temperatures. Thus, data collection using ambient conditions, or moderate cooling (from approximately -40 °C to a few degrees below ambient temperature), is not likely to become obsolete in the near future.

Part 11 describes the processing of X-ray diffraction data from macromolecular crystals. Special associated problems concern dealing with large numbers of observations, large unit cells (hence crowded reciprocal lattices) and diverse factors related to crystal imperfection (large and often anisotropic mosaicity, variability of unit-cell dimensions *etc.*). Various camera geometries have been used in macromolecular crystallography, including precession, Weissenberg, three- and four-circle diffractometry, and oscillation or rotation. The majority of diffraction data sets are collected now *via* the oscillation method and using a variety of detectors. Among the topics covered in Part 11 are autoindexing, intensity integration, space-group assignment, scaling, post refinement, and detection of merohedral twinning. Commonly used software packages for the indexing, integration, scaling and post refinement of crystallographic data are presented. The discussion of intensity integration in Chapter 11.2 includes a description of some of the strategies used by *MOSFLM* and *DENZO*. *DENZO* and *SCALEPACK*, which form the backbone of the *HKL2000* package, are described in detail in Chapter 11.4, and *XDS* is described in Chapter 11.6.

Part 12 describes the theory and practice of the isomorphous replacement method, and begins the portion of Volume F that addresses how the phase problem in macromolecular crystallography can be solved. The isomorphous replacement method was the first technique used for solving macromolecular crystal structures, and will continue to play a central role for the foreseeable future. Chapter 12.1 describes the basic practice of isomorphous replacement, including the selection of heavy-metal reagents as candidate derivatives and crystal-derivatization procedures. Chapter 12.2 surveys some of the techniques used in isomorphous replacement calculations, including the location of heavy-atom sites and use of that information in phasing. Readers are also referred to Chapter 2.4 of *International Tables* Volume B for additional information about the isomorphous replacement method.

Part 13 describes the molecular replacement method and many of its uses in solving macromolecular crystal structures. This part covers general definitions of noncrystallographic symmetry, the use of rotation and translation functions, and phase improvement and extension *via* noncrystallographic symmetry. The molecular replacement method is commonly used to solve macromolecular crystal structures where redundant information is present either in a given crystal lattice or among different crystals. In some cases, phase information is obtained by averaging noncrystallographically redundant electron density either within a single crystal lattice or among multiple crystal lattices. In other cases, atomic models from known structures can be used to help phase unknown crystal structures containing related structures. With the number and type of known protein structures rising exponentially in recent years (see Chapter 24.1), molecular replacement has become the most common way to solve phases for an unknown protein structure. Molecular-replacement phasing is often used in conjunction with other phasing methods, including isomorphous replacement and density-modification methods. *International Tables* Volume B, Chapter 2.3 is also a useful reference for molecular-replacement techniques.

Anomalous-dispersion measurements have played an increasingly important role in solving the phase problem for macromolecular crystals. Anomalous dispersion has long been recognized as a source of experimental phase information; for more than three decades, macromolecular crystallographers have been exploiting anomalous-dispersion measurements from crystals containing heavy metals, using even conventional X-ray sources. In the past two decades, synchrotron sources have permitted optimized anomalous-scattering experiments, where the X-ray energy is selected to be near an absorption edge of a scattering element. Chapter 14.1 summarizes applications of anomalous scattering using single wavelengths for macromolecular crystal-structure determination. The multiwavelength anomalous diffraction (MAD) technique, in particular, is used to solve the phase problem for a broad array of macromolecular crystal structures. In the MAD experiment, intensities measured from a crystal at a number of wavelengths permit direct solution of the phase problem, frequently yielding easily interpretable electron-density maps. The theory and practice of the MAD technique are described in Chapters 14.2 and 14.3.

Density modification, discussed in Part 15, encompasses an array of techniques used to aid solution of the phase problem

via electron-density-map modifications. Recognition of usual density-distribution patterns in macromolecular crystal structures permits the application of such techniques as solvent flattening (disordered solvent regions have lower density), histogram matching (normal distributions of density are expected) and skeletonization (owing to the long-chain nature of macromolecules such as proteins). Electron-density averaging, discussed in Chapter 13.4, can be thought of as a density-modification technique as well. Chapter 15.1 surveys the general problem and practice of density modification, including a discussion of solvent flattening, histogram matching, skeletonization and phase-combination methodology. Chapter 15.2 discusses weighting of Fourier terms for calculation of electron-density maps in a more general sense, especially with respect to the problem of minimizing model bias in phase improvement. Electron-density modification techniques can often be implemented efficiently in reciprocal space, too. Chapter 15.3 summarizes the implementation of phase modification in the *DM/DMMULTI* software.

Part 16 describes the use of direct methods in macromolecular crystallography. Some 30 years ago, direct methods revolutionized the practice of small-molecule crystallography by facilitating structure solution directly from intensity measurements. As a result, phase determination of most small-molecule crystal structures has become quite routine. In the meantime, many attempts have been made to apply direct methods to solving macromolecular crystal structures. Prospects in this area are improving, but success has been obtained in only a limited number of cases, often with extremely high resolution data measured from small proteins. Chapter 16.1 surveys progress in the application of direct methods to solve macromolecular crystal structures, and Chapter 16.2 summarizes the application of the maximum-entropy method to crystallography. Chapter 16.3 explores techniques now being developed to extend *ab initio* phasing to low-resolution data.

The use of computer graphics for building models of macromolecular structures has facilitated the efficiency of macromolecular structure solution and refinement immensely (Part 17). Until just a little more than 20 years ago, all models of macromolecular structures were built as physical models, with parts of appropriate dimensions scaled up to our size! Computer-graphics representations of structures have made macromolecular structure models more precise, especially when coupled with refinement methods, and have contributed to the rapid proliferation of new structural information. With continual improvement in computer hardware and software for three-dimensional visualization of molecules (the crystallographer's version of 'virtual reality'), continuing rapid progress and evolution in this area are likely. The availability of computer graphics has also contributed greatly to the magnificent illustration of crystal structures, one of the factors that have thrust structural biology into many prominent roles in modern life and chemical sciences. Chapter 17.1 describes a graphics software product, *Coot*, now widely used by macromolecular crystallographers for visualization and real-space refinement of X-ray crystal structures. Chapter 17.2 surveys the field of computer visualization and animation of molecular structures, with a valuable historical perspective. Chapter 3.3 of *International Tables* Volume B is a useful reference for basics of computer-graphics visualization of molecules.

As in other areas of crystallography, refinement methods are used to obtain the most complete and precise structural information from macromolecular crystallographic data. The often-limited resolution and other factors lead to underdetermination of structural parameters relative to small-molecule crystal structures. In addition to X-ray intensity observations, macromolecular refinement incorporates observations about the normal stereochemistry of molecules, thereby improving the data-to-parameter ratio. Whereas incorporation of geometrical restraints and constraints in macromolecular refinement was initially implemented about 30 years ago, it is now generally a publication prerequisite that this methodology be used in structure refinement. Basic principles of crystallographic refinement, including least-squares minimization, constrained refinement and restrained refinement, are described in Chapter 18.1. Simulated-annealing methods, discussed in Chapter 18.2, can accelerate convergence to a refined structure, and are now widely used in refining macromolecular crystal structures. Structure quality and target parameters for stereochemical constraints and restraints are discussed in Chapter 18.3. High-resolution refinement of macromolecular structures, including handling of hydrogen-atom positions, is discussed in Chapter 18.4. Estimation of coordinate error in structure refinement is discussed in Chapter 18.5. Chapters 18.6 to 18.11 summarize computer programs and packages in common use for macromolecular structure determination, refinement and analysis. These program systems include the *Crystallography & NMR System* (*CNS*, Chapter 18.6), the *TNT* refinement package (Chapter 18.7), *ARP* and *wARP* for automated model construction and refinement (Chapter 18.8), *SHELX* (Chapter 18.9), *PrimeX* and the *Schrödinger* computational chemistry suite (Chapter 18.10), and *PHENIX* (Chapter 18.11). This part concludes with a description of methods for structure determination in the presence of merohedral twinning (Chapter 18.12).

Part 19 is a collection of short reviews of alternative methods for studying macromolecular structure. Each can provide information complementary to that obtained from single-crystal X-ray diffraction methods. In fact, structural information obtained from nuclear magnetic resonance (NMR) spectroscopy or cryo-electron microscopy is now frequently used in initiating crystal structure solution *via* the molecular replacement method (Part 13). Neutron diffraction, discussed in Chapter 19.1, can be used to obtain high-precision information about hydrogen atoms in macromolecular structures. Electron-diffraction studies of thin crystals are yielding structural information to increasingly high resolution, often for problems where obtaining three-dimensional crystals is challenging (Chapter 19.2). Small-angle X-ray (Chapter 19.3) and neutron (Chapter 19.4) scattering studies can be used to obtain information about shape and electron-density contrast even in noncrystalline materials, and are especially informative in cases of large macromolecular assemblies (*e.g.* viruses and ribosomes). Fibre diffraction (Chapter 19.5) can be used to study the structure of fibrous biological molecules. The combination of electron microscopy and crystallography is helping to bridge molecular structure and multi-molecular ultrastructure in living cells. Cryo-electron microscopy and high-resolution electron microscopy have been applied to the study of detailed structures of noncrystalline molecules of increasing complexity (Chapter 19.6). NMR spectroscopy has become a central method in the determination of small and medium-sized protein structures (Chapter 19.7), and yields unique descriptions of molecular interactions and motion in solution. Continuing breakthroughs in NMR technology are expanding greatly the size range of structures that can be studied by NMR. Part 19 concludes with descriptions of software packages in common use for data processing and image reconstruction of single particles using electron-microscopy techni-

ques. These include *SPIDER* and *SPIRE* (Chapter 19.8), *IMAGIC 4D* (Chapter 19.9) and *EMAN* (Chapter 19.10).

Energy and molecular-dynamics calculations already play an integral role in many approaches for refining macromolecular structures (Part 20). Simulation methods hold promise for greater understanding of the time course of macromolecular motion than can be obtained through painstaking experimental approaches. However, experimental structures are still the starting point for simulation methods, and the quality of simulations is judged relative to experimental observables. Chapters 20.1 and 20.2 present complementary surveys of the current field of energy and molecular-dynamics calculations.

Structure validation (Part 21) is an important part of macromolecular crystal structure determination. Owing in part to the low data-to-parameter ratio and to problems of model phase bias, it can be difficult to correct misinterpretations of structure that can occur at many stages of structure determination. Chapters 21.1, 21.2 and 21.3 present approaches to structure validation using a range of reference information about macromolecular structure, in addition to observed diffraction intensities. Structure-validation methods are especially important in cases where unusual or highly unexpected features are found in a new structure. Software packages widely used for structure validation, as well as the detection and correction of errors in structures, are described in Chapters 21.4 (*PROCHECK*), 21.5 (*KiNG* and kinemages) and 21.6 (*MolProbity*).

Part 22 presents a survey of many methods used in the analysis of macromolecular structure. Since macromolecular structures tend to be very complicated, it is essential to extract features, descriptions and representations that can simplify information in helpful ways. Calculations of molecular surface areas, volumes and solvent-accessible surface areas are discussed in Chapters 22.1 and 22.2. Useful generalizations relating surface areas buried at macromolecular interfaces and energies of association have emerged. Chapter 22.3 surveys the occurrence of hydrogen bonds in biological macromolecules. This treatment is complemented by a description of solvent structure in Chapter 23.5, below. Electrostatic interactions in proteins are described in Chapter 22.4. The Cambridge Structural Database is the most complete compendium of small-molecule structural data; its role in assessing macromolecular crystal structures is discussed in Chapter 22.5. Databases for the deposition of macromolecular coordinates are the subject of Part 24, below.

Part 23 surveys current knowledge of protein and nucleic acid architecture. Proliferation of structural data has created problems for classification schemes, which have been forced to co-evolve with new structural knowledge. Methods and computational algorithms for the structural classification of various protein tertiary folds and the recognition of separate structural domains are reviewed in Chapters 23.1 and 23.2. Systematic aspects of ligand binding to macromolecules are discussed in Chapter 23.3. A survey of nucleic acid structure, geometry and classification schemes is presented in Chapter 23.4. Solvent structure in macromolecular crystals is reviewed in Chapter 23.5. Halogen interactions with small organic molecules have been known for some time now, but halogen interactions with macromolecules have only begun to gain recognition in recent years. These interactions tend to be unfamiliar and, at times, non-intuitive to the structural biologist. For example, an iodine or bromine atom (*i.e.*, of a ligand) might interact with the carbonyl oxygen atom of a protein without electronic repulsion, provided the geometry of the interaction allows the halogen atom to serve as a Lewis acid instead of the expected Lewis base. Our current understanding of halogen–macromolecule interactions and the geometric parameters within which they operate are reviewed in Chapter 23.6.

With the proliferation of macromolecular structures, it has been necessary to establish databases as international resources for rapid access to, and archival of, primary structural data. The Protein Data Bank (PDB), which for almost thirty years was the depository for protein crystal (and later NMR) structures, was first established at Brookhaven National Laboratory in the United States and later expanded into the worldwide PDB (wwPDB). The wwPDB is now run by the Research Collaboratory for Structural Bioinformatics (RCSB PDB) based in the United States with additional deposition centres: the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (EBI), now the Protein Data Bank in Europe (PDBe); the PDB Japan (PDBj); and the BioMagResBank (BMRB). Chapter 24.1 describes the organization and features of the wwPDB. The wwPDB permits ready access to the rapidly increasing store of macromolecular structural data *via* the Internet, as well as rapid correlation of structural data with other key life-sciences databases. The Nucleic Acid Database (NDB), containing nucleic acid structures with and without bound ligands and proteins, is presented in Chapter 24.2. The Biological Macromolecule Crystallization Database (BMCD), a repository for macromolecular crystallization data, is described in Chapter 24.3.

Chapter 25.1 provides a detailed history of the structure determination of lysozyme, the first enzyme crystal structure to be solved. This chapter serves as a guide to the process by which the lysozyme structure was solved. Although the specific methods used to determine macromolecular structures have changed, the overall process is similar and the reader should find this account entertaining as well as instructive.

between the E values (analogous to normalized structure factors as above) derived from the observed isomorphous or anomalous differences and those calculated from the substructure model. In contrast to CC(all), however, CC(weak) is calculated for the weak reflections only. As above, x and y are the E values derived from the observed and calculated differences. The E-value cutoff for defining a reflection as weak can be chosen by the user, but a typical threshold value is 1.5, although lower values may be required for low-resolution data (Schneider & Sheldrick, 2002). A value of CC(weak) ≥ 0.15 often indicates that the substructure has been correctly identified (Sheldrick, 2010).

The minimal function, $R(\varphi)$. The minimal function $R(\varphi)$ is a measure of the mean-square difference between the values of the triplets calculated using a particular set of phases and the expected values of the same triplets as given by the ratio of modified Bessel functions. The minimal function is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph (the minimal principle).

$$R(\varphi) = \sum_{H,K} A_{HK}\left\{\cos\varphi_{HK} - [I_1(A_{HK})/I_0(A_{HK})]\right\}^2 / \sum_{H,K} A_{HK}, \tag{2.2.4.1}$$

where $A_{HK} = (2/N^{1/2})|E_H E_K E_{H+K}|$ with N being the number of atoms in the corresponding primitive unit cell. The minimal function is the function minimized in the program *Shake&Bake*, abbreviated as *SnB* (Weeks *et al.*, 1993, 1994).

2.2.5. Quality indicators for phase determination

Once the isomorphous or anomalous substructure has been successfully determined, it can be used as reference point for the calculation of phases. The quality of the resulting phases is dependent on the strength of the isomorphous or anomalous signal and the completeness and correctness of the isomorphous or anomalous substructure.

Cullis R factor, $R_{\rm Cullis}$. The Cullis R factor (Cullis *et al.*, 1961) for phase determination by isomorphous replacement is defined as the ratio between the lack-of-closure error $\varepsilon(\varphi_{\rm P})$ [equation (2.2.5.1*a*) below] and the isomorphous difference $|F_{\rm PH} - F_{\rm P}|$. It is the most useful signal for a usable heavy-atom derivative. Values < 0.6 for centrosymmetric data are excellent, while values < 0.9 are still usable.

$$\varepsilon(\varphi_{\rm P}) = \left|F_{\rm PH} - |F_{\rm P} + F_{\rm H}|\right|, \tag{2.2.5.1a}$$

$$R_{\rm Cullis} = \sum_{hkl}\left|F_{\rm PH} - |F_{\rm P} + F_{\rm H}|\right| / \sum_{hkl} |F_{\rm PH} - F_{\rm P}|. \tag{2.2.5.1b}$$

Anomalous Cullis R factor, $R_{\rm Cullis,ano}$. The Cullis R factor for phase determination by anomalous dispersion is defined as the ratio between the lack-of-closure error and the observed anomalous difference $|F_{\rm PH}(hkl) - F_{\rm PH}(\bar{h}\bar{k}\bar{l})|$. The lack-of-closure error in the anomalous-dispersion case is the difference between the observed anomalous difference and the calculated anomalous difference $2F_{\rm H}\sin\alpha_{\rm P}$, where $\alpha_{\rm P}$ is the protein phase. A value of $R_{\rm Cullis,ano} < 1.0$ suggests that a contribution to the phasing from the anomalous data is likely (*MLPHARE* program documentation; Collaborative Computational Project, No. 4, 1994).

$$R_{\rm Cullis,ano} = \frac{\sum_{hkl}\left||F_{\rm PH}(hkl) - F_{\rm PH}(\bar{h}\bar{k}\bar{l})| - |2F_{\rm H}\sin\alpha_{\rm P}|\right|}{\sum_{hkl}|F_{\rm PH}(hkl) - F_{\rm PH}(\bar{h}\bar{k}\bar{l})|}. \tag{2.2.5.2}$$

Phasing power, $\mathbf{PP_{iso}}$. The isomorphous phasing power $\mathrm{PP_{iso}}$ for phase determination by isomorphous replacement is defined for a particular pair of native and heavy-atom-derivative data sets as the ratio of $|F_{\rm H}|$ and $\varepsilon(\varphi_{\rm P})$, where $|F_{\rm H}|$ is the calculated amplitude of the heavy-atom structure factor and $\varepsilon(\varphi_{\rm P})$ is the lack-of-closure error [equation (2.2.5.1*a*)]. $\mathrm{F_P + F_H}$ is a vector sum of the calculated structure factor $\mathrm{F_H}$ and the structure factor $\mathrm{F_P}$.

$$\mathrm{PP_{iso}} = \sum_{hkl}|F_{\rm H}| / \sum_{hkl}\left|F_{\rm PH} - |F_{\rm P} + F_{\rm H}|\right|. \tag{2.2.5.3}$$

There is another, slightly different, definition of $\mathrm{PP_{iso}}$, which is implemented in the program *SOLVE*. Here, $\mathrm{PP_{iso}}$ is defined as the ratio of the r.m.s. of the $|F_{\rm H}|$ values and the r.m.s. of the lack-of-closure errors $\varepsilon(\varphi_{\rm P})$. For each reflection *hkl*, a weighted average of $\varepsilon(\varphi_{\rm P})$ is calculated by integrating $\varepsilon(\varphi_{\rm P})$ over the whole phase circle.

$$\mathrm{PP_{iso}} = \Big(\sum_{hkl}|F_{\rm H}|^2\Big)^{1/2} / \Big(\sum_{hkl}\langle\left|F_{\rm PH} - |F_{\rm P} + F_{\rm H}|\right|\rangle^2\Big)^{1/2}. \tag{2.2.5.4}$$

Note. Owing to the cancelling out of the factor $N^{1/2}$ in the numerator and denominator, the latter $\mathrm{PP_{iso}}$ formula does not appear as a ratio of r.m.s. values at first sight.

Anomalous phasing power, $\mathbf{PP_{ano}}$. The anomalous phasing power $\mathrm{PP_{ano}}$ for phase determination by anomalous-dispersion methods is defined as the ratio of the sum of calculated anomalous differences $d''_{\rm calc}$ and the sum of estimated standard uncertainties $\sigma(d''_{\rm obs})$ in the measurement of these anomalous differences:

$$\mathrm{PP_{ano}} = \sum_{hkl} d''_{\rm calc} / \sum_{hkl}\sigma(d''_{\rm obs}). \tag{2.2.5.5}$$

As with $\mathrm{PP_{iso}}$ (see above), the program *SOLVE* uses a slightly different definition of $\mathrm{PP_{ano}}$. Here, the anomalous phasing power is defined as the ratio of the r.m.s. of the $d''_{\rm calc}$ values and the r.m.s. of $\sigma(d''_{\rm obs})$. For this, a weighted average of $d''_{\rm calc}$ is computed by integrating over the whole phase circle for each reflection.

$$\mathrm{PP_{ano}} = \Big(\sum_{hkl}\langle d''_{\rm calc}\rangle^2\Big)^{1/2} / \Big[\sum_{hkl}\sigma(d''_{\rm obs})^2\Big]^{1/2}. \tag{2.2.5.6}$$

Note. As above in the $\mathrm{PP_{iso}}$ formula, the factors $N^{1/2}$ in the numerator and denominator cancel out.

Figure of merit (f.o.m.), m. The figure of merit m is a term used in a number of contexts in X-ray crystallography. In its most common use, it is defined as the weight applied to an individual structure-factor amplitude that, in conjunction with its best phase, gives rise, in a Fourier synthesis, to the electron-density map with the minimum level of noise (Blow & Crick, 1959). Typically, m is given as an average value over all reflections in the data set or in a given resolution shell.

$$m = \int P(\alpha)\exp(i\alpha)\,\mathrm{d}\alpha / \int P(\alpha)\,\mathrm{d}\alpha = \langle\cos(\Delta\alpha)\rangle, \tag{2.2.5.7}$$

where $P(\alpha)$ is the probability of the phase α, initial or refined, being the best phase and $\Delta\alpha = \alpha_{\rm best} - \alpha$ is the error in the phase angle at α. The integration is from 0 to 2π and values for m range from 0 to 1.

2.2.6. Quality indicators for density modification and phase improvement

After determination of initial phases, a first electron-density map can be computed. It is expected that this map will contain significant errors and improbable features. Additional information, such as the flatness of the electron density in the solvent region or the similarity of electron-density regions of two or more identical molecules in the asymmetric unit, can be exploited to modify the electron density and hence improve the phases.

Chapter 3.2. Expression and purification of membrane proteins for structural studies

J. A. Ernst, D. G. Yansura and C. M. Koth

3.2.1. Introduction

Integral membrane proteins constitute about a third of the proteome of most organisms but less than 1% of all entries in protein structural databases (Berman *et al.*, 2002). This disparity is largely due to inherent difficulties in their expression, solubilization and purification. The production of sufficient protein for structural studies can be challenging for any target, but several obstacles are unique to membrane proteins. For example, proper insertion into the membrane relies on host cellular machinery that may be limiting or incompatible. Exceeding this capacity can lead to cell death or the accumulation of aggregated and inactive protein within the cell (Geertsma *et al.*, 2008). Also, post-translational modifications such as glycosylation, acylation and sulfation may not be faithfully reproduced (Grisshammer & Tate, 1995). As a further complication, even if sufficient expression can be achieved, most membrane-protein structural studies require that the target be extracted from the cellular membrane using detergents. However, detergents can adversely affect protein structure and function, as well as influence the outcome of crystal trials (Engel *et al.*, 2002; Lemieux *et al.*, 2003; Prive, 2007). Fortunately, if a strategy for purifying sufficient quantities of a given membrane-protein target *can* be established, crystallization strategies largely mimic the standard techniques for soluble proteins (Newby *et al.*, 2009).

Despite the aforementioned challenges, high-resolution structures of almost 200 unique membrane proteins have now been solved (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html), the vast majority using protein produced by recombinant methods (Willis & Koth, 2008). The rate of new structure determinations has also increased dramatically over the last few years, mimicking the exponential growth of soluble protein structures in the early 1980s. Space limitations here preclude a thorough review of all possible membrane-protein production methods. However, an examination of the successful expression, solubilization and purification strategies that have led to membrane-protein structures reveals that, in many cases, remarkably similar methods have been used (Carpenter *et al.*, 2008; Willis & Koth, 2008; Newby *et al.*, 2009). Guided by these methods, the following sections detail a rational and consensus 'first-attempt' strategy that has worked for a broad range of membrane targets with only minor variations in technique (Dobrovetsky *et al.*, 2005, 2007; Lunin *et al.*, 2006). It must be noted that while this represents an evidence-guided approach, the methods provided herein will not work for every membrane protein and, in fact, will fail for many. Thus, prioritized lists of alternative strategies are provided for those targets in which the initial expression or isolation attempts do not succeed or problems are encountered.

3.2.2. A consensus strategy for membrane-protein expression

Historically, the likelihood of successful structure determination has been high for those membrane proteins that can be isolated from readily available abundant natural sources (Sakai & Tsukihara, 1998). Unfortunately, most membrane proteins do not meet this criterion. In reality, the vast majority of high-resolution membrane-protein structures are of prokaryotic targets, expressed in *Escherichia coli* by recombinant methods (Willis & Koth, 2008; Newby *et al.*, 2009). The reasons for this are simple. Attempting recombinant expression in *E. coli* is inexpensive, flexible, simple and easily scaled-up, and many constructs and strains can be screened quickly. As with crystallographic efforts for many soluble proteins (Gräslund *et al.*, 2008), most successful membrane-protein endeavours also commonly screen multiple constructs and/or orthologues for any given target (see, for example, Chang *et al.*, 1998; Doyle *et al.*, 1998), since the greater the number of unique constructs screened, the greater the chance that one will be successfully isolated and crystallized. Given the overwhelming use of *E. coli* for successful membrane-protein structures, a 'first-pass' expression strategy suitable for almost any membrane protein is clear: attempt expression in *E. coli* and, whenever possible, screen multiple constructs, orthologues and strains.

3.2.2.1. Choosing the expression system and affinity tags

In addition to the choice of expression host, one must also consider the expression system and the type and placement of affinity tags or fusion proteins. For *E. coli*, IPTG-inducible T7 polymerase-driven expression systems, such as those based on pET vectors (Studier & Moffatt, 1986) and λDE3 lysogen strains, are the most widely used for membrane proteins, as is the case for soluble targets. Also, it has generally been observed that for most targets, protein expression is optimal at lower temperatures (*i.e.* <20 °C; Christendat *et al.*, 2000; Wang *et al.*, 2003; Dobrovetsky *et al.*, 2005). When structural studies are the desired outcome, the most common tagging strategy for membrane proteins is to engineer a stretch of at least six histidine residues at the amino or carboxyl terminus of the target constructs; this is used for >80% of successful targets (Willis & Koth, 2008). If expression levels are sufficient (*i.e.* $\geq$0.05 mg g^{-1} cell paste), this often permits purification using a general two-step procedure: capture of the tagged protein by immobilized metal affinity chromatography (IMAC), followed by size-exclusion chromatography (SEC) (see below). Remarkably, this basic approach has proved successful for the crystallization of many membrane proteins, as discussed in the following section. If premature termination is observed during expression, the engineering of a carboxyl-terminal tag will ensure that these proteins are not isolated during purification. Also, extending the stretch of histidines to greater than six residues can improve the retention of membrane proteins on immobilized metal affinity resins. This can prove particularly useful, given the modest expression levels and reduced chromatographic resolution and recovery of many membrane targets (Dobrovetsky *et al.*, 2005; Eshaghi *et al.*, 2005; Surade *et al.*, 2006; Lewinson *et al.*, 2008). Other affinity tags are viable options, but are more rarely used. For example, although fusion proteins such as glutathione S-transferase (GST) and thioredoxin are widely used to promote expression and/or simplify the purification of soluble proteins [for an extensive comparison of various affinity tags, see Lichty *et al.* (2005)], they are rarely cited in expression strategies for polytopic membrane proteins.

cules can induce the formation of non-specific aggregates, alter macromolecular solubility, or interfere with nucleation and crystal growth (McPherson *et al.*, 1996, 2004; Moreno *et al.*, 2005; Thomas *et al.*, 1998). These effects are reduced in gel media (Hirschler *et al.*, 1995; Provost & Robert, 1995).

On the other hand, macromolecules that are apparently pure may be microheterogeneous in sequence and/or conformation. Their causes are multiple and can be revealed by analytical tools, such as polyacrylamide gel electrophoresis, isoelectric focusing, nuclear magnetic resonance and mass spectroscopy, but are often overlooked. The most common causes are partial hydrolysis and post-synthetic modifications. Proteolysis represents a major difficulty that must be overcome during protein isolation. In RNAs, hydrolytic cleavages induced by nucleases, metal ions or alkaline pH are common causes of microheterogeneity. These processes can be inhibited by addition of protease or nuclease inhibitors and metal chelators during purification (Ducruix & Giegé, 1999). Heterogeneity in post-synthetic modification patterns in proteins or RNAs can be the result of functional necessity but can also occur when cloned macromolecules are overproduced. Conformational heterogeneity may also originate from ligand binding, intrinsic flexibility of the macromolecule backbones, oxidation of cysteine residues or partial denaturation. Structural homogeneity may be improved by truncation of the flexible parts of the macromolecule under study (see Section 4.1.3). This can also be done *in situ* by addition of trace amounts of protease to crystallization assays (Wernimont & Edwards, 2009). The many reasons that can account for such subtle degradations or modifications explain why altered versions of a macromolecule can be the worst contaminants for its crystallization. Accordingly, the macromolecule itself must be considered as an essential parameter in crystallization (Dale *et al.*, 2003). Control of these phenomena is of crucial importance for the crystallization of macromolecular assemblages, such as the ribosome (Auerbach-Nevo *et al.*, 2005).

Many crystal growers have found a correlation between the outcome of crystallization assays (*i.e.* number of crystals, crystal habit, volume and best diffraction properties) and the quality of macromolecular samples. For this reason, one should never spoil a 'pure' batch by mixing it with another 'pure' one, that may differ as far as microheterogeneities or minute contaminants are concerned. Altogether, purity, good solubility, structural homogeneity and absence of aggregates are good criteria for protein crystallizability (D'Arcy, 1994; Ferré-D'Amaré & Burley, 1997). Dynamic light scattering (DLS) is the appropriate analytical method to verify sample homogeneity, detect aggregates and find solvent conditions that prevent aggregation (Mikol, Hirsch & Giegé, 1990; Borgstahl, 2007; Niesen *et al.*, 2008).

4.1.3. Crystallization arrangements and classical methodologies

Many methods can be used to crystallize macromolecules (Ducruix & Giegé, 1999; McPherson, 1982, 1999). They all aim to bring the macromolecules to an appropriate state of supersaturation (Fig. 4.1.3.1). Although vapour-phase equilibrium and dialysis techniques are favoured, batch and free interface diffusion methods are often used. Besides the physical and chemical variables that affect crystallization (Fig. 4.1.1.1), macromolecular crystal growth is influenced by the crystallization method itself and the geometry of the setup. Generally, in current methods, growth is promoted by the non-equilibrium nature of the crystallization process, which seldom occurs at constant protein concentration. This introduces changes in supersaturation and, hence, may lead to changes in growth mechanism. Crystallization at constant protein concentration, however, can be achieved in special arrangements based on liquid circulation cells (Vekilov & Rosenberger, 1998).

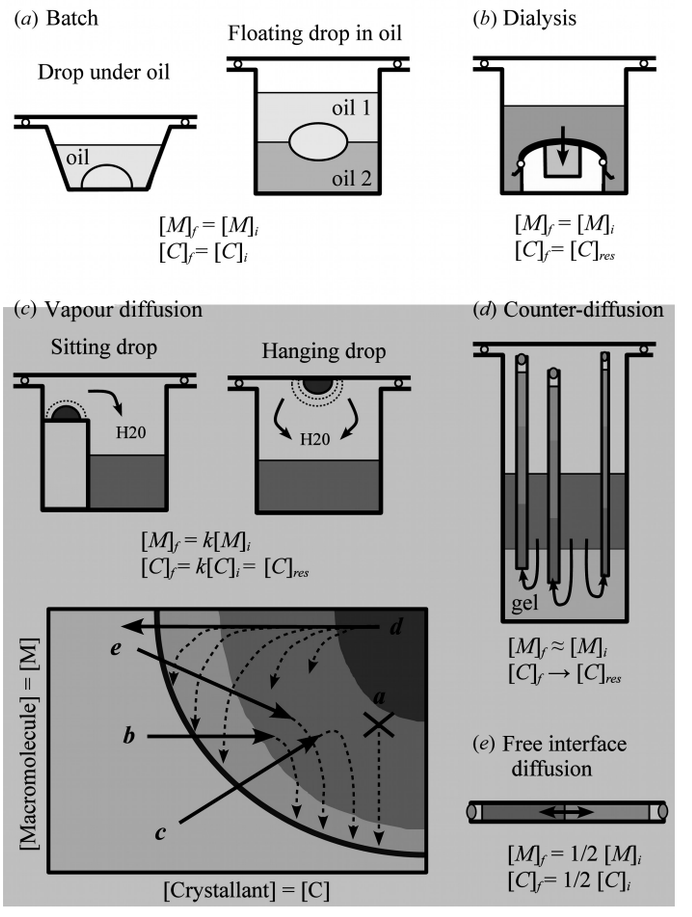


Figure 4.1.3.1
Principles of major methods used to crystallize biological macromolecules. (*a*) Two versions of batch crystallization. (*b*) Dialysis method with Cambridge button. (*c*) Vapour-diffusion crystallization with sitting and hanging drops. (*d*) Counter-diffusion setup in X-ray capillary tubes pinched in a gel layer. (*e*) Crystallization by free-interface diffusion in a capillary where two solutions of equivalent volume are brought into contact. The evolution of the macromolecule concentration, $[M]$, and crystallant concentration, $[C]$, in the different methods is indicated (initial and final concentrations in the crystallization solutions are $[M]_i$, $[M]_f$ and $[C]_i$, $[C]_f$, respectively; $[C]_{res}$ is the concentration of the crystallant in the reservoir, and k is a dilution factor specified by the ratio of the initial concentrations of crystallant in the drop and reservoir. In practice, glass vessels in contact with macromolecules should be silicone-treated in a way to obtain hydrophobic surfaces. Typical equilibration trajectories are illustrated in the phase diagram by black arrows, and the evolution of drops after nucleation and during growth are indicated by dashed arrows.

4.1.3.1. Historical development of methods

Protein crystallization is an old field that started more than 100 years ago (McPherson, 1991). Early methods included protein extractions with salts or organic solvents, or dialysis of salt solutions against water, and they were carried out on the gramme scale. Batch crystallization was the method of choice at that time. A first breakthrough that paralleled the development of X-ray methods occurred in the 1960s with the development of micromethods such as dialysis and vapour diffusion (with protein at the 1–100 mg scale and crystallization assays in the 10–50 µl range). However, it became rapidly apparent that screening of the entire

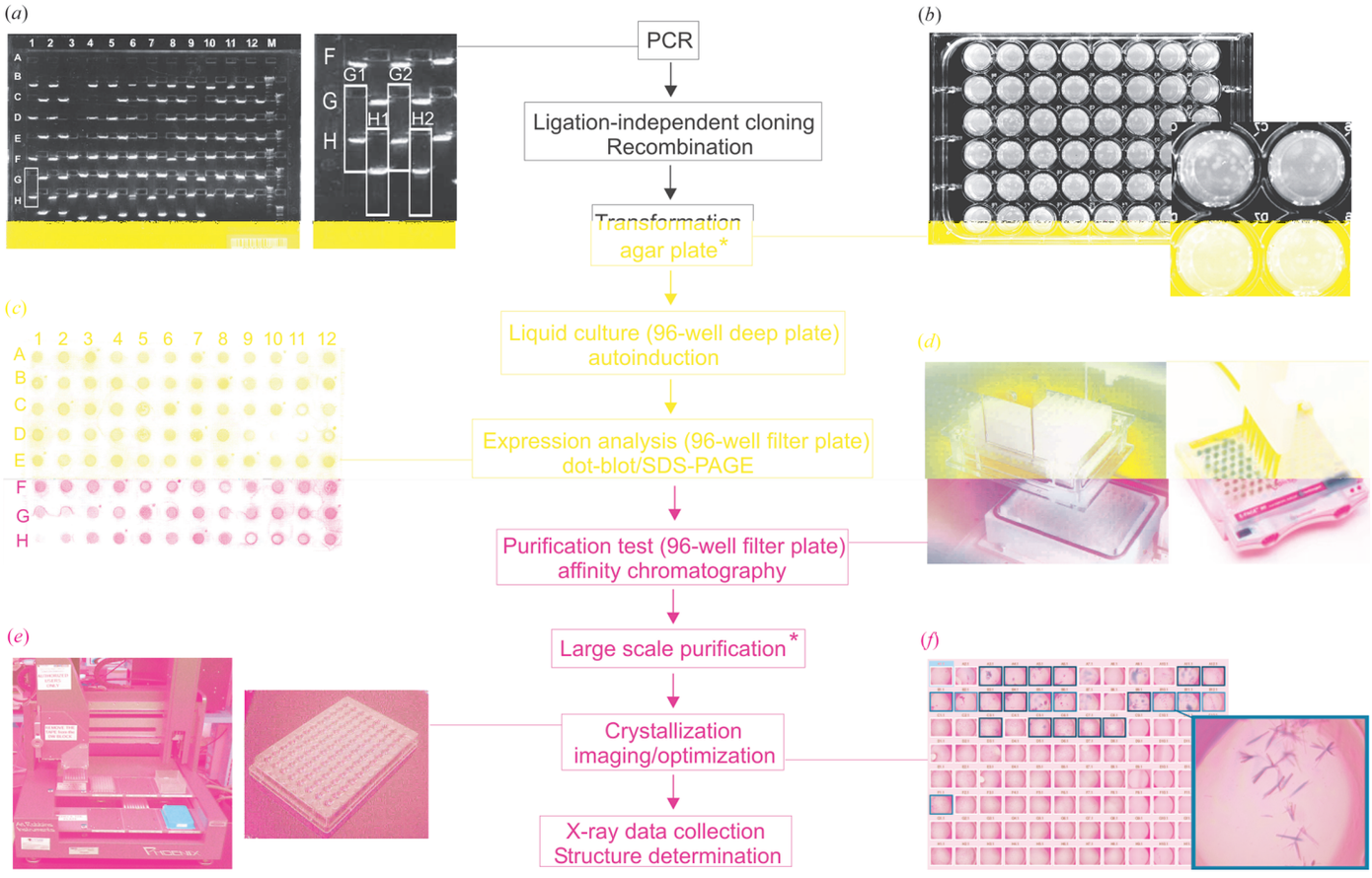


Figure 4.4.3.1
Diagram of HT cloning, expression and purification. All steps can be performed in a 96-well format except the step labelled with an asterisk (*). (*a*) PCR amplification. PCR products are loaded onto an E-gel (1% agarose) in a 96-well format with a molecular marker (lane 'M') and visualized with ethidium bromide. A close-up view of the gel outlining individual wells is also shown. (*b*) Transformation. Transformed cells were plated in 48-well agar plates. (*c*) Protein expression was tested using a western dot blot. H1 and H7 are negative and positive controls, respectively. (*d*) Small-scale purification using affinity resin. A 96-well filter plate in small-scale protein purification steps (left) and a 96-well pre-cast protein gel electrophoresis system (right) are shown. (*e*) Crystallization robot for setting up 96-well crystallization plates (right). (*f*) Crystal imaging and scoring system. A close-up view of an image of well B10 is shown on the right.

expression in *E. coli* is the most common, and will be discussed here.

4.4.3.1. Ligation-independent and recombination cloning

Ligation-independent cloning (LIC) strategies remove the need for restriction enzyme digestion and ligation of PCR products, and are thus ideal for use in an HT cloning procedure. In LIC, PCR primers are designed to append sequences that, after treatment with T4 DNA polymerase, generate 12- to 15-base overhangs that are complementary to overhangs in the vector (Aslanidis & de Jong, 1990). These longer cohesive ends make the insert–vector complex sufficiently stable to allow the transformation of hosts without ligation of the fragments.

The recombination strategy is based on the site-specific recombination reaction involved in bacteriophage λ integration and excision (Hartley *et al.*, 2000). PCR primers are designed to contain the specific sequences of recombination at 5′ and 3′ ends of the target gene, and the resulting PCR product is subcloned into a shuttle vector *via* site-specific recombination in the presence of integrase, integration host factor proteins and excisionase (*e.g.* the Gateway cloning system from Invitrogen). This clone can then either be isolated after transformation or directly used without purification for cloning into an expression vector. The major advantage of recombination cloning is that it provides a convenient way to shuttle an insert from one vector to another and thus is useful to test multiple expression conditions.

4.4.3.2. Practical application

Since the PCR reaction must be synchronized for all 96 wells in a plate, PCR primer design is important for the success of the PCR reaction. All PCR primers should have similar melting temperatures, between 50 and 60 °C. PCR products and the prepared vector are mixed and transferred into competent cells aliquoted into a 96-well plate. The PCR products and other DNA samples during the cloning, including the restriction-enzyme digest or a plasmid preparation (if needed), can be analysed *via* gel electrophoresis using a commercially available 96-well agarose gel (E-gel 96 from Invitrogen) in less than 15 min. The gel is compatible with a 96-well format and can be loaded either with a liquid-handling robot or multi-channel pipettor (Fig. 4.4.3.1*a*). The PCR step usually results in success rates as high as 98%. DNA purification kits are available in a 96-well format that use a vacuum manifold. As described above, the restriction-enzyme digest and ligation steps are not necessary in some cloning strategies.

Following transformation, cells are then plated into two 48-grid agar growth plates mixed with appropriate antibiotics (Fig. 4.4.3.1*b*). The 48-grid agar plate is made by inserting a cloning

chromosome because of morphological changes which occurred during data collection. Hence, full three-dimensional imaging of cells by diffractive methods requires cryogenic protection against radiation damage. Predictions based on a calculation of the cross section for coherent scattering by a smooth dielectric indicate that 10 nm resolution imaging of frozen hydrated organic matter should be possible using soft X-rays at currently available synchrotron sources (Howells *et al.*, 2009). This limit is arrived at through a comparison of the radiation dose required for imaging and the dose at which radiation damage has been empirically observed at different length scales. However, it seems plausible that the presence of many identical particles within a cell could be exploited to provide super-resolution information.

9.3.3.2. Low-dose three-dimensional imaging; low damage potential of stereoscopic viewing

Diffraction imaging in three dimensions proceeds as it does in standard X-ray tomography. That is, two-dimensional data are recorded from many angular orientations of the sample and then assembled into a three-dimensional data set. In this case, however, the data are recorded in reciprocal space and the individual data sets only need to be registered with respect to the angular coordinate due to the Fourier shift theorem. In the absence of any additional information, the angular sampling of reciprocal space is determined by the Crowther resolution,

$$k_C = \frac{1}{\Delta\theta D},$$

where D is the object diameter and $\Delta\theta$ is the angular separation of the two-dimensional data sets. This is the spatial frequency at which the unmeasured Fourier components, those in between the measured Ewald sphere segments, can be properly interpolated from the measured data. Diffraction microscopy, however, requires the addition of information in the form of real-space constraints. This additional information allows for the calculation of not only the missing reciprocal-space phases but also a limited number of missing magnitudes. Chapman *et al.* (2006) showed that $\Delta\theta$ could in fact be up to four times larger than required by the Crowther relation, with k_C matching the numerical aperture of the imaging system. Thus, three-dimensional reconstructions could take place with nearly isotropic diffraction-limited resolution with only about 150 angular orientations of the sample.

Stereoscopic viewing can provide a significant degree of three-dimensional perception of an extended object while only increasing the total radiation exposure by a factor of two. In principle, according to the dose-fractionation theorem of Hegerl and Hoppe, full three-dimensional visualization of a given resolution element should not require a dose any higher than two-dimensional visualization of the same element with the same statistical accuracy (Hegerl & Hoppe, 1976). This theorem provides hope that high-resolution imaging in three-dimensions, perhaps even of dry specimens, is possible, but in practice this is very difficult to achieve and low-dose imaging techniques are only now being explored by the CXDM community. Stereoscopic viewing should be considered the preliminary low-dose technique of choice. One particular advantage is the rapid reconstruction (compared with full three-dimensional reconstructions) which makes possible *in situ* sample inspection. Fig. 9.3.3.3 shows a stereo image of a chemically dried budding yeast cell. When viewed stereoscopically, with the viewer's focus in front of the image, the three-dimensional arrangement of a group of vesicles in the mother cell can be visualized.

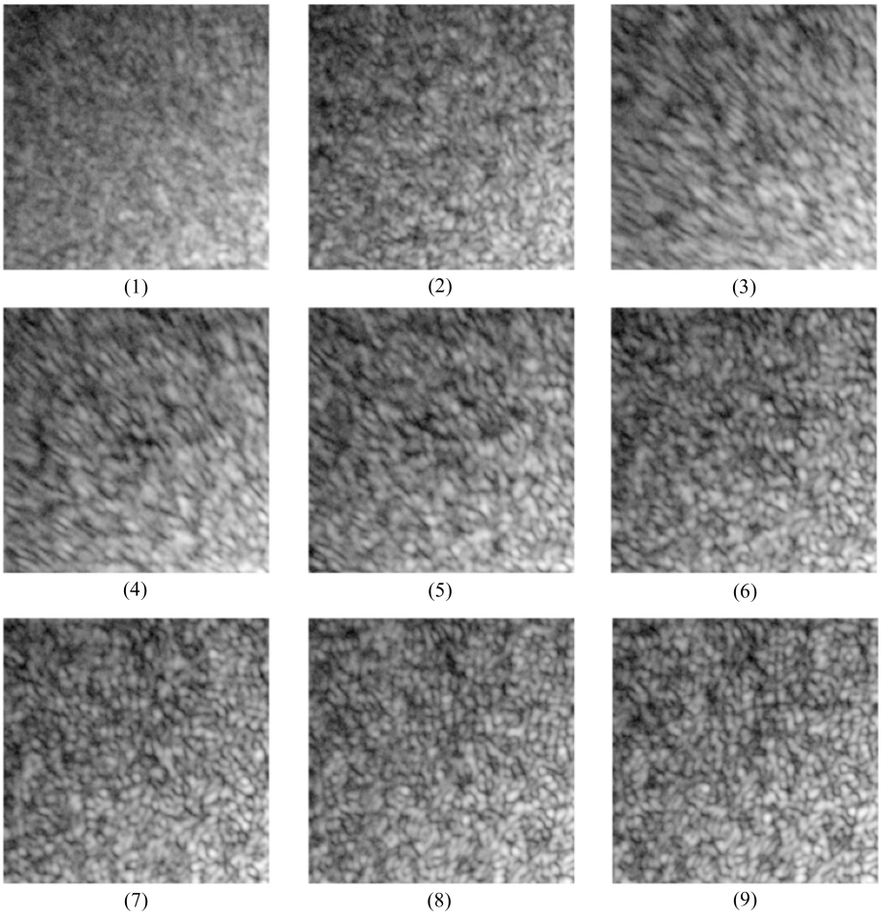


Figure 9.3.3.2
Exposure to ionizing radiation causes shrinkage of organic matter. Each image in this series is a section of a measured diffraction pattern from a freeze-dried yeast cell. The images were taken sequentially and each represents an additional X-ray dose of 5×10^8 Gy. After a cumulative dose of 1×10^9 Gy, the cell undergoes a rapid collapse [apparent from the elongated speckles in images (3)–(5)] followed by continued shrinkage at a reduced rate. The X-rays used had an energy of 750 eV and a dose of 5×10^8 Gy was adequate for reconstruction at 30 nm resolution. (Reproduced from Shapiro, 2004).

9.3.4. Conclusions

CXDM promises to be a highly efficient imaging methodology which can deliver high-resolution and high-contrast images of large non-crystalline biological structures. Radiation-induced shrinkage of dry cells will probably prohibit three-dimensional imaging of such cells at high resolution. However, significant three-dimensionality can be achieved through stereoscopic viewing of a cell, which only doubles the necessary X-ray dose. Even so, in the absence of low-dose diffraction techniques the sample must undergo considerable morphological change prior

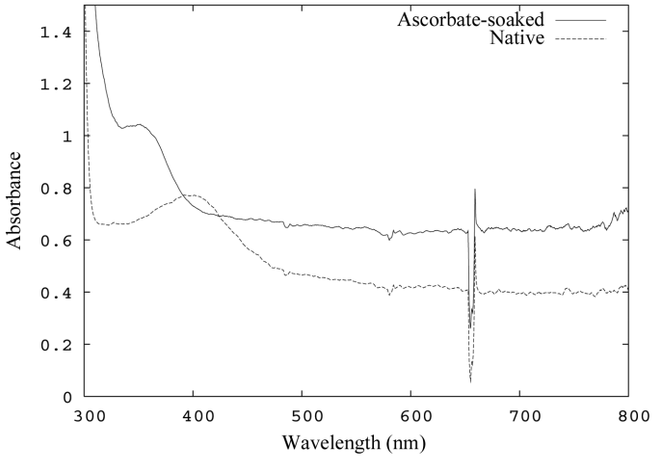


Figure 10.3.4.1
Microspectrophotometer absorption spectra of native and ascorbate-soaked hen egg-white lysozyme crystals at 100 K, showing the disulfide radical 400 nm peak formed on irradiation by a synchrotron X-ray beam, and the suppression of this peak on the addition of ascorbate, which has an absorption peak at 350 nm (Murray & Garman, 2002).

To account for the observed damage effects at 100 K, information must be gleaned from the radiation chemistry literature. Several analyses of the mobility of the various species formed upon X-ray exposure have been carried out and they help to explain the effectiveness of cryocooling for MX. Protons are only known to become mobile in amorphous ice above ∼115 K (Fisher & Devlin, 1995) and although $^{\bullet}$OH radicals are trapped at 77 K in an aqueous glass of 6 *M* CsF (Becker *et al.*, 1994) and are inferred to be trapped in ice at 100 K but are thought to be mobile above 130 K and thus able to recombine (Symons, 1999), they have also been reported to be mobile at 77 K in a glass of DNA (Lange & Httermann, 1995). According to electron spin resonance measurements (Jones *et al.*, 1987), at 77 K positive holes in proteins are rapidly trapped, forming amido radicals on the protein backbone chain, whereas the electrons produced by inelastic interactions have significant mobility. Rao *et al.* (1983) showed that electrons added to proteins at 77 K are able to move efficiently until they encounter S—S bonds, where they are trapped.

Complementary methods are increasingly being employed for the observation of radiation damage in MX and specially designed instruments are being constructed on synchrotron beamlines to facilitate measurements performed simultaneously with X-ray diffraction, including UV–vis microspectroscopy (McGeehan *et al.*, 2009), UV–vis fluorescence, X-ray spectroscopy (Yano *et al.*, 2005) and Raman spectroscopy (Owen *et al.*, 2009). Optical peaks at 400 nm and between 550 and 600 nm are signatures of the formation of the disulfide radical anion, *R*SS*R*$^{\bullet}$ (Fig. 10.3.4.1) (Weik *et al.*, 2002), and the hydrated electron, respectively. The reduction of metal sites by the X-ray beam can also be observed online from their change in optical absorption (Hough *et al.*, 2008).

10.3.5. Mitigating and correcting for radiation damage

Various experimental parameters have been investigated to find ways of reducing the rate of radiation damage at cryotemperatures, including further reducing the temperature, changing the incident wavelength and adding radical scavengers.

Various studies have investigated the rate of global and specific damage infliction with temperature variation, and there is as yet no consensus on this issue. At 40 K, global indicators improved little (Teng & Moffat, 2002), but a study at 8 K found a 25% increase in dose tolerance (Chinte *et al.*, 2007). Another investigation, which also monitored global indicators, reported no improvement for insulin crystal dose tolerance at 15 K, but a 23% improvement for holoferritin, which has a large iron core (Meents *et al.*, 2007). However, significant protection against specific damage (a factor of 30 at 40 K compared to 110 K) has been observed as monitored by the intensity of the photoreduced peak in the X-ray absorption spectrum at the iron *K* edge on an iron-containing metalloprotein (Corbett *et al.*, 2007).

At current fluxes (4×10^{14} photons s^{-1} mm^{-2}), heating of the sample by the beam is predicted to be no more than 15 K (Mhaisekar *et al.*, 2005), so nitrogen cooling to 100 K is adequate to avoid the movement of species other than the already mobile electrons and perhaps positive holes (Jones *et al.*, 1987).

Despite much anecdotal evidence that longer-wavelength incident radiation causes swifter damage than shorter-wavelength incident radiation, systematic studies of this relationship (Shimizu *et al.*, 2007; Weiss *et al.*, 2005) have shown no detectable change in the damage rate with dose over the range 6.2 keV (2 Å) to 33 keV (0.275 Å). Although the absorption is higher at longer wavelengths, the probability of diffraction is also higher, and the ratio of scattering to dose is not predicted to change by more than 20% for a typical protein crystal (no heavy atoms) over the range of energies currently used in MX (7–20 keV).

Two scavengers have been tested at RT in the past (styrene and polyethylene glycol), but neither was found to be very effective and they have not been widely utilized. However, recently, 0.5 *M* ascorbate and 0.5 *M* 1,4-benzoquinone were reported to give an increase by factors of two and nine, respectively, in $D_{1/2}$ at RT. Most interestingly, for both scavengers, the dose dependence of the intensity decay was modified to a linear relationship rather than exhibiting the exponential intensity decay of the native crystal (Barker *et al.*, 2009). The decay of the total diffraction intensity observed at 100 K is predominantly linear (Fig. 10.3.3.1), and this result indicates that the radiation chemistry of the degradation of the crystalline order has been modified by the scavengers in a similar way as it is by temperature if cryocooled.

At 100 K, several scavengers have now been tested: styrene was reported to be ineffective but ascorbate reduced both global and specific damage (Murray & Garman, 2002). Nicotinic acid and 5,5′-dithiobis-2-nitrobenzoic acid (DTNB) have been found to be effective (Kauffmann *et al.*, 2006). A large number of potential scavengers were screened using an online microspectrophotometer (McGeehan *et al.*, 2009) for their ability to quench the formation of the disulfide radical anion 400 nm peak, and only ascorbate, 1,4-benzoquinone and TEMP (2,2,6,6,-tetramethyl-4-piperidone) eradicated it (Southworth-Davies & Garman, 2007). Thus, from the research carried out so far at 100 K, it seems that scavengers deserve further study but are not going to give a large improvement in dose tolerance and do not seem to be as potentially effective as they are at RT.

In addition to these experimental strategies, software development is underway to correct data for radiation damage 'after the event'. These methods include zero-dose extrapolation, which uses multiple measurements of the same reflection to estimate its probable intensity at the beginning of the experiment (Blake & Phillips, 1962; Diederichs *et al.*, 2003), and the occupancy refinement of heavy-atom sites as a function of dose (Schiltz *et al.*, 2004).

Chapter 11.3. Integration, scaling, space-group assignment and post refinement

W. Kabsch

11.3.1. Introduction

The key steps in the processing of diffraction data from single crystals involve (i) modelling of the observed reflection positions in the detector plane, (ii) integration of diffraction intensities, (iii) data correction, scaling and post refinement and (iv) space-group assignment. Much of the theory and many of the methods for carrying out these steps were developed about three decades ago for processing rotation data recorded on film and were subsequently extended in order to fully exploit the capabilities of a variety of electronic area detectors; some CCD (charge-coupled device) and multiwire detectors as well as a new pixel detector specially developed for data collection at synchrotron beamlines allow the recording of finely sliced rotation data because of their fast data read-out. In this article, the principles of the methods are described as employed by the program *XDS* (Chapter 11.6). These apply equally well to rotation images covering small or large oscillation ranges. A large number of other data-processing systems have been developed which differ in the details of the implementations. Some of these packages are described elsewhere in this volume (see Chapters 11.1, 11.2 and 11.4). The theory and practice of processing fine-sliced data have been discussed by Pflugrath (1997).

11.3.2. Modelling rotation images

The observed diffraction pattern, *i.e.* the positions of the reflections recorded on the rotation-data images, is controlled by a small set of parameters which must be accurately determined before integration can start. Approximate values for some of these parameters are given by the experimental setup, whereas others may be completely unknown and must be obtained from the rotation images. This is achieved by the automatic location of strong diffraction spots, the extraction of a primitive lattice basis that yields integer indices for the observed reflections and the subsequent refinement of all parameters to minimize the discrepancies between observed and calculated spot positions in the data images.

11.3.2.1. Coordinate systems and parameters

In the rotation method, the incident-beam wavevector $\mathbf{S}_0$ of length $1/\lambda$ (where λ is the wavelength) is fixed while the crystal is rotated around a fixed axis described by a unit vector $\mathbf{m}_2$. $\mathbf{S}_0$ points from the X-ray source towards the crystal. It is assumed that the incident beam and the rotation axis intersect at one point at which the crystal must be located. This point is defined as the origin of a right-handed orthonormal laboratory coordinate system $\{\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3\}$. This fixed but otherwise arbitrary system is used as a reference frame to specify the setup of the diffraction experiment.

Diffraction data are assumed to be recorded on a fixed planar detector. A right-handed orthonormal detector coordinate system $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\}$ is defined such that a point with coordinates X, Y in the detector plane is represented by the vector $(X - X_0)\mathbf{d}_1 + (Y - Y_0)\mathbf{d}_2 + F\mathbf{d}_3$ with respect to the laboratory coordinate system. The origin X_0, Y_0 of the detector plane is found at a distance $|F|$ from the crystal position. It is assumed that the diffraction data are recorded on adjacent non-overlapping rotation images, each covering a constant oscillation range Δ_φ, with image No. 1 starting at spindle angle φ_0.

Diffraction geometry is conveniently expressed with respect to a right-handed orthonormal goniostat system $\{\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3\}$. It is constructed from the rotation axis and the incident-beam direction such that $\mathbf{m}_1 = (\mathbf{m}_2 \times \mathbf{S}_0)/|\mathbf{m}_2 \times \mathbf{S}_0|$ and $\mathbf{m}_3 = \mathbf{m}_1 \times \mathbf{m}_2$. The origin of the goniostat system is defined to coincide with the origin of the laboratory system.

Finally, a right-handed crystal coordinate system $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ and its reciprocal basis $\{\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*\}$ are defined to represent the unrotated crystal, *i.e.* at rotation angle $\varphi = 0^\circ$, such that any reciprocal-lattice vector can be expressed as $\mathbf{p}_0^* = h\mathbf{b}_1^* + k\mathbf{b}_2^* + l\mathbf{b}_3^*$, where h, k, l are integers.

As shown in Section 11.3.2.2, the location of all diffraction peaks recorded in the data images can be computed from the parameters $\mathbf{S}_0$, $\mathbf{m}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, X_0, Y_0, F, $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$, φ_0 and Δ_φ. In addition, knowledge of the shape and extent of the diffraction spots is required for accurate estimations of their intensities. This can be achieved by a Gaussian model involving two parameters: the standard deviations of the reflecting range, σ_M, and of the beam divergence, σ_D (see Section 11.3.2.3). This leads to an integration region around the spot defined by the parameters δ_M and δ_D, which are typically chosen to be 6–10 times larger than σ_M and σ_D, respectively.

11.3.2.2. Spot prediction

Let $\mathbf{p}_0^*$ denote any arbitrary reciprocal-lattice vector if the crystal has not been rotated, *i.e.* at rotation angle $\varphi = 0^\circ$. Depending on the diffraction geometry, $\mathbf{p}_0^*$ may be rotated into a position fulfilling the reflecting condition. The required rotation angle φ and the coordinates X, Y of the diffracted beam at its intersection with the detector plane can be found from $\mathbf{p}_0^*$ as follows.

$\mathbf{p}_0^*$ can be expressed by its components with respect to the orthonormal goniostat system as

$$\mathbf{p}_0^* = \mathbf{m}_1(\mathbf{m}_1 \cdot \mathbf{p}_0^*) + \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}_0^*) + \mathbf{m}_3(\mathbf{m}_3 \cdot \mathbf{p}_0^*).$$

Rotation by φ around axis $\mathbf{m}_2$ changes $\mathbf{p}_0^*$ into $\mathbf{p}^*$,

$$\begin{aligned}\mathbf{p}^* &= D(\mathbf{m}_2, \varphi)\mathbf{p}_0^* \\ &= \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}_0^*) + [\mathbf{p}_0^* - \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}_0^*)]\cos\varphi + \mathbf{m}_2 \times \mathbf{p}_0^* \sin\varphi \\ &= \mathbf{m}_1(\mathbf{m}_1 \cdot \mathbf{p}_0^* \cos\varphi + \mathbf{m}_3 \cdot \mathbf{p}_0^* \sin\varphi) + \mathbf{m}_2\mathbf{m}_2 \cdot \mathbf{p}_0^* \\ &\quad + \mathbf{m}_3(\mathbf{m}_3 \cdot \mathbf{p}_0^* \cos\varphi - \mathbf{m}_1 \cdot \mathbf{p}_0^* \sin\varphi) \\ &= \mathbf{m}_1(\mathbf{m}_1 \cdot \mathbf{p}^*) + \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}^*) + \mathbf{m}_3(\mathbf{m}_3 \cdot \mathbf{p}^*).\end{aligned}$$

The incident-beam and diffracted-beam wavevectors, $\mathbf{S}_0$ and $\mathbf{S}$, have their termini on the Ewald sphere and satisfy the Laue equations

$$\mathbf{S} = \mathbf{S}_0 + \mathbf{p}^*, \quad \mathbf{S}^2 = \mathbf{S}_0^2 \implies \mathbf{p}^{*2} = -2\mathbf{S}_0 \cdot \mathbf{p}^* = \mathbf{p}_0^{*2}.$$

If $\rho = [\mathbf{p}_0^{*2} - (\mathbf{p}_0^* \cdot \mathbf{m}_2)^2]^{1/2}$ denotes the distance of $\mathbf{p}_0^*$ from the rotation axis, solutions for $\mathbf{p}^*$ and φ can be obtained in terms of $\mathbf{p}_0^*$ as

Chapter 16.3. *Ab initio* phasing of low-resolution Fourier syntheses

V. Y. Lunin, A. G. Urzhumtsev and A. Podjarny

16.3.1. Introduction

Low-resolution macromolecular structural information may be obtained for non-crystalline macromolecular objects by electron microscopy (see Chapter 19.6 and references therein) or small-angle X-ray scattering (see Chapter 19.3 and references therein), or by using X-ray free-electron lasers (Chapman *et al.*, 2011; Seibert *et al.*, 2011) and iterative density-reconstruction techniques (Sayre, 2008). While giving similar structural information, low-resolution crystallographic images are a more natural starting point on the way towards atomic models. The need to work at low resolution may be due to the limited diffraction power of the crystals or the failure of standard phasing methods.

A number of techniques have been suggested to estimate the phases of structure factors using various complementary sources of information, specific to a given crystal:

(i) Phasing by isomorphous replacement or anomalous scattering techniques using heavy atoms or their clusters may lead to good results, ranging from the pioneering structures reported at nearly 6 Å resolution (Green *et al.*, 1954; Perutz *et al.*, 1960; Black *et al.*, 1962) to viruses (for example, Harrison & Jack, 1975) and ribosomes (Ban *et al.*, 1998). However, working at low resolution requires special effort, high data quality and several sets of structure-factor magnitudes.

(ii) Molecular-replacement-type low-resolution searches can be carried out using conventional models, more or less coarsely (Valegård *et al.*, 1991; Jamrog *et al.*, 2003), with simplified models or with molecular envelopes (Jack *et al.*, 1975; Rayment *et al.*, 1982; Urzhumtsev & Podjarny, 1995; Hao, 2006). The simplified models, roughly reproducing molecular shapes obtained previously, may be a spherical shell for viruses (for example Johnson *et al.*, 1976; Chapman *et al.*, 1992), a sphere or a cylindrical shell for proteins and their complexes (Podjarny *et al.*, 1987; Harris, 1995; Andersson & Hovmöller, 1996; Lunin *et al.*, 2001), or several cylinders for α-helical proteins (Kalinin, 1980; Strop *et al.*, 2007).

(iii) Phasing techniques can be based on the different average diffraction power of the proteins, nucleic acids and bulk solvent, naturally or artificially introduced (Bragg & Perutz, 1952; Roth *et al.*, 1984; Carter *et al.*, 1990; Fourme *et al.*, 1995; Shepard *et al.*, 2000). These techniques require several crystals and their corresponding diffraction data sets.

(iv) In some cases, the classical direct phasing methods can be formally tried to obtain a molecular envelope at low resolution (Podjarny *et al.*, 1981; Carter *et al.*, 1990). Similarly, the phase triplets obtained experimentally by a three-beam diffraction experiment lead in some cases to relatively low-resolution images (Hölzer *et al.*, 2000).

In this chapter we discuss only *ab initio* phasing methods. By this term we mean a mathematical or computational procedure to estimate the values of the structure-factor phases using a single set of structure-factor magnitudes and only a general type of information, in contrast with most of the methods mentioned above. Low-resolution *ab initio* phasing estimates the phase values of the structure factors of the lowest resolution for a given crystal, from several tens to several hundreds of reflections in total. Depending on the size of the unit cell, a low-resolution Fourier synthesis calculated with these structure factors can show molecular positions in the unit cell, molecular envelopes (at a resolution $d_{\min} \simeq 15$ Å or lower) or secondary-structure elements (when $d_{\min} \simeq 5$–6 Å).

16.3.2. General features of low-resolution images

Low-resolution *ab initio* phasing is usually a laborious procedure. Some features of low-resolution Fourier syntheses increase the difficulties, as follows:

(i) There is a common belief that low-resolution Fourier syntheses represent molecular envelopes when the cutoff level is relatively low and the centres of the molecules when this level is high. In practice, density peaks are often shifted from molecular centres toward regions of close intermolecular contact. Bulk-solvent correction decreases the contrast between two neighbouring peaks and can result in their merging. Also, due to the relatively small number of reflections used in the calculation, a small change in phase can significantly modify the image and the positions of the peaks.

(ii) Low-resolution envelopes cannot represent sharp molecular features accurately enough and thus cannot cover all macromolecular atoms, even when the synthesis is calculated with the exact structure-factor values.

(iii) An increase in the resolution of Fourier maps does not always help to interpret them. For example, an increase in resolution from 16–25 Å to 10–12 Å often makes maps even less suitable for visual inspection, since they stop showing molecular envelope features but do not reveal secondary-structure elements.

(iv) Usually, visual inspection of low-resolution maps does not allow one to choose the correct enantiomer. In addition, the overall features of the flipped map $-\rho(\mathbf{r})$ at very low resolution are often very similar to those of the direct map $\rho(\mathbf{r})$. This might complicate the choice of the correct sign of the map.

(v) Low-resolution Fourier maps are very sensitive to missing reflections, even if their number is small.

(vi) Maps calculated with too few reflections may show a superposition of images corresponding to different choices of the origin.

(vii) The determination of twinning and the correct choice of space group are especially complicated at low resolution.

16.3.3. Low-resolution phasing

If additional experimental information cannot be used to solve the phase problem, then the search for the correct structure has to be based on some general features of the true phase set. Such features can be formulated as a selection criterion ('score

PART 17. MODEL BUILDING AND COMPUTER GRAPHICS

Chapter 17.1. Macromolecular model building and validation using *Coot*

P. Emsley, B. Lohkamp and K. Cowtan

17.1.1. Introduction

Macromolecular model building using X-ray data is to some extent an interactive task, involving the iterative application of various optimization algorithms with the evaluation of the model and interpretation of the electron density by the scientist. *Coot* is an interactive three-dimensional molecular-modelling program particularly designed for building and validation of protein structures by facilitating the steps of the process.

In recent years, the initial construction of the protein chain has often been carried out using automatic model-building tools such as *ARP/wARP* (Langer *et al.*, 2008), *SOLVE/RESOLVE* (Wang *et al.*, 2004) and more recently *Buccaneer* (Cowtan, 2006). Therefore, relatively more time and emphasis has been placed on model validation than had previously been the case (Dauter, 2006). The refinement and validation steps become increasingly important and also more time consuming as data sets become poorer. *Coot* aims to provide convenient access to as many of the tools required in the iterative refinement and validation of a macromolecular structure as possible, to facilitate those aspects of the process which cannot be performed automatically. The software is also designed with a goal of being easy to learn, in order to provide a low barrier for scientists who are beginning to work with X-ray data (Fig. 17.1.1.1).

The principal tasks of the software are the visualization of macromolecular structures and data, the building of models into electron density, and the validation of existing models; these will be considered in the next three sections. The remaining sections of the chapter will deal with more technical aspects of the software, including interactions with external software, scripting and testing.

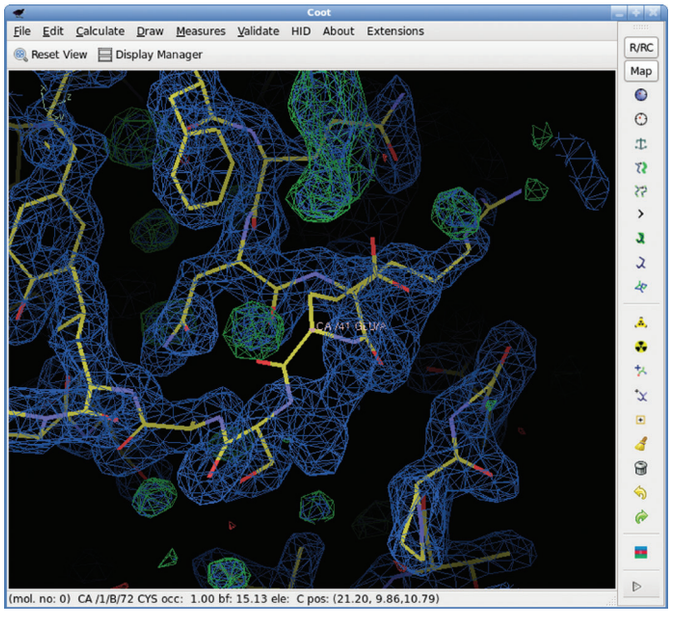


Figure 17.1.1.1
The *Coot* main window. The main display area shows a molecule and electron density. At the top of the window is a menu bar providing access to most of the tools. Commonly used model-manipulation tools are also available through the toolbar on the right. Below the menu bar is an area for user-definable buttons.

17.1.2. Model building

Initial building of protein structures from experimental phasing is usually accomplished by automated methods. The main focus in *Coot*, therefore, is the completion of initial models generated by either molecular replacement or automated model building. However, the following features are provided for cases where an initial model is not available.

17.1.2.1. Tools for general model building

17.1.2.1.1. Cα baton mode

Baton building, introduced by Jones [see for example Kleywegt & Jones (1994)] allows a protein main chain to be built by using a 3.8 Å 'baton' to position successive α-carbons at the correct spacing. In *Coot*, this facility is reproduced and coupled with an electron-density ridge-trace skeleton (Greer, 1974). First, a skeleton is calculated following the ridges of the electron density. Then, the user selects baton-building mode, which places an initial baton with one end at the current screen centre. Candidate positions for the next α-carbon are highlighted as crosses, selected from those points on the skeleton which lie at the correct distance from the start point. The user can cycle through a list of candidate positions using the 'Try Another' button, or, alternatively, rotate the baton freely by use of the mouse. Additionally, the length of the baton can be changed to accommodate moderate inaccuracies in the α-carbon positions. Once a new position is accepted, the baton moves so that its base is on the new α-carbon. In this way, a chain may be traced manually at a rate of between 1 and 10 residues per minute.

Having placed the α-carbons, the rest of the main-chain atoms may be generated automatically using an implementation of the method of *CALPHA* (Esnouf, 1997).

17.1.2.1.2. Find secondary structure

Protein secondary-structure elements, including α-helices and β-strands, can be located by their repeating electron-density features, which lead to high and low electron-density values in characteristic positions relative to the consecutive α-carbons. The 'Find Secondary Structure' tool performs a six-dimensional rotational and translational search to find the likely positions of helical and strand elements within the electron density. This search has been highly optimized in order to achieve interactive performance for moderately sized structures, and, as a result, is less exhaustive than the corresponding tools employed in the automated model-building packages. However, it can provide a

averaged. Either of these methods can be applied in a composite manner where small portions of the structure are omitted systematically over the entire map volume. Finally, simulated-annealing omit maps can be calculated by deleting the atoms to be omitted from the model, followed by simulated-annealing refinement and automatic map generation.

18.10.4.3. Real-space tools

18.10.4.3.1. Loop refinement

Polypeptide segments, both internal (loops) and external (tails), can be built from the target sequence as stored in the program. The program builds polypeptides according to an algorithm derived from the protein-modelling program *Prime* (Jacobson *et al.*, 2004; Zhu, Shirts & Friesner, 2007). Using the end of the existing structure as an anchor point, a residue is added and sampled in various conformations. All conformations that demonstrate even a modest fit to the electron density are kept. During the next addition, all stored conformations are used as starting points for sampling the conformation of the next residue. The total number of conformations may reach many thousands as additional residues are added. When all amino acids have been added, the conformations are clustered to reduce the redundancy of the set, treated with real-space minimization, scored by conformation (OPLS energy) and fitted to the electron density. When building loops, the structures are built from both ends of the structure gap. Only those pairs of built polypeptides that meet to form a contiguous loop are clustered, refined and evaluated. Typically, one to three of the highest scoring structures are returned by the program, depending on the quality of the electron density.

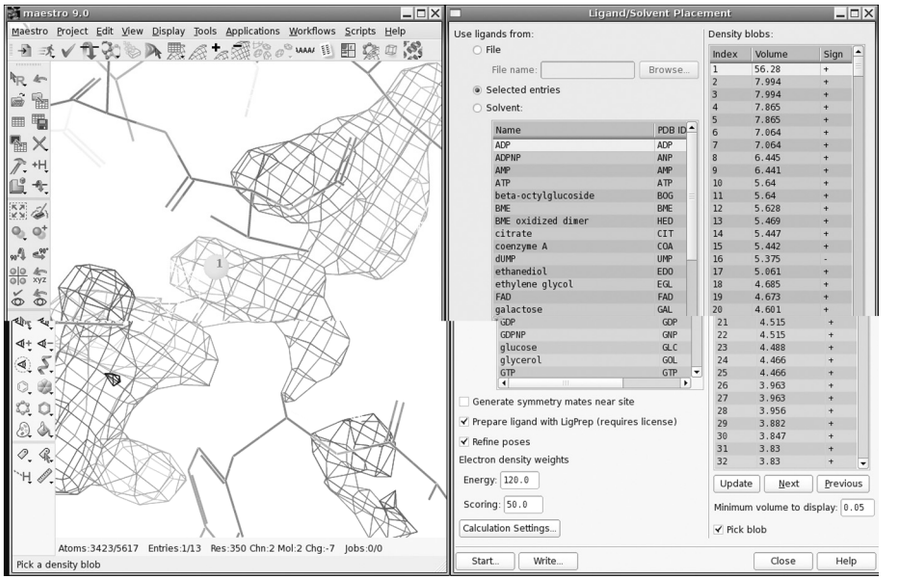


Figure 18.10.4.1
The ligand/solvent placement GUI provides a list of electron-density blobs from a map with coefficients $F_o - F_c$, together with their volume, and several ways to specify the molecules to be placed into the electron density. Multiple molecules may be included in a single placement operation at a single site. Ligand-placement options include the generation of charge and tautomer states of the ligand with the program *LigPrep* and the inclusion of symmetry-related molecules in the calculation at the ligand site.

18.10.4.3.2. Side-chain placement

Also based on *Prime* technology, one or multiple side chains may be simultaneously evaluated and placed into electron density. Even if two side chains have overlapping conformations at the start of the process, they will both be placed into appropriate electron density. As an alternative, various common rotamers of side chains may be explored interactively in *Maestro* and the best fit to the electron density selected visually or automatically.

18.10.4.3.3. Minimization

Real-space minimization provides an additional method of optimizing model coordinates according to the electron density and force field, as well as playing an integral part in the building of polypeptide chains, the placement of amino-acid side chains and the placement of ligands.

18.10.4.4. Water placement

Waters may be added by *PrimeX* to difference electron-density maps for positive peaks exceeding a selected level of significance. Prospective water sites in the electron-density map are screened by three different criteria related to distance from other atoms. Water molecules that refine to a B factor above a selectable cutoff are deleted.

18.10.4.5. Ligand placement

Ligand placement can be accessed through a convenient GUI (Fig. 18.10.4.1). The algorithm for ligand placement was derived from the docking program *Glide* (Friesner *et al.*, 2004). In brief, the program abstracts the protein model as an energy grid in which the properties of the residues are encoded. Multiple conformations are generated for the ligand, and then a search is made for the best possible position and orientation in the vicinity of the electron density that the user has identified. The many alternative solutions are reduced in number with the application of a dual scoring system: (i) the GlideScore is an evaluation of the chemical complementarity between the ligand and the protein; (ii) the DensScore is a measure of the fit of the ligand to the observed electron density. Both of these scores are used to rank the fitting of each ligand conformation in each orientation. The weighting of the DensScore *versus* the GlideScore is under user control and it strongly favours the DensScore by default. Surviving candidates are subjected to rigid-body and torsion-angle refinement against an energy function that includes the OPLS energy and the DensScore. The weighting of the DensScore *versus* the OPLS energy is also under user control, and it strongly favours the DensScore by default. Finally, the top few dozen candidates are treated with real-space refinement and are ranked by real-space R factor. The top five poses for the ligand are returned with the protein structure for evaluation by the user.

18.10.4.6. Validation

PrimeX provides three main validation tools. 'Protein reports' lists geometric variations from expected values in table form. This table can be sorted by residue or by the magnitude of the deviation, to highlight parts of the structure most in need of attention. A density fit table (Fig. 18.10.4.2) lists the real-space R factors for each main chain and side chain and for the entire residue. This table can also be sorted by any column. Clicking on any row of either table centres the molecular display on that

Chapter 18.12. Structure determination in the presence of twinning by merohedry

T. O. Yeates and M. R. Sawaya

18.12.1. Introduction

Twinning by merohedry (see Chapter 11.7) presents challenges for structure determination because the correct intensities are not observed directly. This is a key problem, since the ultimate goal of structure determination is to arrive at agreement between the observed crystallographic quantities and the corresponding values calculated from a model. Common strategies for handling this problem at different stages of the structure-determination process are considered here. In the following discussion, it is helpful to note that, in any given strategy, one of three approaches is generally being taken. Either (i) the effects of twinning are ignored and treated essentially as a source of noise, (ii) the effect twinning has on the observed intensities is reversed – a process called 'detwinning' – in an attempt to recover the true underlying crystallographic intensities, or (iii) the effects of twinning are applied to the model in order to match the twinning present in the observed data. The following discussions apply primarily to twinning by hemihedry, in which just two domain orientations are present. For the treatment of higher forms of twinning (Yu *et al.*, 2009; Barends *et al.*, 2005) or other types of disorder (Wang *et al.*, 2005; Tsai *et al.*, 2009; Hare *et al.*, 2009), the reader is referred to the recent literature.

18.12.2. Detwinning based on observed intensities

Equation (11.7.4.1) in Chapter 11.7 describes how the observed twinned intensities arise from linear combinations of true crystallographic intensities. By inverting that system of equations, one can obtain the crystallographic intensities from the observed intensities, assuming the twin fraction α is known:

$$I(\mathbf{h}_1) = \big[(1-\alpha)I_{\rm obs}(\mathbf{h}_1) - \alpha I_{\rm obs}(\mathbf{h}_2)\big]/(1-2\alpha),$$
$$I(\mathbf{h}_2) = \big[-\alpha I_{\rm obs}(\mathbf{h}_1) + (1-\alpha)I_{\rm obs}(\mathbf{h}_2)\big]/(1-2\alpha),$$

where I are the true crystallographic intensities for two reflections $\mathbf{h}_1$ and $\mathbf{h}_2$ related by the twin operation, and $I_{\rm obs}$ are the corresponding observed intensities. The form of this equation makes it clear that the observed data cannot be detwinned when $\alpha = 1/2$ (*i.e.* perfectly twinned). When $\alpha \neq 1/2$, the crystallographic intensities can be calculated but measurement errors in the observed intensities are magnified according to the term $1/(1-2\alpha)$ (Fisher & Sweet, 1980; Grainger, 1969). Therefore, detwinning based on observed intensities alone tends to be problematic when the twin fraction is high. If the twin fraction is low then detwinning is robust, although in those cases the effects of twinning are the least important to consider. As a result, detwinning based only on observed intensities is not practised often. The situation is different once calculated intensities from a model become available, as discussed below.

18.12.3. Molecular replacement with twinning

Even in the case of perfect twinning, it is often possible to arrive at a correct molecular-replacement solution without taking twinning into account, as demonstrated by Redinbo & Yeates (1993). At the rotation-function stage, the effects of twinning are much like the effects of high crystal symmetry; one is attempting to detect the correct orientation of a single search model in the context of noise due to other molecules in distinct orientations within the crystal. In favourable cases such a search can be successful. At the translation-function stage, recognizing the true space-group symmetry is critical. As described in Chapter 11.7, even the correct point-group symmetry can be misclassified in cases of high or perfect twinning. Therefore, it is generally advisable to evaluate potential molecular-replacement solutions in multiple space groups consistent with the data. In practice, twinning is usually ignored throughout the molecular-replacement process (Breyer *et al.*, 1999; Contreras-Martel *et al.*, 2001; Luecke *et al.*, 1998; Redinbo & Yeates, 1993; Larsen *et al.*, 2002; Heffron *et al.*, 2006; Yuan *et al.*, 2003; Lee *et al.*, 2003; Anand *et al.*, 2007), although some workers have detwinned the data and then proceeded with molecular replacement (Rabijns *et al.*, 2001; Taylor *et al.*, 2000). In evaluating potential molecular-replacement solutions, it could be useful to consider the contributions of both domain orientations to the observed intensities; this has generally not been done.

18.12.4. Multiple isomorphous replacement and anomalous phasing with twinning

Strategies for phasing twinned data by multiple isomorphous replacement (MIR) and anomalous-scattering methods have been reviewed (Terwisscha van Scheltinga *et al.*, 2003; Dauter, 2003). To date, several successful solutions of twinned structures by MIR and anomalous phasing approaches have ignored the effects of twinning, at least during the steps of heavy-atom substructure determination and phasing (Rees & Lipscomb, 1980; Mueller, Muller *et al.*, 1999; Mueller, Schübel *et al.*, 1999; MacRae & Doudna, 2007). As noted above, such an approach effectively treats the contributions from the alternate (minor) twin domain as noise. If a correct heavy-atom substructure can be obtained, from Patterson analysis for example, that solution reflects only the heavy-atom contents of one twin domain, presumably the dominant one. Additional peaks not directly accounted for by the identified heavy-atom positions will be present in the native Patterson map in twin-related positions. Adding to the potential confusion, cross-peaks between the heavy-atom positions identified and those related by the twin operation are not present in a Patterson map. Likewise, contributions of the minor twin domain to the measured differences between native and derivative structure-factor amplitudes are ignored when estimating the best phase for the native structure factor. In favourable cases, useful heavy-atom substructures and phases can be obtained despite these serious oversimplifications. In less favourable cases, some workers have overcome twinning in MIR experiments by detwinning the native and derivative data sets, then executing isomorphous replacement with detwinned structure factors (Declercq & Evrard, 2001; Terwisscha van Sheltinga *et al.*, 2001; Hillig & Renault, 2006; Ban *et al.*, 1999). In a few cases, results indicate that detwinning the data produces maps of a quality similar to or worse than the observed twinned data (Rudolph *et al.*, 2003; Yang *et al.*, 2000; Dong *et al.*, 2001).

Chapter 19.8. Use of *SPIDER* and *SPIRE* in image reconstruction

A. Leith, W. Baxter and J. Frank

19.8.1. Introduction

SPIDER (System for Processing Image Data in Electron microscopy and Related fields) was one of the earliest image-processing packages for single-particle reconstruction and electron tomography. It was first released in 1978, and has been continually improved and updated since then (Frank *et al.*, 1981, 1996; Baxter *et al.*, 2007; Shaikh *et al.*, 2008; Yang *et al.*, 2007). It is currently available as free open-source code under Gnu Public License (GPL).

In its most common usage, *SPIDER* allows a researcher to create a three-dimensional (3D) reconstruction of a macromolecule from a collection of transmission electron micrographs, interpreted as projections showing the molecule in different orientations. Using cryo-electron microscopy these methods, pioneered by Joachim Frank at the Wadsworth Center in Albany (see Frank, 2006), have recently allowed researchers to determine the 3D structures of several macromolecular complexes at near-atomic resolution (~3.8–4.5 Å; Zhou, 2008). These methods provide a way of studying the structure of such complexes in a more natural environment than is possible using X-ray crystallography and of studying conformational changes of molecular machines as they perform their work (Saibil, 2000).

One way in which *SPIDER* differs from most other software for single-particle reconstruction [*Imagic* (van Heel *et al.*, 1996; Chapter 19.9), *Frealign* (Grigorieff, 2007), *EMAN2* (Tang *et al.*, 2007; Chapter 19.10), *SPARX* (Hohn *et al.*, 2007) and *Xmipp* (Sorzano *et al.*, 2004)] is that it provides alternative methods for many procedures and thus can serve as a developmental platform for testing different approaches to single-particle reconstruction. For instance, alternative alignment operations and statistical analysis operations can be combined inside scripts to investigate different approaches to overcoming molecular heterogeneity.

The prime reasons for the continued success of *SPIDER* are its validated reputation for providing the highest-resolution reconstructions, its ease of installation and its comprehensive documentation, which is constantly being updated (Shaik *et al.*, 2008). *SPIDER*'s procedure language makes it easy to use for controlling the flow of a complex train of operations and addressing problems related to heterogeneity of particles. With the Quick Start Guide and tutorials as an introduction, a new user can quickly acquire competence in the procedure language with only a few hours' introduction.

The *SPIDER* system consists of six major components:

(i) *SPIDER*: the Fortran program with a command interpreter recognizing *SPIDER* commands and procedure calls.
(ii) *SPIDER* procedures: text files containing *SPIDER* commands, parameters, and script-specific operations for conditional execution and looping (see Section 19.8.3).
(iii) *Web*: a graphical user interface (GUI) for use in Linux and OS X (Frank *et al.*, 1996). *Web* is available in two different versions, the original X-Window version written in C and a newer Java version.
(iv) *SPIDER* reconstruction engine (*SPIRE*), which represents a metastructure that enables procedure files, file numbering and directories required for a single-particle reconstruction project to be managed from a simple GUI (see Section 19.8.4; Baxter *et al.*, 2007).
(v) *PubSub*: a set of Perl programs that enable *SPIDER* procedures to be run in parallel on a computer cluster.
(vi) *SPIDER* documentation: an extensive collection of more than 800 HTML-based documents.

The system is available in versions for Linux, AIX and OS X. Both source code and precompiled binaries for popular platforms are provided. *SPIDER* is available for free download as a GPL Open Source distribution. Its documentation is available under a Creative Commons Attribution 2.5 Licence.

19.8.2. Basic philosophy of single-particle reconstruction

'Single-particle reconstruction' is the term used for the reconstruction of a biological macromolecule from images of a specimen in which the molecule exists in many 'copies' in the form of single isolated particles, *i.e.* with no contact with neighbouring molecules. Since there is no need for crystallization, there is in principle no restriction on the kinds of macromolecules that can be reconstructed, except that they must be above a critical size required for accurate alignment. Combined with cryogenic electron microscopy (cryo-EM), the method is capable of visualizing molecules in their native states. Although the method was originally conceived for a homogeneous population, the introduction of powerful classification techniques is now allowing heterogeneous populations to be disentangled and represented by a series of reconstructions which, suitably ordered, may reflect the development of a system of interacting molecules (a molecular machine) over time.

Each particle image is interpreted as a noisy projection of the three-dimensional Coulomb potential representing the molecule, which is, for practical purposes, identical to the electron-density distribution rendered by X-ray crystallography. The noise (signal-to-noise ratio $\simeq 0.1$) is due mainly to the low exposure required to avoid radiation damage. For reconstruction, all projections must be placed in a common coordinate frame. Since, unlike the case in electron tomography, the angles are initially unknown, the most challenging and computationally intensive task in single-particle reconstruction is the determination of particle orientations, usually done in an iterative manner with increasing angular resolution. In the following, it is assumed that a reference density map of a closely related molecular complex is already available. For instance, a map of an empty ribosome may be used as reference for a data set obtained from a ribosome complexed with EF-G, mRNA and tRNA. For *ab initio* reconstructions of an unknown structure, random-conical and common-lines techniques are available (see Shaikh *et al.*, 2008).

Owing to the oscillatory behaviour of the contrast transfer function (CTF), an entire defocus series must be collected in order to cover the whole range of information in Fourier space. For each micrograph, the defocus is determined by computing the power spectrum and matching it with the CTF. Two strategies are in use in the field: CTF correction is done either at the stage of the raw micrograph, by phase flipping and pooling all data for

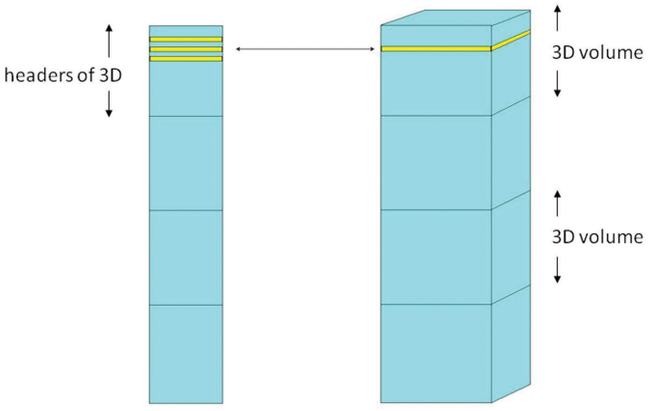


Figure 19.9.3.1
IMAGIC 4D file format. The 4D file format is not very different to the original file format with one header record per 2D image in a stack (van Heel & Keegstra, 1981). The main difference is that specific header locations now indicate the additional organization of a sequence of 2D images in the stack belonging together in a 3D volume. Substantial changes were required in the software, however, to make programs loop over the 3D volumes in one 4D *IMAGIC* file in the way that the programs looped over the 2D images in a stack. For full format details, see www.imagescience.de.

three-dimensional volumes to remain in core and to be manipulated efficiently by single subroutine calls without I/O (input/output) overhead. The file format for '4D' processing has hardly needed any updating other than defining some new header parameters. The reason is that a stack of three-dimensional (3D) volumes (a '4D' data format) is still just a stack of two-dimensional (2D) images, the traditional *IMAGIC* data format (Fig. 19.9.3.1). This '4D' upgrade, however, did require a radical overhaul of all programs, since all relevant programs needed an extra loop over a set of 3D volumes. At the same time, these 3D programs have all become much faster, more compact and more easily maintained because they rely on the new in-core 3D libraries and are now freed of all excess I/O calls.

19.9.4. Software parallelization

In cryo-EM all averaging and information extraction from all individual molecular images takes place *in silico*; the computational requirements are thus huge. Computers are never fast enough for the most demanding single-particle approaches. Over the last 15 years much emphasis has been placed on the parallelization of critical *IMAGIC* code (van Heel *et al.*, 2000), mainly using a message passing interface (MPI; Gropp *et al.*, 1994) to take advantage of modern 'cluster' computing environments. Other software packages have since followed the same parallelization path (Smith & Carragher, 2008). The *IMAGIC* software is implemented such that the same code will run on all machines from a single-CPU notebook computer up to large cluster systems with hundreds of CPUs.

Recently, a 'GPU' library (see www.nvidia.com/object/cuda_home.html) has been implemented in *IMAGIC* to exploit the properties of cheap graphic processors. However, for most 'standard library' operations, like 2D fast Fourier transforms, which can be almost directly linked to existing programs, most time gained in GPU processing is still lost in transporting data to and from the limited memory of the GPUs. The question is whether the time invested in software development for exploiting specific properties of any specific parallel computing hardware will pay off in the longer term for the specific needs of cryo-EM. In the case of the GPUs, the amount of memory typically available per GPU core is rather low.

19.9.5. Full 2D (parallel) astigmatic contrast transfer function correction

The *IMAGIC* contrast transfer function (CTF) estimation and correction programs (van Heel *et al.*, 2000) have recently been upgraded to work fully in two dimensions, enabling the accurate detection of all CTF parameters including astigmatism. This has been carried out in order to accomplish two goals. Firstly, the programs are now capable of operating on an entire data set of charge-coupled device (CCD) images or patches of micrographs. This now allows the use of the parallel MSA programs (see below) to classify sets of amplitude spectra (and create class averages thereof) prior to the precise determination of the defocus and astigmatism parameters. This enables the simple and largely automatic determination of CTF parameters, and subsequent CTF correction *via* phase flipping of entire data-set stacks. Secondly, the ability to accurately detect even extreme levels of astigmatism enables the use of highly astigmatic images, unlocking extremely close-to-focus defoci in order to push the achievable resolution. These programs have been used for the processing of a highly astigmatic data set of the *Limulus polyphemus* hemocyanin, resulting in a ∼4 Å reconstruction from only 15 000 raw molecular images, corresponding to 60 000 asymmetric units for this $C2$ point-group symmetry structure (Grant *et al.*, 2011).

19.9.6. Parallel automatic particle picking

Parallel processing is of increasing importance for collecting the large raw data sets required for 4D cryo-EM. If we are interested in a structure that represents 1% of all molecular complexes, we need to increase the size of the data set 100-fold in order to achieve the same resolution we had for a monodisperse data set. This implies that we need to be able to rapidly process data sets of the order of 1 Tbyte in size. One of the fastest automatic particle-selection procedures is still one of the oldest ones, based on the local variance (van Heel, 1982) in the raw micrographs or CCD images. It is applied immediately after the full-data-set CTF correction discussed above. Very important for the calculation of the variance image is the choice of the frequency range used for discriminating the presence of a particle with respect to the background. The sensitivity of the approach is approximately as good as that of particle searching using the cross-correlation function (CCF) (Saxton & Frank, 1977). However, like all CCF alignments (see below), the CCF particle picking requires templates and may tend to bias the particle selection towards the references (Boekema *et al.*, 1986; Stewart & Grigorieff, 2004). Modulation-image particle picking is a variant of the original variance image detection (van Heel, 1982) which is based on the local standard deviation rather than the local variance and avoids squaring of amplitudes. This new algorithm is often the method of choice for an unbiased automatic particle selection. Once a reliable first 3D structure has been calculated, the particle-picking program can then look for all possible views in all possible orientations in an extensive stack of input images (the massive calculations involved exploit MPI parallelization).

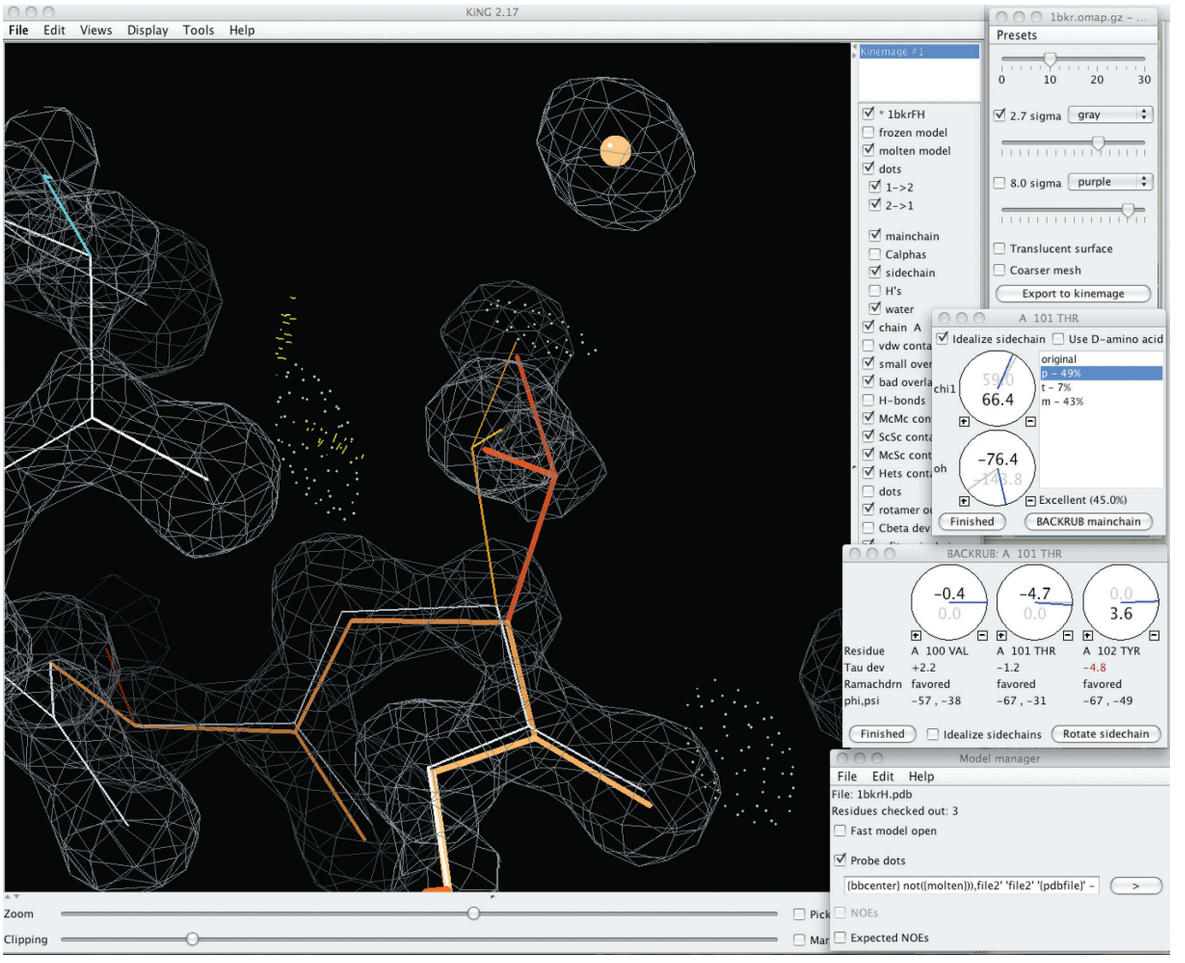


Figure 21.5.2.2
Screen capture of side-chain rebuilding in *KiNG* using the side-chain rotator and backrub tools, with display of electron density, all-atom contacts and rotamer quality. In this example from the 1.1 Å resolution crystal structure of the calponin homology domain of β-spectrin (PDB code: 1BKR) (Banuelos *et al.*, 1998), a threonine was fitted with its side chain flipped over by 180°. The original (in gold since it is a rotamer outlier) has large bond-angle outliers and steric clashes, and its Cβ is entirely out of density. Using the side-chain rotator tool in combination with the backrub tool allows a much better fit to the density (the thicker orange 'mobile' side chain), with an excellent rotamer, ideal geometry and two hydrogen bonds (pale green dot pillows) for the Thr OH. H atoms and van der Waals dots are turned off, for clarity in two dimensions.

21.5.2.4. Cystallographic rebuilding tools

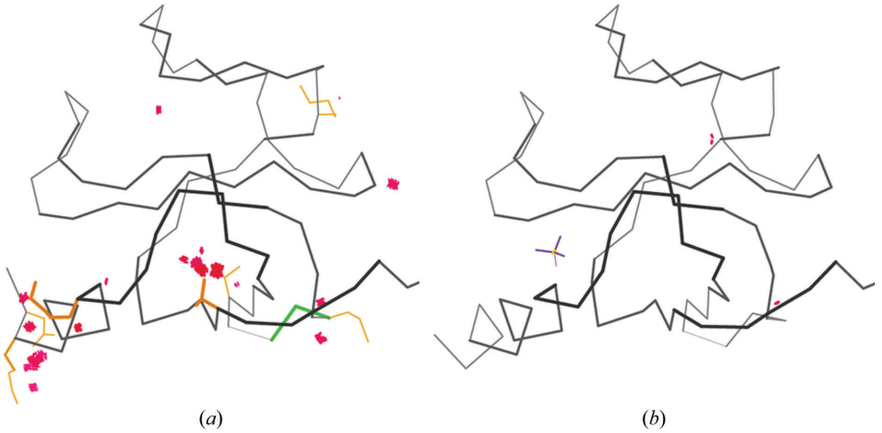
In crystallography, the most important use of all-atom contact analysis is for quickly finding, and frequently for fixing, problems with molecular geometry during fitting and refinement. All-atom contacts add independent new information to that process, since the hydrogen atoms make almost no contributions either to the electron density or to the energetic component of refinement as presently done, yet they are undeniably present and cannot significantly overlap other atoms (except in hydrogen bonds). The steric constraints implied by all-atom contacts are significantly more stringent than those based only on non-hydrogen atoms, yet they are obeyed almost perfectly by low-*B* regions of structures at resolutions near 1 Å, even when hydrogen atoms were not used in refinement.

At any stage of a structure determination, contact dots for the entire molecule or molecules can be calculated by *Probe* and examined in *KiNG*, or a list of the severe clashes and other outliers can be generated, for printout or as a scripted 'to-do' list for rebuilding in *Coot* (Emsley & Cowtan, 2004; Emsley *et al.*, 2010).

The most common way to do *MolProbity* analysis on a structure is on the web site (http://molprobity.biochem.duke.edu) (Davis *et al.*, 2007; Chen *et al.*, 2010; Chapter 21.6), which guides the entire process through to both chart and kinemage analysis of the local quality of a macromolecular structure. The resulting 'multi-criterion' kinemage flags all the individual outliers on the three-dimensional structure (Fig. 21.5.2.1): serious clashes ($\geq$0.4 Å overlap), poor side-chain rotamers, backbone bond-length or angle outliers, Cβ deviations, Ramachandran outliers and, for RNA, sugar-pucker and backbone conformer outliers (Richardson *et al.*, 2008). That kinemage can be viewed in *KiNG* online, but for rebuilding to correct the diagnosed problems a user would download the multi-criterion kinemage and the H- and flip-optimized PDB file for off-line work in *KiNG*. The multi-criterion kinemage shows where work is needed. However, the most powerful use of all-atom contact information is directly and interactively in the process of fitting and rebuilding. Therefore, the rebuilding tools in *KiNG* can call *Probe* on-the-fly during refitting, for display of clashes, hydrogen bonds and favourable van der Waals contacts (as well as updated rotamer quality and checks for φ, ψ and τ outliers) (Lovell *et al.*, 2000, 2003). This gives instant feedback on how user-proposed changes improve or worsen the model locally. These tools, including a side-chain mutation tool, a side-chain rotator tool, a backrub tool and an RNA rotator tool for adjusting RNA suite conformations,

not methyl groups) are rotationally optimized and His protonation is chosen within each local hydrogen-bond network, including interactions with the first shell of explicit waters.

A common problem is that the side-chain ends of Asn, Gln and His are easily fitted 180° backwards, since the electron density alone cannot usually distinguish the correct choice of orientation. *REDUCE* can automatically diagnose and correct these types of systematic errors by considering all-atom steric overlaps as well as hydrogen bonding within each local network. Automatic correction of Asn/Gln/His flips is the default option in *MolProbity* during addition of H atoms. *MolProbity* presents each potential flip correction to the user in kinemage view so they have the option of inspecting the before-and-after effects of each flip and approving (or rejecting) each correction. Fig. 21.6.2.1 shows an example of a simple Gln flip that is unquestionably correct but that could not have been decided on the basis of hydrogen bonding alone. Other examples can be much more complex, with rotatable OH positions, large hydrogen-bond networks and multiple competing interactions evaluated exhaustively.



(*a*) (*b*)

Figure 21.6.1.2
Two multi-criterion validation kinemages illustrating the successful outcome of an overall process of *MolProbity* diagnosis and structure improvement. (*a*) The original 1lpl Cap-Gly structure (Li *et al.*, 2002) shows three major clusters of clash, rotamer and Ramachandran problems plus a few isolated outliers. (*b*) The corrected 1tov structure (Arendall *et al.*, 2005) has essentially no outliers, a 4% lower $R_{\rm free}$, a bound sulfate and an additional turn of helix at the N-terminus.

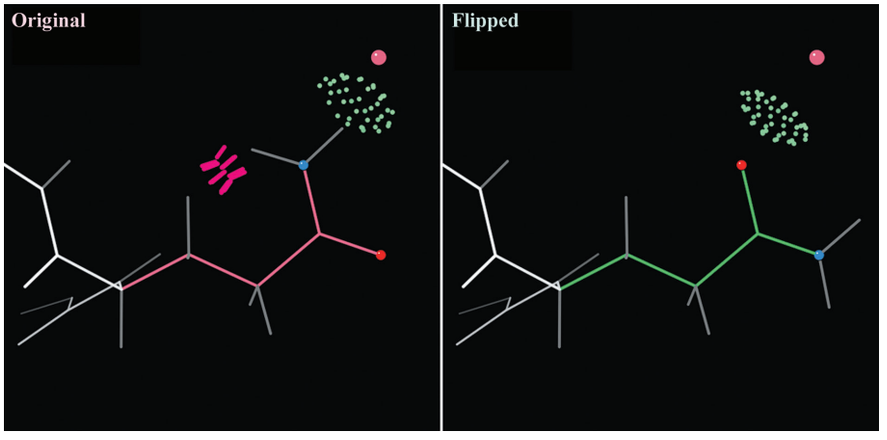


Figure 21.6.2.1
The simple 'flip' correction of a Gln side-chain amide in the 2dq4 threonine 3-dehydrogenase structure (R. Omi, T. Yao, M. Goto, I. Miyahara & K. Hirotsu, unpublished work), a better-than-average 2.5 Å resolution structure. Both orientations make a hydrogen bond to the crystallographic water, but the original has a serious internal clash of the NH_2 group with its own C^{β} hydrogen.

Users can also choose to add H atoms without Asn/Gln/His flips, which is useful in evaluating the atomic coordinates as they were deposited, but which rejects the easiest and most robustly correct improvement that can be made in a crystallographic model (Word, Lovell, Richardson & Richardson, 1999; Higman *et al.*, 2004). If flips are performed, the user needs to download and use the corrected PDB file (either with or without the H atoms) in order to benefit.

21.6.2.2. All-atom contact analysis

Once H atoms have been added to (or detected in) a structure, then the complete 'Analyze all-atom contacts and geometry' option is enabled. A main feature of this option is the all-atom contact analysis, which is performed by the program *PROBE* (Word, Lovell, LaBean *et al.*, 1999). *PROBE* operates by, in effect, rolling a 0.5 Å diameter ball around the van der Waals surfaces of atoms to measure the amount of overlap between pairs of nonbonded atoms. When non-donor–acceptor atoms overlap by more than 0.4 Å, *PROBE* denotes the contact as a serious clash, which is included in the reported clashscore and is shown in kinemage format as a cluster of hot-pink spikes in the overlap region (Fig. 21.6.1.1). Such large overlaps cannot occur in the actual molecule, but mean that at least one of the two atoms is modelled incorrectly. *MolProbity* allows users to select any combination of clashes, hydrogen bonds and van der Waals contacts to calculate and display on the structure. By default, all three are enabled for structures that are not excessively large; for large structures, van der Waals contacts are deselected.

The 'clashscore' is the number of serious clashes per 1000 atoms. It is reported in the *MolProbity* summary (top of Fig. 21.6.2.2), with a red/yellow/green colour coding for absolute quality. The structure's percentile rank for clashscore value within the relevant resolution range is also given. In the detailed sortable 'multi-chart' (an extract is shown below the summary in Fig. 21.6.2.2), the worst clash ≥0.4 Å is listed for each residue and highlighted in pink.

21.6.2.3. Torsion-angle combinations: updated Ramachandran and rotamer analyses

Also included in the 'Analyze all-atom contacts and geometry' option is an evaluation of where residues fall in the multi-dimensional distributions of Ramachandran backbone φ, ψ angles and side-chain rotamer χ angles. The reference distributions are currently from 100 000 residues in 500 files, quality-filtered at both the file and the residue level. The Ramachandran plots are separated for Gly, Pro and pre-Pro residue types; the general plot has only one in 2000 residues outside the 'allowed' contour, which is the same probability as a 3.5σ outlier in a normal distri-

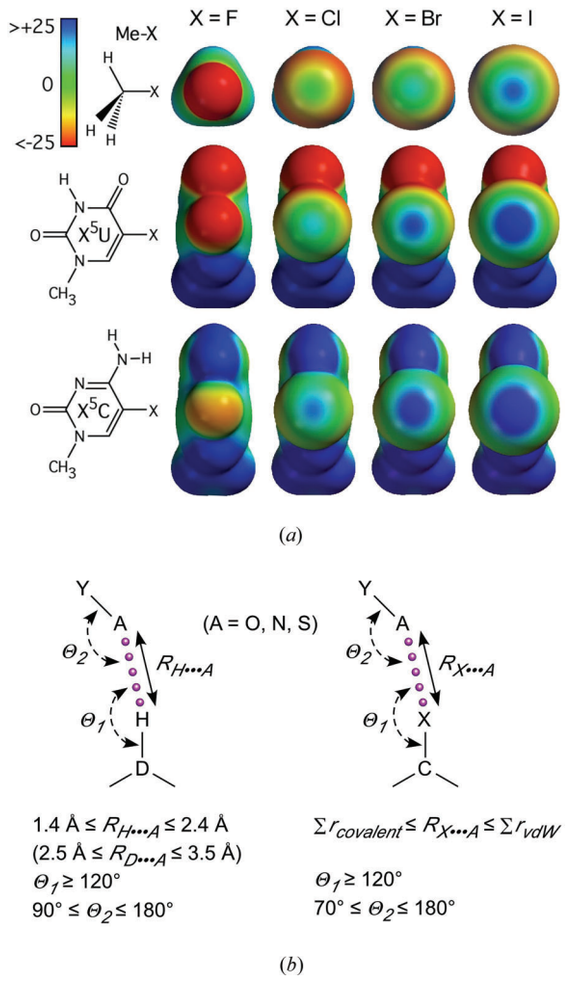


Figure 23.6.3.2
Properties of halogen bonds. (*a*) *Ab initio* calculations on model compounds (halomethane top, 5-halouracil middle, 5-halocytosine bottom) show an increase in polarizability of molecular halogens (*X*), with I > Br > Cl > F [adapted from Auffinger *et al.* (2004)]. The molecular surfaces are shown looking down the C–*X* bond, with electropositive and electronegative potentials shown in blue and red, respectively. (*b*) Comparison of the geometries of hydrogen bonds (left) and halogen bonds (right) [adapted from Voth & Ho (2007)]. In each case, the weak interactions are characterized by the distance between the donor (*D*) and acceptor (*A*) atoms ($R_{H\cdots A}$ and $R_{X\cdots A}$), which is longer than the sum of the covalent radii ($r_{\rm covalent}$) but shorter than the sum of the van der Waals radii ($r_{\rm vdW}$), and by the defined angles for the approach of the acceptor towards the donor (θ_1) and the donor towards the acceptor (θ_2).

The type of molecular interaction described by Hassel involves halogens serving as Lewis acids rather than as Lewis bases, and has seen a resurgence of interest in materials science. These interactions, initially called 'charge-transfer bonds' (Hassel, 1972), could, for example, draw the Br atoms of the Br_2 molecule to within 2.7 Å of the O atoms in dioxane (~0.5 Å shorter than $\sum r_{\rm vdW}$ of the interacting atoms). This would initially appear to be contrary to the expectation that electron-rich atoms should repel each other, but, as we shall see, halogens are not uniform spheres of electron density. In halogens larger than fluorine, the electron distributions are polarized such that an electropositive crown develops along the covalent (σ) C—*X* bond (Fig. 23.6.3.2*a*), and it is this 'σ hole' (Politzer, Murray & Lane, 2007) that accommodates the stabilizing electrostatic interactions between a molecular halogen and an electron-rich Lewis base (Fig. 23.6.3.2*b*). Such interactions are now referred to as halogen bonds (or *X* bonds) to acknowledge their similarities with the better known hydrogen bond (Metrangolo *et al.*, 2005; Metrangolo & Resnati, 2008). We can thus define the *X*-bond acceptor (by analogy with hydrogen-bond acceptors) as any electron-rich Lewis base, and the *X*-bond donor as the electropositive crown of the polarized halogen. Although they share similar acceptors (Metrangolo & Resnati, 2001) and, as we will discuss in greater detail, similar energies of interaction, *X* bonds are more directional than hydrogen bonds (Lommerse *et al.*, 1996).

The 'strength' of the *X* bond is expected to be dependent on the polarizability of the halogen donor, with I > Br > Cl in both polarizability and *X*-bonding potential (F atoms are not generally polarizable and therefore not typically considered *X*-bond donors) (Fig. 23.6.3.2). This is consistent with the understanding that the *X* bond is primarily an electrostatic interaction, although dispersion and charge transfer do contribute to some extent (Metrangolo *et al.*, 2005). Today, *X* bonds are exploited in chemistry to control molecular crystallization and supramolecular assembly (Metrangolo *et al.*, 2005; Metrangolo & Resnati, 2008; Politzer, Murray & Concha, 2007).

With this long history and a rekindled interest in applying *X* bonds for the design and synthesis of new materials (Fourmigué, 2009), it is surprising that this interaction has been largely overlooked in structural biology, but this could be attributed, until recently, to the very limited number of biological structures that included Group VII atoms. In 2004, we surveyed the occurrence of *X* bonds in the PDB and found 113 unique *X*-bond interactions in 66 protein–ligand complexes and six nucleic acid structures (Auffinger *et al.*, 2004). The results were dominated by *X* bonds to O atoms, with carbonyl O atoms of the peptide backbone (Figs. 23.6.1.1*a*,*b*) in protein–ligand complexes representing ~70% of these interactions. A survey of the 2008 PDB entries by Lu *et al.* (2009) found 258 *X* bonds to O-atom acceptors in the structures of protein–ligand complexes alone. Clearly, *X*-bonding interactions are important for ligand binding to proteins and this list of *X*-bonding interactions is expected to grow with the PDB, while becoming more prominent in the literature as the significance of this interaction becomes better recognized. A short Br···O *X* bond is estimated to contribute to a more than 1000-fold increase in selectivity (equivalent to ~4 kcal mol^{-1} in energy; 1 kcal mol^{-1} = 4.184 kJ mol^{-1}) for the binding of an inhibitor by aldose reductase *versus* the related aldehyde reductase enzyme (Muzet *et al.*, 2003). In addition to oxygen, other electron-rich Lewis bases are also seen as *X*-bond acceptors, including the N atoms of various basic amino-acid side chains (Voth & Ho, 2007) and the S atoms of Cys or Met residues (Liu *et al.*, 2009). Much of the remainder of this discussion will be on the dominant interactions with O-atom acceptors of various types.

The geometry of *X* bonds provides an understanding of what is involved in the formation and stabilization of *X* bonds in biological systems. By analogy with the structure of *X* bonds between small organic molecules (Cody & Murray-Rust, 1984; Lommerse *et al.*, 1996; Ouvrard *et al.*, 2003), the geometry of biological *X* bonds is characterized by the angles of approach of the acceptor to the halogen donor (θ_1 relative to the C—*X* bond; Fig. 23.6.3.2*b*) and the halogen towards the base acceptor (θ_2; Fig. 23.6.3.2*b*) (Auffinger *et al.*, 2004). The θ_1 angle is essentially linear, consistent with this being an electrostatic interaction of the base acceptor, with the electropositive crown, or σ hole, being aligned along the C—*X* bond, as seen previously in organic complexes. The θ_2 angle is ~110–120°, suggesting that the *X* bonds involve the non-bonding electrons of the acceptor. However, an extended analysis showed that most *X* bonds in protein–ligand complexes are aligned perpendicular to the