INTERNATIONAL TABLES FOR CRYSTALLOGRAPHY

Volume F

CRYSTALLOGRAPHY OF BIOLOGICAL MACROMOLECULES

Edited by

MICHAEL G. ROSSMANN AND EDDY ARNOLD

Advisors and Advisory Board

Contributing authors

E. E. Abola: The Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA. [24.1]

P. D. Adams: The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [18.2, 25.2.3]

F. H. Allen: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [22.4, 24.3]

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, 1.4.1, 13.4, 25.1]

E. N. Baker: School of Biological Sciences, University of Auckland, Private Bag 92-109, Auckland, New Zealand. [22.2]

T. S. Baker: Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907-1392, USA. [19.6]

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

J. Berendzen: Biophysics Group, Mail Stop D454, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [14.2.2]

H. M. Berman: The Nucleic Acid Database Project, Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8077, USA. [21.2, 24.2, 24.5]

T. N. Bhat: National Institute of Standards and Technology, Biotechnology Division, 100 Bureau Drive, Gaithersburg, MD 20899, USA. [24.5]

C. C. F. Blake: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.§ [26.1]

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]

P. E. Bourne: Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537, USA. [24.5]

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

A. T. Brunger: Howard Hughes Medical Institute, and Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences, and Stanford Synchrotron Radiation Laboratory (SSRL), Stanford University, 1201 Welch Road, MSLS P210, Stanford, CA 94305, USA. [18.2, 25.2.3]

A. Burgess Hickman: Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0560, USA. [4.3]

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. [22.1.2]

† 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.

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

J. C. Cole: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [22.4]

M. L. Connolly: 1259 El Camino Real #184, Menlo Park, CA 94025, USA. [22.1.2]

K. D. Cowtan: Department of Chemistry, University of York, York YO1 5DD, England. [15.1, 25.2.2]

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

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: National Cancer Institute, Brookhaven National Laboratory, Building 725A-X9, Upton, NY 11973, USA. [9.1, 18.4]

D. R. Davies: Laboratory of Molecular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0560, USA. [4.3]

W. L. DeLano: Graduate Group in Biophysics, Box 0448, University of California, San Francisco, CA 94143, USA. [25.2.3]

R. E. Dickerson: Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095-1570, USA. [23.3]

J. Ding: Biomolecular Crystallography Laboratory, CABM & Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854-5638, USA, and Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Yue-Yang Road, Shanghai 200 031, People's Republic of China. [25.1]

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

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

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

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]

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]

Z. Feng: The Nucleic Acid Database Project, Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8077, USA. [24.2, 24.5]

R. H. Fenn: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.‡‡ [26.1]

W. Furey: Biocrystallography Laboratory, VA Medical Center, PO Box 12055, University Drive C, Pittsburgh, PA 15240, USA, and Department of Pharmacology, University of Pittsburgh School of Medicine, 1340 BSTWR, Pittsburgh, PA 15261, USA. [25.2.1]

M. Gerstein: Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA. [22.1.1]

R. Giegé: Unité Propre de Recherche du CNRS, Institut de Biologie Moléculaire et Cellulaire, 15 rue René Descartes, F-67084 Strasbourg CEDEX, France. [4.1]

†† Present address: 105 Moss Lane, Alderley Edge, Cheshire SK9 7HW, England.

‡‡ Present address: 2 Second Avenue, Denvilles, Havant, Hampshire PO9 2QP, England.

G. L. Gilliland: 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. [24.4, 24.5]

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

P. Gros: Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. [25.2.3]

R. W. Grosse-Kunstleve: The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [25.2.3]

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

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

H. A. Hauptman: Hauptman–Woodward Medical Research Institute, Inc., 73 High Street, Buffalo, NY 14203-1196, USA. [16.1]

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. [14.2.1]

A. E. Hodel: Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA. [23.2]

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. [23.1.2]

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

V. J. Hoy: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [24.3]

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]

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

J.-S. Jiang: Biology Department, Bldg 463, Brookhaven National Laboratory, Upton, NY 11973-5000, USA. [24.1, 25.2.3]

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.‡ [26.1]

T. A. Jones: Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, Box 596, SE-751 24 Uppsala, Sweden. [17.1]

W. Kabsch: Max-Planck-Institut für medizinische Forschung, Abteilung Biophysik, Jahnstrasse 29, 69120 Heidelberg, Germany. [11.3, 25.2.9]

M. Kjeldgaard: Institute of Molecular and Structural Biology, University of Aarhus, Gustav Wieds Vej 10c, DK-8000 Aarhus C, Denmark. [17.1]

G. J. Kleywegt: Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, Box 596, SE-751 24 Uppsala, Sweden. [17.1, 21.1]

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.§ [26.1]

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

P. J. Kraulis: Stockholm Bioinformatics Center, Department of Biochemistry, Stockholm University, SE-106 91 Stockholm, Sweden. [25.2.7]

J. E. Ladner: 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. [24.4]

V. S. Lamzin: European Molecular Biology Laboratory (EMBL), Hamburg Outstation, c/o DESY, Notkestr. 85, 22603 Hamburg, Germany. [25.2.5]

R. A. Laskowski: Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, England. [25.2.6]

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

D. Lin: Biology Department, Bldg 463, Brookhaven National Laboratory, Upton, NY 11973-5000, USA. [24.1]

M. W. MacArthur: Biochemistry and Molecular Biology Department, University College London, Gower Street, London WC1E 6BT, England. [25.2.6]

A. McPherson: Department of Molecular Biology & Biochemistry, University of California at Irvine, Irvine, CA 92717, USA. [4.1]

P. Main: Department of Physics, University of York, York YO1 5DD, England. [15.1, 25.2.2]

G. A. Mair: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.§ [26.1]

N. O. Manning: Biology Department, Bldg 463, Brookhaven National Laboratory, Upton, NY 11973-5000, 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. [23.4]

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., 73 High Street, Buffalo, NY 14203-1196, 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]

G. N. Murshudov: Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, England, and CLRC, Daresbury Laboratory, Daresbury, Warrington, WA4 4AD, England. [18.4]

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

A. C. T. North: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.¶ [26.1]

J. W. H. Oldham.† [26.1]

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. [23.1.1]

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. [25.2.3]

A. Perrakis: European Molecular Biology Laboratory (EMBL), Grenoble Outstation, c/o ILL, Avenue des Martyrs, BP 156, 38042 Grenoble CEDEX 9, France. [25.2.5]

D. C. Phillips.† [26.1]

R. J. Poljak: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.†† [26.1]

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]

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

J. Prilusky: Bioinformatics Unit, Weizmann Institute of Science, Rehovot 76100, Israel. [24.1]

‡ Present address: Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford OX1 3QU, England.

§ Present address unknown.

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

† Deceased.

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

F. A. Quiocho: Howard Hughes Medical Institute and Department of Biochemistry, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [23.2]

R. J. Read: Department of Haematology, University of Cambridge, Wellcome Trust Centre for Molecular Mechanisms in Disease, CIMR, Wellcome Trust/ MRC Building, Hills Road, Cambridge CB2 2XY, England. [15.2, 25.2.3]

L. M. Rice: Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [18.2, 25.2.3]

F. M. Richards: Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA. [22.1.1]

D. C. Richardson: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710-3711, USA. [25.2.8]

J. S. Richardson: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710-3711, USA. [25.2.8]

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. [23.4]

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

M. G. Rossmann: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA. [1.1, 1.2, 1.4.2, 11.1, 11.5, 13.4]

C. Sander: MIT Center for Genome Research, One Kendall Square, Cambridge, MA 02139, USA. [23.1.2]

V. R. Sarma: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.‡ [26.1]

B. Schneider: The Nucleic Acid Database Project, Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8077, USA. [24.2]

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

K. A. Sharp: E. R. Johnson Research Foundation, Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA 19104-6059, USA. [22.3]

G. M. Sheldrick: Lehrstuhl für Strukturchemie, Universität Göttingen, Tammannstrasse 4, D-37077 Göttingen, Germany. [16.1, 25.2.10]

I. N. Shindyalov: San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537, USA. [24.5]

T. Simonson: Laboratoire de Biologie Structurale (CNRS), IGBMC, 1 rue Laurent Fries, 67404 Illkirch (CU de Strasbourg), France. [25.2.3]

J. L. Smith: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA. [14.2.1]

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]

J. L. Sussman: Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel. [24.1]

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

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, 25.2.4]

T. C. Terwilliger: Bioscience Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [14.2.2]

J. M. 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. [23.1.1, 25.2.6]

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. [25.2.4]

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. [24.4]

I. Usón: Institut für Anorganisch Chemie, Universität Göttingen, Tammannstrasse 4, D-37077 Göttingen, Germany. [16.1]

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

M. L. Verdonk: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [22.4]

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.† [26.1]

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., 73 High Street, Buffalo, NY 14203-1196, USA. [16.1]

H. Weissig: San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537, USA. [24.5]

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

J. Westbrook: The Nucleic Acid Database Project, Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8077, USA. [24.2, 24.5]

K. S. Wilson: Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, England. [9.1, 18.4. 25.2.5]

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]

T. O. Yeates: UCLA–DOE Laboratory of Structural Biology and Molecular Medicine, Department of Chemistry & Biochemistry and Molecular Biology Institute, UCLA, Los Angeles, CA 90095-1569, USA. [21.3]

C. Zardecki: The Nucleic Acid Database Project, Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8077, 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, 25.2.2]

J.-Y. Zou: Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, Box 596, SE-751 24 Uppsala, Sweden. [17.1]

† Deceased.

Contents

SAMPLE PAGES

1. INTRODUCTION

1.1. Overview

BY E. ARNOLD AND M. G. ROSSMANN

As the first *International Tables* volume devoted to the crystallography of large biological molecules, Volume F is intended to complement 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 considerably greater depth in other volumes of *International Tables*. Much of the information in the latter portion of this volume is more specifically related to macromolecular structure. This chapter is intended to serve as a basic guide to the contents of this book and to how the information herein relates to material in the other *International Tables* volumes.

Chapter 1.2 presents a brief history of the field of macromolecular crystallography. This is followed by an article describing many of the connections of crystallography with the field of medicine and providing an exciting look into the future possibilities of structure-based design of drugs, vaccines and other agents. Chapter 1.4 provides some personal perspectives on the future of science and crystallography, and is followed by a complementary response suggesting how crystallography could play a central role in unifying diverse scientific fields in the future.

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. All of the existing *International Tables* volumes are central references for basic crystallography.

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. This is complemented by Chapter 4.3, which describes molecular-engineering approaches for enhancing the likelihood of 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 to play a central role in coaxing more macromolecules to form crystals suitable for structure determination in the future. 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 provides a brief introduction to the theory and practice of generating X-rays and neutrons for diffraction experiments. Chapter 6.1 describes the basic 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 excellent articles in other volumes of *International Tables* go into more depth in these areas and the reader is referred in particular to Volume C, Chapter 4.2. Chapter 6.2 describes the generation and definition of neutron beams; related articles in other *International Tables* volumes include those in Volume C, Chapter 4.4.

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 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 also 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.

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. 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 successfully to cryogenic temperatures. Thus, data collection using ambient conditions, or moderate cooling (from approximately -40 °C to a few degrees below ambient temperature), are 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, integration, space-group assignment, scaling and post refinement.

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 very 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. 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 been long 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 Chapter 14.2.

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.

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.

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 is likely. The availability of computer graphics has also contributed greatly to the magnificent illustration of crystal structures, one of the factors that has thrust structural biology into many prominent roles in modern life and chemical sciences. 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.

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. 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). The combination of electron microscopy and crystallography is helping to bridge molecular structure and multi-molecular ultrastructure in living cells. 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.

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.

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 Chapter 22.1. Useful generalizations relating surface areas buried at macromolecular interfaces and energies of association have emerged. Chapter 22.2 surveys the occurrence of hydrogen bonds in biological macromolecules. Electrostatic interactions in proteins are described in Chapter 22.3. 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 22.4.

Part 23 surveys current knowledge of protein and nucleic acid structures. Proliferation of structural data has created problems for classification schemes, which have been forced to co-evolve with new structural knowledge. Methods of protein structural classification are described in Chapter 23.1. Systematic aspects of ligand binding to macromolecules are discussed in Chapter 23.2. A survey of nucleic acid structure, geometry and classification schemes is presented in Chapter 23.3. Solvent structure in macromolecular crystals is reviewed in Chapter 23.4.

With the proliferation of macromolecular structures, it has been necessary to have databases as international resources for rapid access to, and archival of, primary structural data. The functioning of the former Brookhaven Protein Data Bank (PDB), which for almost thirty years was the depository for protein crystal (and later NMR) structures, is summarized in Chapter 24.1. Chapter 24.5 describes the organization and features of the new PDB, run by the Research Collaboratory for Structural Bioinformatics, which superseded the Brookhaven PDB in 1999. The PDB permits rapid 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 described in Chapter 24.2. The Cambridge Structural Database (CSD), which is the central database for small-molecule structures, is described in Chapter 24.3. The Biological Macromolecule Crystallization Database (BMCD), a repository for macromolecular crystallization data, is described in Chapter 24.4.

Part 25 summarizes computer programs and packages in common use in macromolecular structure determination and analysis. Owing to constant changes in this area, the information in this chapter is expected to be more volatile than that in the remainder of the volume. Chapter 25.1 presents a survey of some of the most popular programs, with a brief description and references for further information. Specific programs and program systems summarized include *PHASES* (Section 25.2.1); *DM*/*DMMULTI* (Section 25.2.2); the *Crystallography & NMR System* or *CNS* (Section 25.2.3); the *TNT* refinement package (Section 25.2.4); *ARP* and *wARP* for automated model construction and refinement (Section 25.2.5); *PROCHECK* (Section 25.2.6); *MolScript* (Section 25.2.7); *MAGE*, *PROBE* and kinemages (Section 25.2.8); *XDS* (Section 25.2.9); and *SHELX* (Section 25.2.10).

Chapter 26.1 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.

well as great gaps in our structural knowledge of proteins from humans and human pathogens.

1.3.3. Crystallography and genetic diseases

Presently, an immense number of genetic diseases have been characterized at the genetic level and archived in OMIM [On-line Mendelian Inheritance in Man. Center for Medical Genetics, Johns Hopkins University (Baltimore, MD) and the National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 1999. URL: http://www.ncbi.nlm.nih.gov/omim/], with many more discoveries to occur in the next decades. Biomolecular crystallography has been very successful in explaining the cause of numerous genetic diseases at the atomic level. The stories of sickle cell anaemia, thalassemias and other deficiencies of haemoglobin set the stage (Dickerson & Geis, 1983), followed by numerous other examples (Table 1.3.3.1). Given the frequent

Table 1.3.3.1. *Crystal structures and genetic diseases*

Crystal structure	Disease	Reference
Acidic fibroblast growth factor receptor	Familial Pfeiffer syndrome	[1]
Alpha-1-antitrypsin	Alpha-1-antitrypsin deficiency	[2]
Antithrombin III	Hereditary thrombophilia	[3], [4]
Arylsulfatase A	Leukodystrophy	[5]
Aspartylglucosaminidase	Aspartylglusominuria	[6]
Beta-glucuronidase	Sly syndrome	[7]
Branched-chain alpha-keto acid dehydrogenase	Maple syrup urine syndrome, type Ia	[39]
Carbonic anhydrase II	Guibaud–Vainsel syndrome, Marble brain disease	[8]
p53	Cancer	[9], [10]
Ceruloplasmin	Hypoceruloplasminemia	[11]
Complement C3	C3 complement component 3 deficiency	[12]
Cystatin B	Progressive myoclonus epilepsy	[13]
Factor VII	Factor VII deficiency	[14]
Factor VIII	Factor VIII deficiency	[40]
Factor X	Factor X deficiency (Stuart–Prower factor deficiency)	[15]
Factor XIII	Factor XIII deficiency	[16]
Fructose-1,6-bisphosphate aldolase	Fructose intolerance (fructosemia)	[41]
Gelsolin	Amyloidosis V	[17]
Growth hormone	Growth hormone deficiency	[18]
Haemochromatosis protein HFE	Hereditary haemochromatosis	[19]
Haemoglobin	Beta-thalassemia, sickle-cell anaemia	[20]
Tyrosine hydroxylase	Hereditary Parkinsonism	[21]
Hypoxanthine–guanine phosphoribosyltransferase	Lesch–Nyhan syndrome	[22]
Insulin	Hyperproinsulinemia, diabetes	[42]
Isovaleryl–coenzyme A dehydrogenase	Isovaleric acid CoA dehydrogenase deficiency	[23]
Lysosomal protective protein	Galactosialidosis	[24]
Ornithine aminotransferase	Ornithine aminotransferase deficiency	[25]
Ornithine transcarbamoylase	Ornithine transcarbamoylase deficiency	[43]
p16INK4a tumour suppressor	Cancer	[26]
Phenylalanine hydroxylase	Phenylketonuria	[27]
Plasminogen	Plasminogen deficiency	[28], [29], [30]
Protein C	Protein C deficiency	[31]
Purine nucleotide phosphorylase	Purine nucleotide phosphorylase deficiency	[32]
Serum albumin	Dysalbuminemic hyperthyroxinemia	[33]
Superoxide dismutase (Cu, Zn-dependent)	Familial amyotrophical lateral sclerosis	[34]
Thrombin	Hypoprothrombinemia, dysprothrombinemia	[35]
Transthyretrin	Amyloidosis I	[36]
Triosephosphate isomerase	Triosephosphate isomerase deficiency	[37]
Trypsinogen	Hereditary pancreatitis	[38]

References: [1] Blaber *et al.* (1996); [2] Loebermann *et al.* (1984); [3] Carrell *et al.* (1994); [4] Schreuder *et al.* (1994); [5] Lukatela *et al.* (1998); [6] Oinonen *et al.* (1995); [7] Jain *et al.* (1996); [8] Liljas *et al.* (1972); [9] Cho *et al.* (1994); [10] Gorina & Pavletich (1996); [11] Zaitseva *et al.* (1996); [12] Nagar *et al.* (1998); [13] Stubbs *et al.* (1990); [14] Banner *et al.* (1996); [15] Padmanabhan *et al.* (1993); [16] Yee *et al.* (1994); [17] McLaughlin *et al.* (1993); [18] DeVos *et al.* (1992); [19] Lebron *et al.* (1998); [20] Harrington *et al.* (1997); [21] Goodwill *et al.* (1997); [22] Eads *et al.* (1994); [23] Tiffany *et al.* (1997); [24] Rudenko *et al.* (1995); [25] Shah *et al.* (1997); [26] Russo *et al.* (1998); [27] Erlandsen *et al.* (1997); [28] Mulichak *et al.* (1991); [29] Mathews *et al.* (1996); [30] Chang, Mochalkin *et al.* (1998); [31] Mather *et al.* (1996); [32] Ealick *et al.* (1990); [33] He & Carter (1992); [34] Parge *et al.* (1992); [35] Bode *et al.* (1989); [36] Blake *et al.* (1978); [37] Mande *et al.* (1994); [38] Gaboriaud *et al.* (1996); [39] Ævarsson *et al.* (2000); [40] Pratt *et al.* (1999); [41] Gamblin *et al.* (1990); [42] Bentley *et al.* (1976); [43] Shi *et al.* (1998).

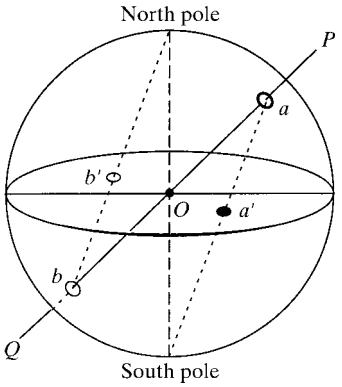


Fig. 2.1.3.1. How to construct a stereographic projection. Imagine a sphere around the crystal with O as the centre. O is also the origin of the coordinate system of the crystal. Symmetry elements of the point groups pass through O. Line OP is normal to a crystal plane. It cuts through the sphere at point a. This point a is projected onto the horizontal plane through O in the following way: a vertical dashed line is drawn through O normal to the projection plane and connecting a north and a south pole. Point a is connected to the pole on the other side of the projection plane, the south pole, and is projected onto the horizontal plane at a'. For a normal OQ intersecting the lower part of the sphere, the point of intersection b is connected to the north pole and projected at b'. For the symmetry elements, their points of intersection with the sphere are projected onto the horizontal plane.

morphic space groups. (Enantiomorphic means the structure is not superimposable on its mirror image.) Apparently, some of these space groups supply more favourable packing conditions for proteins than others. The most favoured space group is $P2_12_12_1$ (Table 2.1.2.1). A consequence of symmetry is that multiple copies of particles exist in the unit cell. For instance, in space group $P2_1$ (space group No. 4), one can always expect two exactly identical entities in the unit cell, and one half of the unit cell uniquely represents the structure. This unique part of the structure is called the asymmetric unit. Of course, the asymmetric unit does not necessarily contain one protein molecule. Sometimes the unit cell contains fewer molecules than anticipated from the number of asymmetric units. This happens when the molecules occupy a position on a crystallographic axis. This is called a special position.

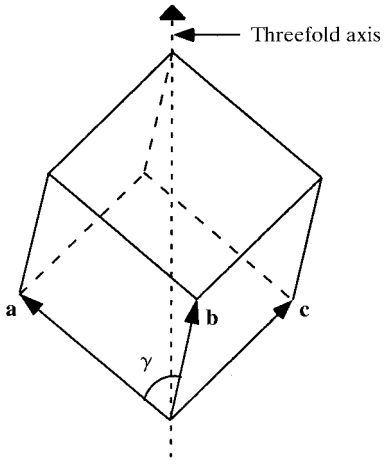


Fig. 2.1.3.2. A rhombohedral unit cell.

In this situation, the molecule itself obeys the axial symmetry. Otherwise, the molecules in an asymmetric unit are on general positions. There may also be two, three or more equal or nearly equal molecules in the asymmetric unit related by noncrystallographic symmetry.

2.1.3. Point groups and crystal systems

If symmetry can be recognised in the external shape of a body, like a crystal or a virus molecule, corresponding symmetry elements have no translations, because internal translations (if they exist) do not show up in macroscopic properties. Moreover, they pass through one point, and this point is not affected by the symmetry operations (point-group symmetry). For idealized crystal shapes, the symmetry axes are limited to one-, two-, three-, four- and sixfold rotation axes because of the space-filling requirement for crystals. With the addition of mirror planes and inversion centres, there are a total of 32 possible crystallographic point groups.

Not all combinations of axes are allowed. For instance, a combination of two twofold axes at an arbitrary angle with respect to each other would multiply to an infinite number of twofold axes. A twofold axis can only be combined with another twofold axis at 90°. A third twofold axis is then automatically produced perpendicular to the first two (point group 222). In the same way, a threefold axis can only be combined with three twofold axes perpendicular to the threefold axis (point group 32).

For crystals of biological macromolecules, point groups with mirrors or inversion centres are not allowed, because these molecules are chiral. This restricts the number of crystallographic point groups for biological macromolecules to 11; these are the enantiomorphic point groups and are presented in Table 2.1.3.1.

Although the crystals of asymmetric molecules can only belong to one of the 11 enantiomorphic point groups, it is nevertheless important to be aware of the other point groups, especially the 11 centrosymmetric ones (Table 2.1.3.2). This is because if anomalous scattering can be neglected, the X-ray diffraction pattern of a crystal is always centrosymmetric, even if the crystal itself is asymmetric (see Sections 2.1.7 and 2.1.8).

The protein capsids of spherical virus molecules have their subunits packed in a sphere with icosahedral symmetry (532). This is the symmetry of a noncrystallographic point group (Table 2.1.3.3). A fivefold axis is allowed because translation symmetry does not apply to a virus molecule. Application of the 532 symmetry leads to 60 identical subunits in the sphere. This is the simplest type of spherical virus (triangulation number $T = 1$). Larger numbers of subunits can also be incorporated in this icosahedral surface lattice, but then the subunits lie in quasi-equivalent environments and T assumes values of 3, 4 or 7. For instance, for $T = 3$ particles there are 180 identical subunits in quasi-identical environments.

On the basis of their symmetry, the point groups are subdivided into crystal systems as follows. For each of the point groups, a set of axes can be chosen displaying the external symmetry of the crystal as clearly as possible, and, in this way, the seven crystal systems of Table 2.1.3.4 are obtained. If no other symmetry is present apart from translational symmetry, the crystal belongs to the triclinic system. With one twofold axis or screw axis, it is monoclinic. The convention in the monoclinic system is to choose the b axis along the twofold axis. The orthorhombic system has three mutually perpendicular twofold (screw) axes. Another convention is that in tetragonal, trigonal and hexagonal crystals, the axis of highest symmetry is labelled c. These conventions can deviate from the guide rules for unit-cell choice given in Section 2.1.1.

The seven crystal systems are based on the point-group symmetry. Except for the triclinic unit cell, all other cells can

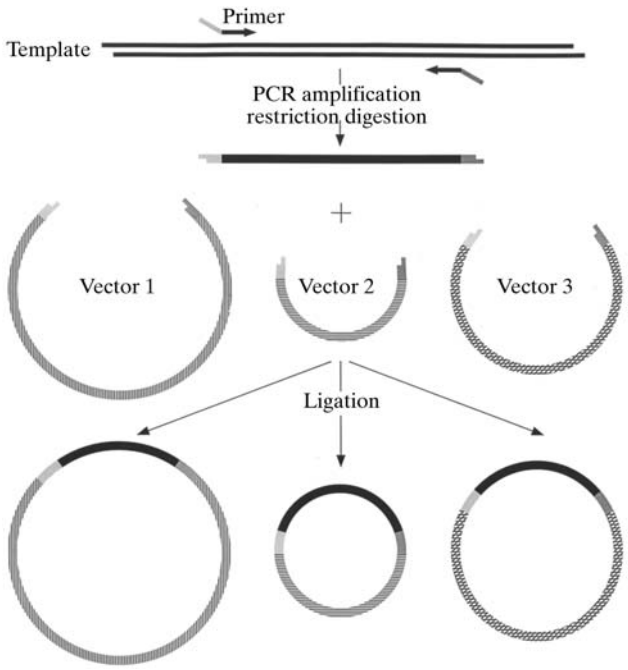


Fig. 3.1.3.1. Creating an expression construct. PCR can be used to amplify the coding region of interest, providing that a suitable template is available. PCR primers should be designed to contain one or more restriction sites that can be conveniently used to subclone the fragment into the desired expression vector. It is often possible to choose vectors and primers such that a single PCR product can be ligated to multiple vectors. The ability to test several expression systems simultaneously is advantageous, since it is impossible to predict which vector/host system will give the most successful expression of a specific protein.

3.1.3.3. *Addition of tags or domains*

In some cases it is useful to add a small peptide tag or a larger protein to either the amino or carboxyl terminus of the protein of interest (Nilsson *et al.*, 1992; LaVallie & McCoy, 1995). As will be discussed in more detail below, such fused elements can be used for affinity chromatography and can greatly simplify the purification of the recombinant protein. In addition to aiding purification, some protein domains used as tags, such as the maltose-binding protein, thioridazine, and protein A, can also act as molecular chaperones to aid in the proper folding of the recombinant protein (LaVallie *et al.*, 1993; Samuelsson *et al.*, 1994; Wilkinson *et al.*, 1995; Richarme & Caldas, 1997; Sachdev & Chirgwin, 1998). Tags range in size from several amino acids to tens of kilodaltons. Numerous tags [including hexahistidine (His_6), biotinylation peptides and streptavidin-binding peptides (Strep-tag), calmodulin-binding peptide (CBP), cellulose-binding domain (CBD), chitin-binding domain (CBD), glutathione S-transferase (GST), maltose-binding protein (MBP), protein A domains, ribonuclease A S-peptide (S-tag) and thioridazine (Trx)] have already been engineered into expression vectors that are commercially available. Additional systems are constantly being introduced. While these systems provide some advantages, there are also drawbacks, including expense, which can be considerable when both affinity purification and specific proteolytic removal of the tag are performed on a large scale.

If a sequence tag or a fusion protein is added to the protein of interest, one problem is solved but another is created, *i.e.* whether or not to try to remove the fused element. During the past year, there have been numerous reports of crystallization of proteins containing His-tags, but there are also unpublished anecdotes about cases where removal of the tag was necessary to obtain crystals. In a small number of cases, additional protein domains present in fusion proteins appear to have aided crystallization (see Chapter 4.3). Experiences with tags appear to be protein specific. There are a number of relevant issues, including the protein, the tag and the length and composition of the linker that joins the two. If the tag is to be removed, it is usually necessary to use a protease. To avoid unwanted cleavage of the desired protein, 'specific' proteases are usually used. When the expression system is designed, the tag or fused protein is separated from the desired protein by the recognition site for the protease. While this procedure sounds simple and straightforward, and has, in some cases, worked exactly as outlined here, there are a number of potential pitfalls. Proteases do not always behave exactly as advertised, and there can be unwanted cleavages in the desired product. Since protease cleavage efficiency can be quite sensitive to structure, it may be more difficult to cleave the fusion joint than might be expected. Unless cleavage is performed with an immobilized protease, additional purification is necessary to separate the protease from the desired protein product. A variation of the classic tag-removal procedure is provided by a system in which a fusion domain is linked to the protein of interest by a protein self-cleaving element called an intein (Chong *et al.*, 1996, 1997).

3.1.4. Expression systems

3.1.4.1. *E. coli*

If the desired protein does not have extensive post-translational modifications, it is usually appropriate to begin with an *E. coli* host–vector system (for an extensive review of expression in *E. coli*, see Makrides, 1996). Both plasmid-based and viral-based (M13, λ *etc.*) expression systems are available for *E. coli*. Although viral-based vector systems are quite useful for some purposes (expression cloning of cDNA strands, for example), in general, for expression of relatively large amounts of recombinant protein, they are not as convenient as plasmid-based expression systems. Although there are minor differences in the use of viral expression systems and plasmid-based systems, the rules that govern the design of the modified segment are the same and we will discuss only plasmid-based systems. We will first consider general issues related to design of the plasmid, then continue with a discussion of fermentation conditions, and finally address some of the problems commonly encountered and potential solutions.

Basically, a plasmid is a small circular piece of DNA. To be retained by *E. coli*, it must contain signals that allow it to be successfully replicated by the host. Most of the commonly used *E. coli* expression plasmids are present in the cell in multiple copies. Simply stated, in the selection of *E. coli* containing the plasmid, the plasmids carry selectable markers, which usually confer resistance to an antibiotic, typically ampicillin and/or kanamycin. Ampicillin resistance is conferred by the expression of a β-lactamase that is secreted from cells and breaks down the antibiotic. It has been found that, in typical liquid cultures, most of the ampicillin is degraded by the time cells reach turbidity (approximately 10^7 cells ml^{-1}), and cells not harbouring plasmids can overgrow the culture (Studier & Moffatt, 1986). For this reason, kanamycin resistance is being used as the selectable marker in many recently constructed expression plasmids.

There are literally dozens, if not hundreds, of expression plasmids available for *E. coli*, so a comprehensive discussion of the available plasmids is neither practical nor useful. Fortunately, this broad array of choices means that considerable effort has been expended in developing *E. coli* expression systems that are efficient and easy to use (for a concise review, see Unger, 1997). In most cases, it is possible to find expression and/or fermentation conditions that result in the production of a recombinant protein

3.1.5. Protein purification

3.1.5.1. *Conventional protein purification*

Those of us old enough to remember the task of purifying proteins from their natural sources, using conventional (as opposed to affinity) chromatography, where a 5000-fold purification was not unusual and the purifications routinely began with kilogram quantities (wet weight) of *E. coli* paste or calves' liver, are most grateful to those who developed efficient systems to express recombinant proteins. In most cases, it is possible to develop expression systems that limit the required purification to, at most, 20- to 50-fold, which vastly simplifies the purification procedure and concomitantly reduces the amount of starting material required to produce the 5–10 mg of pure protein needed to begin crystallization trials. This does not mean, however, that the process of purifying recombinant proteins is trivial. Fortunately, advances in chromatography media and instrumentation have improved both the speed and ease of protein purification. A wide variety of chromatography media (and prepacked columns) are commercially available, along with technical bulletins that provide detailed recommended protocols for their use. Purification systems (such as Pharmacia's FPLC and ÄKTA systems, PerSeptive Biosystems' BioCAD workstations and BioRad's BioLogic systems) include instruments for sample application, pumps for solvent delivery, columns, sample detection, fraction collection and information storage and output into a single integrated system, but such systems are relatively expensive. Several types of high capacity, high flow rate chromatography media and columns (for example, Pharmacia's HiTrap products and PerSeptive Biosystems' POROS Perfusion Chromatography products) have been developed and are marketed for use with these systems. However, the use of these media is not restricted to the integrated systems; they can be used effectively in conventional chromatography without the need for expensive instrumentation.

In designing a purification protocol, it is critically important that careful thought be given to the design of the protocol and to a proper ordering of the purification steps. In most cases, individual purification steps are worked out on a relatively small scale, and an overall purification scheme is developed based on an ordering of these independently developed steps. However, the experimentalist, in planning a purification scheme, should keep the amount of protein needed for the project firmly in mind. In general, crystallography takes a good deal more purified protein than conventional biochemical analyses. Scaling up a purification scheme is an art; however, it should be clear that purification steps that can be conveniently done in batch mode (precipitation steps) should be the earliest steps in a large-scale purification, chromatographic steps that involve the absorption and desorption of the protein from columns (ion-exchange, hydroxyapatite, hydrophobic interaction, dye-ligand and affinity chromatography) should be done as intermediate steps, and size exclusion, which requires the largest column volumes relative to the amount of protein to be purified, should generally be used only as the last step of purification. If reasonably good levels of expression can be achieved, most recombinant proteins can be purified using a relatively simple combination of the previously mentioned procedures (Fig. 3.1.5.1), requiring a limited number of column chromatography steps (generally two or three).

All protein purification steps are based on the fact that the biochemical properties of proteins differ: proteins are different sizes, have different surface charges and different hydrophobicity. With the exception of a small number of cases involving proteins that have unusual solubility characteristics, batch precipitation steps usually do not provide substantial increases in purity. However, precipitation is often used as the first step in a purification procedure, in part because it can be used to separate protein from

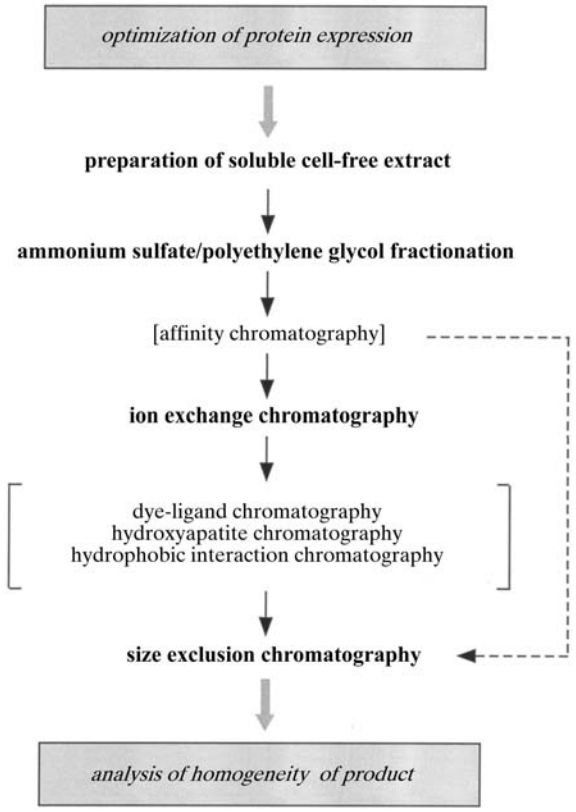


Fig. 3.1.5.1. Protein purification strategy. Purification of proteins expressed at reasonably high levels typically requires only a limited number of chromatographic steps. Additional chromatography columns (indicated in brackets) can be included as necessary. Affinity chromatography can allow efficient purification of fusion proteins or proteins with well defined ligand-binding domains.

nucleic acids. Nucleic acids are highly charged polyanions; the presence of nucleic acid in a protein extract can dramatically decrease the efficiency of column chromatography, for example by saturation of anion-exchange resins. If the desired protein binds to nucleic acids and the nucleic acids are not removed, ion-exchange chromatography can be compromised by the interactions of the protein and the nucleic acid and by the interactions of the nucleic acid and the column. The most commonly used precipitation reagents are ammonium sulfate and polyethylene glycols. With little effort, the defined range of these reagents needed to precipitate the protein of interest can be determined. However, if the precipitation range is broad, it may be only marginally less efficient simply to precipitate the majority of proteins by addition of ammonium sulfate to 85% saturation or 30% polyethylene glycol 6000. Precipitation can be a useful method for concentrating proteins at various steps during purification and for storing proteins that are unstable upon freezing or upon storage in solution.

Column chromatography steps in which the protein is absorbed onto the resin under one set of conditions and then eluted from the column under a different set of conditions can produce significant purification. Anion-exchange chromatography is usually a good starting point. Most proteins have acidic pIs, and conditions can often be found that allow binding of the protein to anion-exchange matrices. Elution of the protein in an optimized gradient often yields greater than tenfold purification. If conditions cannot be found under which the protein binds to an anion-exchange resin, a

Table 4.1.2.2. *Crystallizing agents for protein crystallization*

(*a*) Salts.

Chemical		No. of macromolecules	No. of crystals
Ammonium salts:	sulfate	802	979
	phosphate	20	21
	acetate	13	13
	chloride, nitrate, citrate, sulfite, formate, diammonium phosphate	1–3	1–3
Calcium salts:	chloride	12	12
	acetate	6	8
Lithium salts:	sulfate	33	34
	chloride	17	19
	nitrate	2	2
Magnesium salts:	chloride	32	32
	sulfate	13	14
	acetate	6	7
Potassium salts:	phosphate	42	79
	chloride	15	17
	tartrate, citrate, fluoride, nitrate, thiocyanate	1–3	1–3
Sodium salts:	chloride	148	186
	acetate	43	46
	citrate	34	36
	phosphate	28	36
	sulfate, formate, nitrate, tartrate	3–10	3–10
	acetate buffer, azide, citrate–phosphate, dihydrogenphosphate, sulfite, borate, carbonate, succinate, thiocyanate, thiosulfate	1 or 2	1 or 2
Other salts:	sodium–potassium phosphate	60	65
	phosphate (counter-ion not specified)	33	39
	caesium chloride	18	24
	phosphate buffer	10	11
	trisodium citrate, barium chloride, sodium–potassium tartrate, zinc(II) acetate, cacodylate (arsenic salt), cadmium chloride	1 or 2	1–3

(*b*) Organic solvents.

Chemical	No. of macromolecules	No. of crystals
Ethanol	63	93
Methanol, isopropanol	27 or 25	31 or 28
Acetone	13	13
Dioxane, 2-propanol, acetonitrile, DMSO, ethylene glycol, *n*-propanol, tertiary butanol, ethyl acetate, hexane-1,6-diol	2–11	3–11
1,3-Propanediol, 1,4-butanediol, 1-propanol, 2,2,2-trifluoroethanol, chloroform, DMF, ethylenediol, hexane-2,5-diol, hexylene-glycol, *N*,*N*-bis(2-hydroxymethyl)-2-aminomethane, *N*-lauryl-*N*,*N*-dimethylamine-*N*-oxide, *n*-octyl-2-hydroxyethylsulfoxide, pyridine, saturated octanetriol, *sec*-butanol, triethanolamine–HCl	1	1

(*c*) Long-chain polymers.

Chemical	No. of macromolecules	No. of crystals
PEG 4000	238	275
PEG 6000	189	251
PEG 8000	185	230
PEG 3350	48	54
PEG 1000, 1500, 2000, 3000, 3400, 10 000, 12 000 or 20 000; PEG monomethyl ether 750, 2000 or 5000	2–18	2–20
PEG 3500, 3600 or 4500; polygalacturonic acid; polyvinylpyrrolidone	1	1

Table 4.2.1.1. *Compilation of membrane proteins with known structures, including crystallization conditions and key references for the structure determinations*

This table is continuously updated and can be inspected at http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html. The membrane proteins listed are divided into polytopic membrane proteins from inner membranes of bacteria and mitochondria (*a*), membrane proteins from the outer membrane of Gram-negative bacteria (*b*) and monotopic membrane proteins [(*c*); these are proteins that are only inserted into the membrane, but do not span it]. Within parts (*a*), (*b*) and (*c*) the membrane proteins are listed in chronological order of structure determination.

(*a*) Polytopic membrane proteins from inner membranes of bacteria and mitochondria.

Membrane protein	Crystallization conditions (detergent/additive/ precipitating agent)	Key references (and pdb reference code, if available)
Photosynthetic reaction centre		
from *Rhodopseudomonas viridis*	*N*,*N*-Dimethyldodecylamine-*N*-oxide/heptane-1,2,3-triol/ammonium sulfate	[1], [2] (1PRC), [3], [4] (2PRC, 3PRC, 4PRC, 5PRC, 6PRC, 7PRC)
from *Rhodobacter sphaeroides*	*N*,*N*-Dimethyldodecylamine-*N*-oxide/heptane-1,2,3-triol/polyethylene glycol 4000	[5] (4RCR)
	Octyl-β-D-glucopyranoside/polyethylene glycol 4000	[6] (2RCR)
	N,*N*-Dimethyldodecylamine-*N*-oxide/heptane-1,2,3-triol, dioxane/potassium phosphate	[7] (1PCR)
	Octyl-β-D-glucopyranoside/benzamidine, heptane-1,2,3-triol/polyethylene glycol 4000	[8] (1AIG, 1AIJ)
Bacteriorhodopsin		
from *Halobacterium salinarium*	(Electron crystallography using naturally occuring two-dimensional crystals)	[9] (1BRD), [10] (2BRD), [11] (1AT9)
	(Type I crystal grown in lipidic cubic phases)	[12] (1AP9), [13] (1BRX)
	Octyl-β-D-glucopyranoside/benzamidine/sodium phosphate (epitaxic growth on benzamidine crystals)	[14] (1BRR)
Light-harvesting complex II		
from pea chloroplasts	(Electron crystallography of two-dimensional crystals prepared from Triton X100 solubilized material)	[15]
Light-harvesting complex 2		
from *Rhodopseudomonas acidophila*	Octyl-β-D-glucopyranoside/benzamidine/ phosphate	[16] (1KZU)
from *Rhodospirillum molischianum*	*N*,*N*-dimethylundecylamine-*N*-oxide/heptane-1,2,3-triol/ammonium sulfate	[17] (1LGH)
Cytochrome *c* oxidase		
from *Paracoccus denitrificans*,		
four-subunit enzyme complexed with antibody Fv fragment	Dodecyl-β-D-maltoside/polyethylene glycol monomethylether 2000	[18]
two-subunit enzyme complexed with antibody Fv fragment	Undecyl-β-D-maltoside/polyethylene glycol monomethylether 2000	[19] (1AR1)
from bovine heart mitochondria	Decyl-β-D-maltoside with some residual cholate/polyethylene glycol 4000	[20], [21] (1OCC), [22] (2OCC, 1OCR)
Cytochrome bc_1 complex		
from bovine heart mitochondria	Decanoyl-*N*-methylglucamide or diheptanoyl phosphatidyl choline/polyethylene glycol 4000	[23] (1QRC), [24]
	Octyl-β-D-glucopyranoside/polyethylene gycol 4000	[25]
	Pure dodecyl-β-D-maltoside or mixture with methyl-6-*O*-(*N*-heptylcarbamoyl)-α-D-glucopyranoside/polyethylene glycol 4000	[26]
from chicken heart mitochondria	Octyl-β-D-glucopyranoside/polyethylene glycol 4000	[25] (1BCC, 3BCC)
Potassium channel		
from *Streptomyces lividans*	*N*,*N*-Dimethyldodecylamine/polyethylene glycol 400	[27] (1BL8)
Mechanosensitive ion channel		
from *Mycobacterium tuberculosis*	Dodecyl-β-D-maltoside/triethylene glycol	[28]

4.3. Application of protein engineering to improve crystal properties

BY D. R. DAVIES AND A. BURGESS HICKMAN

4.3.1. Introduction

There is accelerating use of protein engineering by protein crystallographers. Site-directed mutations are being used for a variety of purposes, including solubilizing the protein, developing new crystal forms, providing sites for heavy-atom derivatives, constructing proteolysis-resistant mutants and enhancing the rate of crystallization. Traditionally, if the chosen protein failed to crystallize, a good strategy was to examine a homologous protein from a related species. Now, the crystallographer has a variety of tools for directly modifying the protein according to his or her choice. This is owing to the development of techniques that make it easy to produce a large number of mutant proteins in a timely manner (see Chapter 3.1).

The relevance to macromolecular crystallography of these mutational procedures rests on the assumption that the mutations do not produce conformation changes in the protein. It is often possible to measure the activity of the protein *in vitro* and, therefore, test directly whether mutation has affected the protein's properties. Several observations suggest that changes of a small number of surface residues can be tolerated without changing the three-dimensional structure of a protein. The work on haemoglobins demonstrated that mutant proteins generally have similar topologies to the wild type (Fermi & Perutz, 1981). The systematic study of T4 phage lysozyme mutants by the Matthews group (Matthews, 1993; Zhang *et al.*, 1995) has confirmed and significantly extended these studies and has provided a basis for mutant design. This work revealed that, for monomeric proteins, 'Substitutions of solvent-exposed amino acids on the surfaces of proteins are seen to have little if any effect on protein stability or structure, leading to the view that it is the rigid parts of proteins that are critical for folding and stability' (Matthews, 1993). It was also concluded that point mutants do not interfere with crystallization unless they affect crystal contacts. The corollary from this is that if the topology of the protein is known from sequence homology with a known structure, the residues that are likely to be located on the surface can be defined and will provide suitable targets for mutation. Fortunately, even in the absence of such information, it is usually possible to make an informed prediction of which residues (generally charged or polar) will, with reasonable probability, be found on the surface.

Here, we shall outline some of the procedures that have been used successfully in protein crystallography. We have tried to provide representative examples of the variety of techniques and creative approaches that have been used, rather than attempting to assemble a comprehensive review of the field. The identification of appropriate references is a somewhat unreliable process, because information regarding these attempts is usually buried in texts; we apologize in advance for any significant omissions.

There have been several reviews on the general topic of the application of protein engineering to crystallography. An overview of the subject is provided by D'Arcy (1994), while Price & Nagai (1995) 'focus on strategies either to obtain crystals with good diffraction properties or to improve existing crystals through protein engineering'. In addition to attempts at a rational approach to protein engineering, it is worth emphasizing the role of serendipity in achieving the goal of diffraction-quality crystals. One example is given by the structure of GroEL (Braig *et al.*, 1994), where better crystals were obtained by the accidental introduction of a double mutation, which arose from a polymerase error during the cloning process. The second example is provided by the search for crystals of the complex between the U1A spliceosomal protein and its RNA hairpin substrate (Oubridge *et al.*, 1995). Initially, only poorly diffracting crystals (7–8 Å) could be obtained, which were similar in morphology to those of the protein alone. A series of mutations were made, designed to improve the crystal contacts, but the end result was a new crystal form that diffracted to 1.7 Å.

Dasgupta *et al.* (1997), in an informative review, have compared the contacts formed between molecules in crystal lattices and in protein oligomerization. They found that there are more polar interactions in crystal contacts, while oligomer contacts favour aromatic residues and methionine. Arginine is the only residue prominent in both, and for a protein that is difficult to crystallize, they recommend replacing lysine with arginine or glutamine. Carugo & Argos (1997) also examined crystal-packing contacts between protein molecules and compared these with contacts formed in oligomers. They observed that the area of the crystal contacts is generally smaller, but that the amino-acid composition of the contacts is indistinguishable from that of the solvent-accessible surface of the protein and is dramatically different from that observed in oligomer interfaces.

4.3.2. Improving solubility

Frequently, a protein is so insoluble that there is only a small probability of direct crystallization. Not only does the limited amount of protein hinder crystallization, but the departure from optimal solubility conditions by the addition of almost any crystallization medium frequently results in rapid precipitation of the protein from solution. When this happens, it is sometimes possible to find surface mutations that enhance solubility. Two strategies have been successfully applied, depending on whether or not the overall topology is known.

An early investigation of the effects of surface mutations (McElroy *et al.*, 1992) involved the crystallization of human thymidylate synthase, where the *Escherichia coli* enzyme structure was known, but the human enzyme could only be crystallized in an apo form unsuitable for studying inhibitors owing to disorder in the active site. The effect of surface mutations was systematically explored by making 12 mutations in 11 positions, and it was found that some of the mutations dramatically changed the protein solubility. Some of the mutant proteins were easier to crystallize than the wild type, and, furthermore, three crystal forms were obtained that differed from that of the wild type.

A second example of the rational design of surface mutations based on prior knowledge of the structure of a related protein is demonstrated by the studies of the trimethoprim-resistant type S1 hydrofolate reductase (Dale *et al.*, 1994). This protein was rather insoluble and precipitated at concentrations greater than 2 mg ml^{-1}. The authors changed four neutral, amide-containing side chains to carboxylates and examined the expressed proteins for improved solubility. Three of the four mutant proteins were more soluble than the wild-type protein, and a double mutant, Asn48 → Glu and Asn130 → Asp, was particularly soluble; this mutant protein crystallized in thick plates, ultimately enabling the structure to be determined.

In the absence of any knowledge of the structure, more heroic procedures are required, as illustrated by the crystallization of the HIV-1 integrase catalytic domain (residues 50–212). This domain had been a focus of intensive crystallization attempts, which were hindered by the low solubility of the protein. The strategy used was to replace all the single hydrophobic residues with lysine and to replace groups of adjacent hydrophobic amino acids with alanines (Jenkins *et al.*, 1995). A simple assay for improved solubility based on the overexpression of the protein was employed, which did not require isolating the purified protein; cell lysis followed by

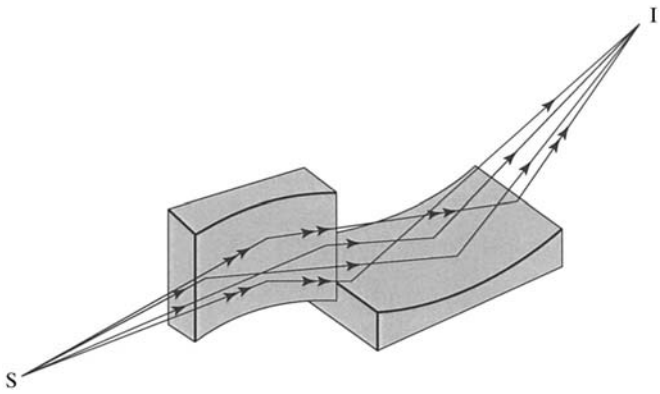


Fig. 6.1.4.1. Production of a point focus by successive reflections at two orthogonal curved mirrors. Arrangement due to Kirkpatrick & Baez (1948) and to Franks (1955).

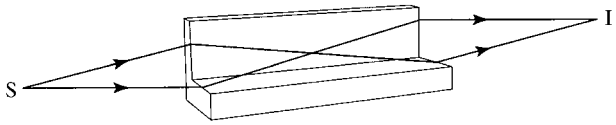


Fig. 6.1.4.2. The 'catamegonic' arrangement of Montel (1957), in which two confocal mirrors with orthogonal curvatures lie side-by-side.

calls a 'catamegonic roof' (Fig. 6.1.4.2). The mirrors are then best made from thicker material, and the reflecting surfaces are ground to the appropriate curvature. The same arrangement has been used by Osmic Inc. (1998) for their Confocal Max-Flux Optics, in which the curved surfaces are coated with graded-spacing multilayers.

Flat mirror plates can be bent elastically to a desired curvature by applying appropriate couples. Fig. 6.1.4.3 shows the bending method adopted by Franks (1955). A cylindrical curvature results from a symmetrical arrangement that produces equal couples at both ends. With appropriate unequal couples applied at the two ends of the plate, the curvature can be made parabolic or elliptical. Precision elliptical mirrors have been produced by Padmore *et al.*

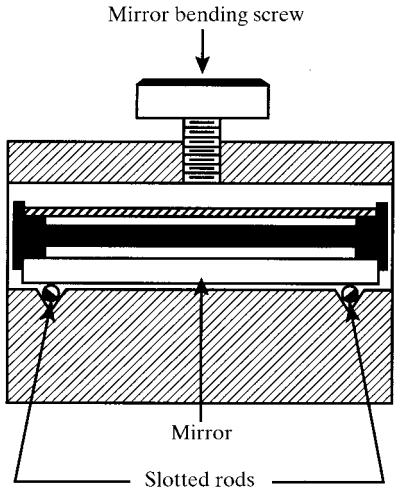


Fig. 6.1.4.3. Mirror bender (after Franks, 1955). The force exerted by the screw produces two equal couples which bend the mirror into a circular arc. The slotted rods act as pivots and also as beam-defining slits.

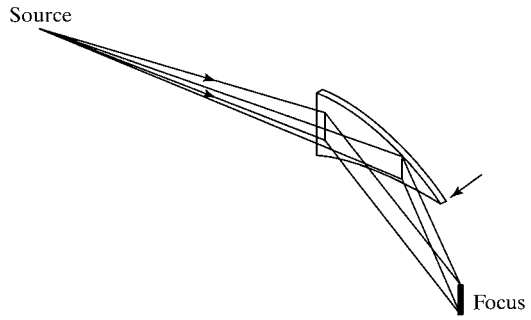


Fig. 6.1.4.4. Triangular mirror bender as described by Lemonnier *et al.* (1978) for crystal plates and by Milch (1983) for glass mirrors. The base of the triangular plate is clamped and the bending force is applied at the apex along the arrow.

(1997); unequal couples are applied in this way. Cylindrically curved mirrors can be produced by applying a force at the tip of a triangular plate whose base is firmly anchored (Fig. 6.1.4.4). Lemonnier *et al.* (1978) first used this method for making curved-crystal monochromators. Milch (1983) described X-ray mirrors made in this way; the effect of the linear increase of the bending moment along the plate is compensated by the linear increase of the plate section so that the curvature is constant. An elliptical or a parabolic curvature results if either the width or the thickness of the plate is made to vary in an appropriate way along the length of the plate. Arndt, Long & Duncumb (1998) described a monolithic mirror-bending block in which the mirror plates are inserted into slots cut to an elliptical curvature by ion-beam machining. The solid angle of collection is made four times larger than for a two-mirror arrangement by providing a pair of horizontal mirrors and a pair of vertical mirrors in tandem in one block (Fig. 6.1.4.5).

Mirror plates for these benders are usually made from highly polished glass, quartz, or silicon plates which are coated with nickel, gold, or iridium.

Mirrors for synchrotron beam lines that focus the radiation in the vertical plane are most often ground and polished to the correct shape, rather than bent elastically. Much longer mirrors can be made in this way.

The collecting efficiency of specularly reflecting mirrors depends on the reflectivity of the surface and on the solid angle of collection; this, in turn, is a function of the maximum glancing angle of incidence, which is the critical angle for total external reflection, θ_c.

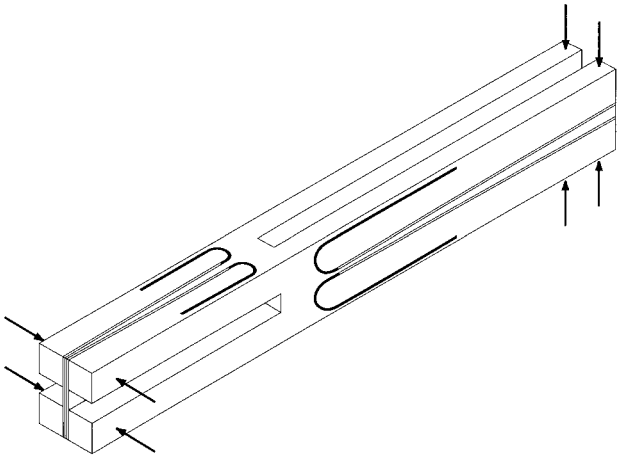


Fig. 6.1.4.5. Mirror holder with machined slots for two orthogonal pairs of curved mirrors (after Arndt, Duncumb *et al.*, 1998).

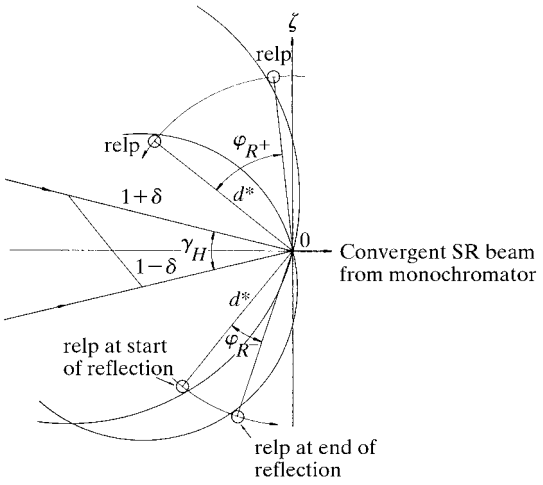


Fig. 8.1.7.3. The rocking width of an individual reflection for the case of Fig. 8.1.7.1(*c*) and a vertical rotation axis. From Greenhough & Helliwell (1982). Copyright (1982) International Union of Crystallography.

$$E \simeq \varphi_R/2L \qquad (8.1.7.11)$$

(Greenhough & Helliwell, 1982).

In Fig. 8.1.7.3, the relevant parameters are shown. The diagram shows $(\delta\lambda/\lambda)_{\rm corr} = 2\delta$ in a plane, usually horizontal with a perpendicular (vertical) rotation axis, whereas the formula for φ_R above is for a horizontal axis. This is purely for didactic reasons since the interrelationship of the components is then much clearer.

8.1.8. Scientific utilization of SR in protein crystallography

There are a myriad of applications and results of the use of SR in crystallography. Helliwell (1992) has produced an extensive survey and tabulations of SR and macromolecular crystallography applications; Chapter 9 therein concentrates on anomalous scattering and Chapter 10 on high resolution, large unit cells, small crystals, weak scattering efficiency and time-resolved data collection. The field has expanded so dramatically, in fact, that an equivalent survey today would be vast. Table 8.1.4.1 lists the home

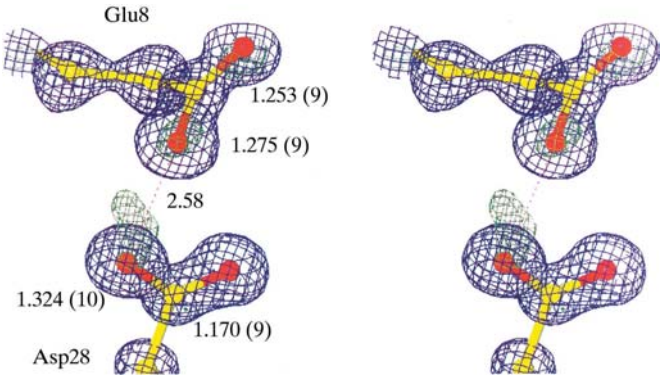


Fig. 8.1.8.1. Determination of the protonation states of carboxylic acid side chains in proteins *via* hydrogen atoms and resolved single and double bond lengths. After Deacon *et al.* (1997) using CHESS. Reproduced by permission of The Royal Society of Chemistry.

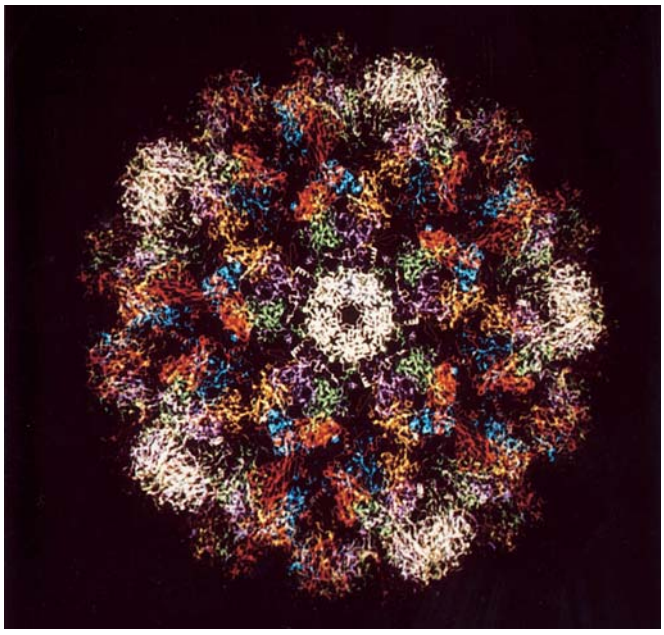


Fig. 8.1.8.2. A view of SV40 virus (based on Liddington *et al.*, 1991) determined using data recorded at the SRS wiggler station 9.6 (Fig. 8.1.4.1*a*).

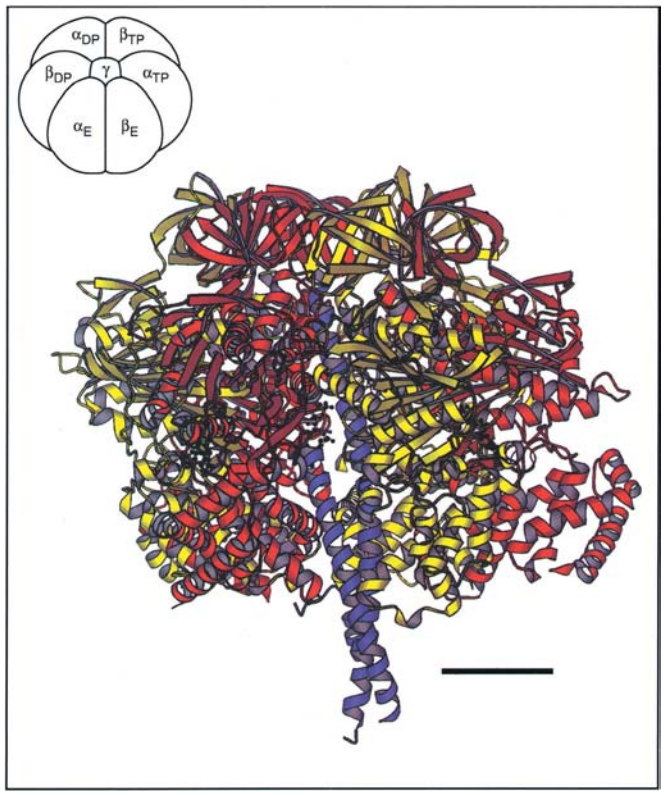


Fig. 8.1.8.3. The protein crystal structure of F_1 ATPase, one of the largest non-symmetrical protein structure complexes, solved using SR data recorded at the SRS wiggler 9.6, Daresbury. The scale bar is 20 Å long. Reprinted with permission from *Nature* (Abrahams *et al.*, 1994). Copyright (1994) MacMillan Magazines Limited.

9. MONOCHROMATIC DATA COLLECTION

9.1. Principles of monochromatic data collection

BY Z. DAUTER AND K. S. WILSON

9.1.1. Introduction

X-ray data collection is the central experiment in a crystal structure analysis. For small-molecule structures, the availability of intensity data to atomic resolution, usually around 0.8 Å, means that the phase problem can be solved directly and the atomic positions refined with a full anisotropic model. This results in a truly automatic structure solution for most small molecules.

Macromolecular crystals pose much greater problems with regard to data collection. The first arise from the size of the unit cell, resulting in lower average intensities of individual reflections coupled with a much greater number of reflections (Table 9.1.1.1). Secondly, the crystals usually contain considerable proportions of disordered aqueous solvent, giving further reduction in intensity at high resolution and, in the majority of cases, restricting the resolution to be much less than atomic. Thirdly, again mostly owing to the solvent content, the crystals are sensitive to radiation damage. Such problems have severe implications for all subsequent steps in a structure analysis. Solution of the phase problem is generally not possible through direct methods, except for a small number of exceptionally well diffracting proteins. The refined models require the imposition of stereochemical constraints or restraints to maintain an acceptable geometry. Recent advances, such as the use of synchrotron beamlines, cryogenic cooling and high-efficiency two-dimensional (2D) detectors, have made data collection technically easier, but it remains a fundamental scientific procedure underpinning the whole structural analysis. Therefore, it is essential to take the greatest care over this key step. The aim of this chapter is to indicate procedures for optimizing data acquisition. Overviews on several issues related to this topic have been published recently (Carter & Sweet, 1997; Turkenburg *et al.*, 1999).

9.1.2. The components of a monochromatic X-ray experiment

To collect X-ray data from single crystals, the following elements are required:

(1) a source of X-rays;
(2) optical elements to focus the X-rays onto the sample;
(3) a monochromator to select a single wavelength;
(4) a collimator to produce a beam of defined dimension;
(5) a shutter to limit the exposure of the sample to X-rays;
(6) a goniostat with associated sample holder to allow rotation of the crystal; and
(7) the crystalline sample itself.

Other desirable elements are:

(1) a cryogenic cooling device for frozen crystals;
(2) an efficient, generally 2D, detector system;
(3) software to control the experiment and store and display the X-ray images;
(4) data-processing software to extract intensities and associated standard uncertainties for the Bragg reflections in the images.

Many of these are discussed elsewhere in this volume. This chapter aims to provide guidance in those areas where choices are to be made by the experimenter and is concerned with the interrelations between parameters and how they conspire for or against different strategies of data collection.

Table 9.1.1.1. *Size of the unit cell and number of reflections*

Compound	Unit cell		Reflections	Average intensity
	Edge (Å)	Volume (Å^3)		
Small organic	10	1000	2000	1
Supramolecule	30	25000	30000	1/25000
Protein	100	1000000	100000	1/1000000
Virus	400	100000000	1000000	1/100000000

9.1.3. Data completeness

The advantage of diffraction methods over spectroscopy is that they provide a full 3D view of the object. Diffraction methods are theoretically limited by the wavelength of the radiation used, but, in practice, every diffraction experiment is further limited by the aperture and quality of the lens. In the X-ray experiment, the aperture corresponds to the resolution limit and the quality of the 'lens' to the completeness and accuracy of the measured Bragg reflection intensities.

In this context, completeness has two components, the first of which is geometric and hence quantitative. It is necessary to rotate the crystal so that all unique reciprocal-lattice points pass through the Ewald sphere and the associated intensities are recorded on the detector. Ideally, the intensities of 100% of the unique Bragg reflections should be measured. The second component is qualitative and statistical: for each *hkl*, the intensity, I_{hkl}, should be significant, with its accuracy correctly estimated in the form of an associated standard uncertainty, $\sigma(I)$. The data should be significant in terms of the $I/\sigma(I)$ ratio throughout the resolution range. This point will be returned to below, but it is especially important that the data at low resolution are complete and not overloaded on the detector, and that there is not an extensive set of essentially zero-level intensities in the higher-resolution shells.

9.1.4. X-ray sources

There are two principal sources of X-rays appropriate for macromolecular data collection: rotating anodes and synchrotron storage rings. These are discussed briefly here and in more detail in Chapters 6.1 and 8.1.

9.1.4.1. *Conventional sources*

Rotating anodes were initially developed for biological scattering experiments on muscle samples and have the advantage of higher intensity compared to sealed-tube generators. They usually have a copper target providing radiation at a fixed wavelength of 1.542 Å. Alternative targets, such as silver or molybdenum, provide lower intensities at short wavelengths, but have not found general applications to macromolecules. Historically, rotating anodes were first used with nickel filters to give monochromatic Cu $K\alpha$ radiation. Current systems are equipped with either graphite

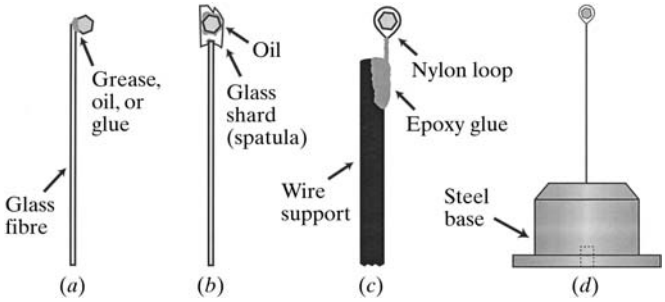


Fig. 10.2.3.1. Different crystal mounts for flash cooling and cryogenic data collection. (*a*) Crystal mounted on a thin glass fibre with adhesive, grease, or oil. (*b*) Crystal placed in a hydrocarbon oil and then scooped onto a thin glass shard. (*c*) Crystal suspended in a film of aqueous solution within a nylon loop. The loop is attached to a thin (~0.25 mm diameter) wire support. (*d*) A diagram of the entire loop-mount assembly. The base is made of plain steel or a magnetic alloy and has two holes, one for the wire post and one for a locating pin, which reproducibly positions the assembly on the goniometer.

The technique is quick and straightforward, remarkably gentle to the crystal, and provides a large surface area for cooling.

The loops are generally formed from nylon fibre, although glass wool is useful for larger versions because its rigidity keeps them from collapsing under the surface tension of the suspended film. Both types of fibres should have a diameter of approximately 10 μm. This small cross section reduces absorption and scattering from the material itself and also minimizes the thickness of the film in the loop. Several methods of making the loops have been described in detail (Rodgers, 1997; Garman & Schneider, 1997), and nylon loops of different sizes are available commercially. The loop is usually glued to a thin metal wire or other heat-conductive post. The ability to conduct heat rapidly is required to minimize ice formation at the point where the wire or post exits the cold gas stream of the cryostat, which occurs in some orientations of the loop assembly. This post is in turn attached to a steel base, which is used with the magnetic transfer system described below.

Crystals are placed in the loop as shown in Fig. 10.2.3.2. They can be mounted directly from the crystallization drop or after harvesting into any convenient container. Under a stereomicroscope, the crystal is teased to the surface of the solution, usually with the loop itself. Once at the surface, the crystal is carried through the interface by first resting it on the bottom of the loop and then moving the assembly vertically to pull it out of the solution. A practiced experimentalist can usually capture the crystal in the first few tries. The plane of the loop should be kept near the vertical to increase the chance of catching the crystal and to minimize the amount of liquid drawn up with it. An alternative technique is to use a small pipette to place the crystal and a drop of cryosolvent into the loop and then draw off the excess solution with filter paper. In either case, it can be difficult to form a film in the loop with solutions high in organic solvent due to the lack of surface tension. For these solutions, adding PEG up to a few per cent usually allows a stable film to form. Fig. 10.2.3.3 is a photograph of a crystal mounted in a nylon loop. If the diameter of the loop is chosen so that it just accommodates the crystal, mounting is easier and the amount of extra scattering material in the X-ray beam is reduced. Also, asymmetric crystals can then be oriented relative to the assembly by preforming the loop into the appropriate shape.

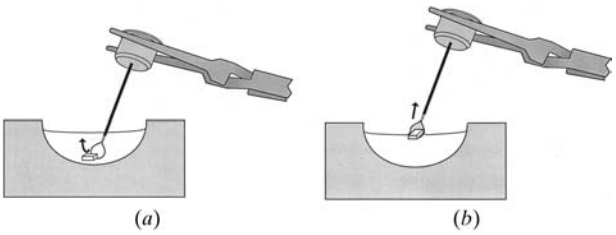


Fig. 10.2.3.2. Mounting a crystal in a loop. (*a*) While viewing with a stereomicroscope, the crystal is teased to the surface of the liquid using the loop. (*b*) It is then drawn through the interface and into the loop. The sizes of the loop and crystal have been exaggerated. Reproduced with permission from Rodgers (1997). Copyright (1997) Academic Press.

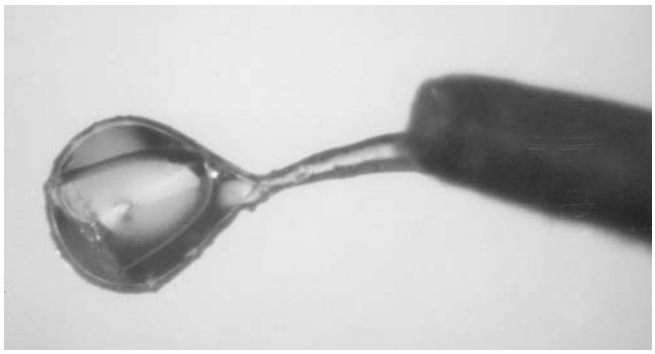


Fig. 10.2.3.3. Photograph of a flash-cooled crystal mounted in a nylon loop. The wire post holding the loop is visible on the right. Reprinted from Rodgers (1994) with permission from Elsevier Science.

The loop-mounting technique can also be used for data collection above cryogenic temperatures by sealing the loop and pin in a large diameter (3 mm) glass or quartz X-ray capillary (Fig. 10.2.3.4). A guard composed of stiff wax or a plastic plug cemented to the pin helps to guide the capillary over the sample before sealing it to the base with high vacuum grease or a cement low in volatile solvent. Loop mounting can be less damaging for many crystals than capillary mounting, and it results in a more uniform X-ray absorption surface.

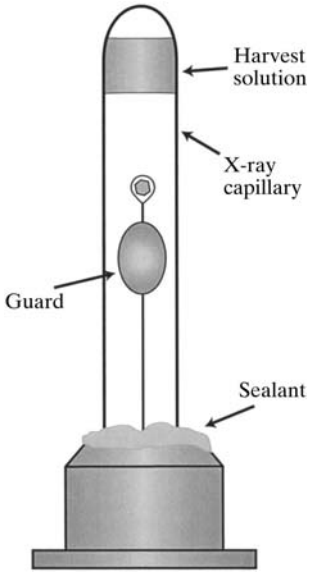


Fig. 10.2.3.4. Arrangement for using the loop-mounting technique at non-cryogenic temperatures.

Table 12.1.3.1. *Useful pH ranges of some heavy-atom reagents derived from the heavy-atom data bank*

No. of entries	Minimum	Average	Maximum	Compound
159	3.0	6.7	9.1	Potassium tetrachloroplatinum(II)
63	4.2	6.6	9.0	Potassium dicyanoaurate(I)
53	4.2	6.9	9.5	Mercury(II) chloride
59	2.8	6.7	9.0	Mercury(II) acetate
52	4.7	6.7	9.3	4-(Chloromercurio)benezenesulfonic acid
57	2.0	6.5	9.3	Potassium tetraiodomercurate(II)
36	5.4	6.7	8.5	Ethylmercurythiosalicylate (EMTS)
46	4.0	6.0	8.0	Potassium pentafluorooxyuranate(VI)
2	8.2	8.4	8.5	Barium(II) chloride
22	4.0	6.2	8.1	Lead(II) acetate
13	4.5	6.6	7.5	Lead(II) nitrate
1	6.5	6.5	6.5	Strontium(II) acetate
3	6.3	6.8	7.5	Thallium(I) acetate
2	5.9	6.6	7.2	Thallium(III) chloride
5	5.0	5.8	6.8	Gadolinium(III) chloride
9	4.9	6.7	7.5	Samarium(III) nitrate
7	4.9	6.6	8.7	Neodymium(III) chloride
64	4.1	6.3	8.6	Uranium(VI) oxyacetate

Thus the number and occupancy of sites can be manipulated by varying the pH, often after cross-linking the crystals to stabilize them.

Extremes in pH can give rise to considerable difficulties in establishing suitable derivatives, as hydrogen and hydroxyl ions compete with the metal ion/complex for the protein and with the protein for the metal ion/complex. At extremely high pH values metals in solution tend to form insoluble hydroxides. The ranges of pH values that are useful for metal ions are given in Table 12.1.3.1.

Varying the reactivity of amino-acid side chains by manipulation of the pH can enable the same heavy-atom ion/complex to bind at different sites, thus producing more than one derivative useful for phase determination.

12.1.3.5. *Effect of precipitants and buffers on heavy-atom binding*

Components present in the heavy-atom solution can have a profound effect on protein–heavy-atom interactions. The salting in/out agent (precipitant) and buffer are the principal sources of alternative ligands for the heavy-atom reagents, while protons compete with the heavy-atom ion/complex for the reactive amino-acid side chains.

Ammonium sulfate is the most successful precipitant in protein crystallization experiments (Gilliland *et al.*, 1994). However, its continued presence in the mother liquor can cause problems by interfering with protein–heavy-atom interactions. At high hydrogen-ion concentrations, the NH_3 group is protonated (*i.e.* NH_4^+), but as the pH rises the proton is lost, typically around pH 6.0–7.0, enabling the group to compete with the protein for the heavy-atom reagent by an S_{N2} reaction.

The nucleophilic strength of potential ligands follows the order

$$NH_3 > Cl^- > H_2O.$$

The anionic complex $PtCl_4^{2-}$ is present in excess ammonia at pH > 7.0 and it will react:

$$PtCl_4^{2-} \rightarrow \textit{cis}\text{-}PtCl_2(NH_3)_2 \rightarrow Pt(NH_3)_4^{2+}.$$

The resultant cationic complex is less susceptible to reaction due to the *trans* effect of NH_3. Pd, Au, Ag and Hg complexes react in a similar way. Decreasing the pH of the solution reduces the amount of free ammonia available through protonation (Sigler & Blow, 1965). Such a technique may give rise to other problems (*e.g.* cracked crystal, decreased nucleophilicity of the protein ligands).

Changing the precipitant to sodium/potassium phosphate or magnesium sulfate may alleviate the situation, but it may also present other problems. For instance, PO_4^{3-} displaces Cl^- from $PtCl_4^{2-}$, thus increasing the negative charge. Both PO_4^{3-} and SO_4^{2-} form insoluble complexes with class A metals (*e.g.* lanthanide and uranyl cations) (Petsko *et al.*, 1978). Both acetate and citrate form complexes with class A metals, but citrate, a chelating ion, binds more strongly. Tris buffer is probably preferable; it binds many cations, but the complexes formed tend to be relatively unstable.

12.1.3.6. *Solubility of heavy-atom compounds*

The solubility of a heavy-atom compound will depend upon the precipitant, buffer and pH. Typically, the component present in the highest concentration is the precipitant, either as salts (*e.g.* ammonium sulfate) or as an organic-based reagent (*e.g.* ethanol, MPD, PEG). Heavy-atom compounds that are essentially covalent and organic in character will be more soluble in ethanol, MPD, PEGs and other organic precipitants.

Although the solubility of tetrakis(acetoxymercurio)methane (TAMM) is higher than most multiple-heavy-atom compounds in aqueous solutions, the presence of glycylglycine or charged mercaptans, such as cysteamine or penicillamine, can increase solubility further (Lipka *et al.*, 1976). The ratio of TAMM to solubilization agent (*e.g.* glycylglycine) is typically 1:10. Even so, the final solubility of TAMM depends on the concentration of competing anions (*e.g.* chloride) (O'Halloran *et al.*, 1987).

Many organometallic compounds are relatively insoluble in aqueous solutions, but their solubility may be increased by predissolving in an aprotic solvent such as acetonitrile.

Iodine and several inorganic iodide salts are insoluble in aqueous solutions. This can be rectified by dissolving the heavy-atom compounds in an aqueous solution of KI.

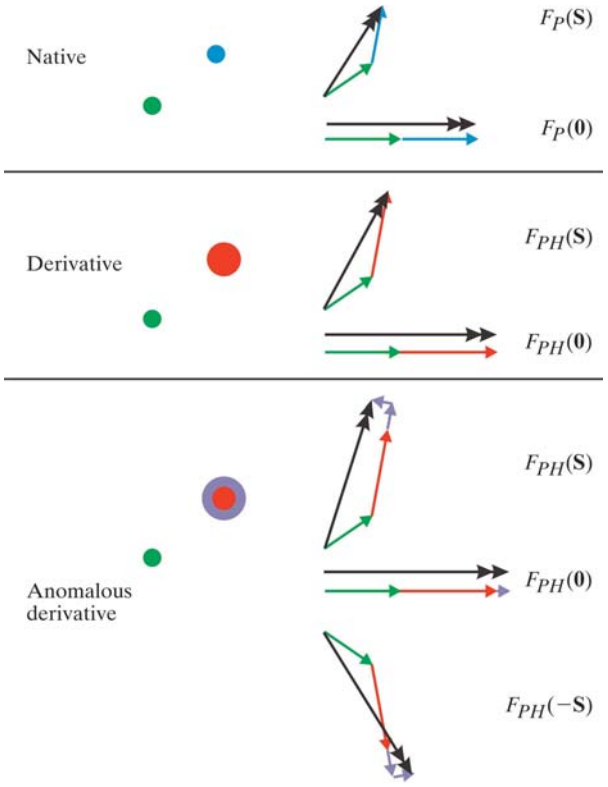


Fig. 12.2.1.2. The effect of introducing a heavy atom or anomalous scatterer. The native two-atom structure gives rise to two diffraction vectors (green and blue) of equal magnitude but different phase (see Chapter 2.1), with a resultant diffraction vector F_P (black). Isomorphous replacement of the blue atom by the larger red one gives rise to a diffraction vector of greater magnitude but equivalent phase (red), causing a change in the resultant magnitude F_{PH} (and hence the intensity) *and* in the phase. Introduction of an anomalous scatterer results in a phase shift (lilac) of the diffraction vector, resulting in differing amplitudes and phases for $F_{PH}(\mathbf{S})$ and $F_{PH}(-\mathbf{S})$.

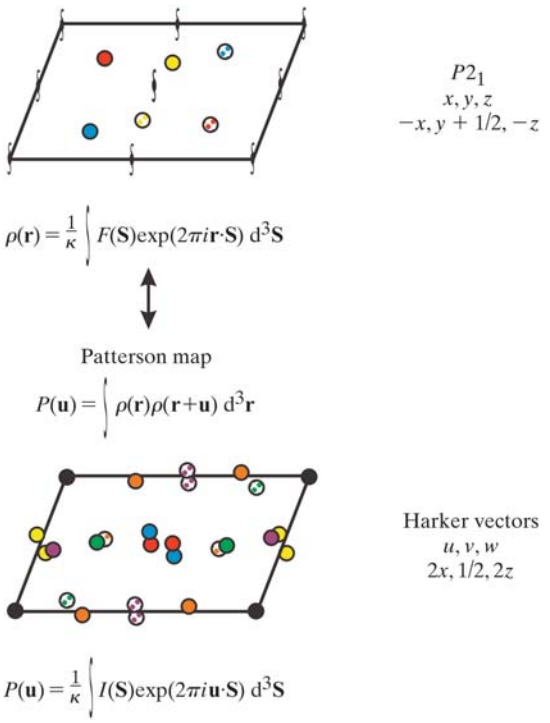


Fig. 12.2.2.1. The Patterson map with symmetry. When the crystal unit cell contains more than one molecule, then additional cross vectors will be formed between differing molecules. If these are related by crystallographic symmetry, there is a geometrical relationship between cross peaks. In this diagram, the peaks of Fig. 12.2.1.1 are supplemented by those between atoms of symmetry-related molecules. The red, yellow and blue peaks of the resulting Patterson function represent those between same atoms (*i.e.* red to red, yellow to yellow and blue to blue) related by symmetry. These peaks are found on a Harker section.

In order to obtain phase information from isomorphous replacement (or from anomalous dispersion), it is necessary to locate the atomic positions of the heavy-atom (or anomalous) scatterers.

12.2.2. The Patterson function

Although the set of measured intensities contains no information regarding the phases, the Fourier transform of the intensities, the so-called Patterson function, contains valuable information. Patterson (1934) showed that the inverse Fourier transform of the intensity,

$$P(uvw) = (1/V)\sum_{hkl} I(hkl) \exp\{-2\pi i(hu + kv + lw)\},$$

is related to the electron density by

$$P(\mathbf{u}) = \int \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})\, \mathrm{d}^3\mathbf{r}.$$

The Patterson function $P(\mathbf{u})$ is an autocorrelation function of the density. For every vector $\mathbf{u}$ that corresponds to an interatomic vector, $P(\mathbf{u})$ will contain a peak (Fig. 12.2.1.1). These are some properties of the Patterson function:

(1) Every atom makes an 'interatomic vector' with itself, and therefore the *origin peak*, $P(\mathbf{0}) = \sum \rho^2(\mathbf{r})$, dominates the Patterson function. This origin peak can be 'removed' through subtraction of the average intensity from $I(hkl)$ before Fourier transformation.

(2) For every vector between $\rho_i(\mathbf{r}_i)$ and $\rho_j(\mathbf{r}_j)$, the same value (*i.e.* their product) is found for $\rho_j(\mathbf{r}_j)$ to $\rho_i(\mathbf{r}_i)$, and so the Patterson map is *centrosymmetric*.

(3) For a structure consisting of n atoms, there are $n(n-1)/2$ cross vectors, and so the Patterson function is extremely crowded.

For simple crystals, the Patterson map can be used to solve the structure directly. For macromolecular structures, the Patterson map provides a vehicle for solving the phase problem.

If the crystal contains rotational symmetry elements, then the cross vectors between $\rho_i(\mathbf{r}_i)$ and its symmetry mate lie on a plane perpendicular to the symmetry axis – the *Harker section* (Harker, 1956). By way of example, the space group $P2_1$ has two symmetry-related positions (Fig. 12.2.2.1),

$$(x, y, z) \text{ and } (-x, y + \tfrac{1}{2}, -z).$$

Cross vectors between symmetry-related points will therefore have the form

$$(2x, \tfrac{1}{2}, 2z),$$

i.e. all cross vectors lie on the plane $v = \frac{1}{2}$. For space group $P2_12_12_1$, the general coordinates

$$(x, y, z), (x + \tfrac{1}{2}, -y + \tfrac{1}{2}, -z), (-x + \tfrac{1}{2}, -y, z + \tfrac{1}{2}),$$
$$(-x, y + \tfrac{1}{2}, -z + \tfrac{1}{2})$$

give rise to cross vectors

$$(\tfrac{1}{2}, 2y + \tfrac{1}{2}, 2z), (2x + \tfrac{1}{2}, 2y, \tfrac{1}{2}), (2x, \tfrac{1}{2}, 2z + \tfrac{1}{2}),$$

13.2. Rotation functions

BY J. NAVAZA

13.2.1. Overview

We will discuss a technique to find either the relative orientations of homologous but independent subunits connected by noncrystallographic symmetry (NCS) elements or the absolute orientations of these subunits if the structure of a similar molecule or fragment is available. The procedure makes intensive use of properties of the rotation group, so we will start by recalling some properties of rotations. More advanced results are included in Appendix 13.2.1.

13.2.2. Rotations in three-dimensional Euclidean space

A rotation **R** is specified by an oriented axis, characterized by the unit vector **u**, and the spin, χ, about it. Positive spins are defined by the right-hand screw sense and values are given in degrees. An almost one-to-one correspondence between rotations and parameters (χ, **u**) can be established. If we restrict the spin values to the positive interval $0 \leq \chi \leq 180$, then for each rotation there is a unique vector $\chi\mathbf{u}$ within the sphere of radius 180. However, vectors situated at opposite points on the surface correspond to the same rotation, *e.g.* (180, **u**) and (180, $-$**u**).

When the unit vector **u** is specified by the colatitude ω and the longitude φ with respect to an orthonormal reference frame (see Fig. 13.2.2.1*a*), we have the spherical polar parameterization of rotations (χ, ω, φ). The range of variation of the parameters is

$$0 \leq \chi \leq 180;\ 0 \leq \omega \leq 180;\ 0 \leq \varphi < 360.$$

Rotations may also be parameterized with the Euler angles (α, β, γ) associated with an orthonormal frame (**x**, **y**, **z**). Several conventions exist for the names of angles and definitions of the axes involved in this parameterization. We will follow the convention by which (α, β, γ) denotes a rotation of α about the z axis, followed by a rotation of β about the nodal line n, the rotated y axis, and finally a rotation of γ about **p**, the rotated z axis (see Fig. 13.2.2.1*b*):

$$\mathbf{R}(\alpha,\beta,\gamma) = \mathbf{R}(\gamma,\mathbf{p})\mathbf{R}(\beta,\mathbf{n})\mathbf{R}(\alpha,\mathbf{z}). \qquad (13.2.2.1)$$

The same rotation may be written in terms of rotations around the fixed orthonormal axes. By using the group property

$$\mathbf{T}\mathbf{R}(\chi,\mathbf{u})\mathbf{T}^{-1} = \mathbf{R}(\chi,\mathbf{T}\mathbf{u}), \qquad (13.2.2.2)$$

which is valid for any rotation **T**, we obtain (see Appendix 13.2.1)

$$\mathbf{R}(\alpha,\beta,\gamma) = \mathbf{R}(\alpha,\mathbf{z})\mathbf{R}(\beta,\mathbf{y})\mathbf{R}(\gamma,\mathbf{z}). \qquad (13.2.2.3)$$

The parameters (α, β, γ) take values within the parallelepiped

$$0 \leq \alpha < 360;\ 0 \leq \beta \leq 180;\ 0 \leq \gamma < 360.$$

Here again, different values of the parameters may correspond to the same rotation, *e.g.* (α, 180, γ) and $(\alpha-\gamma, 180, 0)$.

Although rotations are abstract objects, there is a one-to-one correspondence with the orthogonal matrices in three-dimensional space. In the following sections, **R** will denote a 3×3 orthogonal matrix. An explicit expression for the matrix which corresponds to the rotation (χ, **u**) is

$$\begin{bmatrix} \cos\chi + u_1u_1(1-\cos\chi) & u_1u_2(1-\cos\chi) - u_3\sin\chi & u_1u_3(1-\cos\chi) + u_2\sin\chi \\ u_2u_1(1-\cos\chi) + u_3\sin\chi & \cos\chi + u_2u_2(1-\cos\chi) & u_2u_3(1-\cos\chi) - u_1\sin\chi \\ u_3u_1(1-\cos\chi) - u_2\sin\chi & u_3u_2(1-\cos\chi) + u_1\sin\chi & \cos\chi + u_3u_3(1-\cos\chi) \end{bmatrix} \qquad (13.2.2.4)$$

or, in condensed form,

$$\mathbf{R}(\chi,\mathbf{u})_{ij} = \delta_{ij}\cos\chi + u_iu_j(1-\cos\chi) + \sum_{k=1}^{3}\varepsilon_{ikj}u_k\sin\chi, \qquad (13.2.2.5)$$

where δ_{ij} is the Kronecker tensor, u_i are the components of **u**, and ε_{ijk} is the Levi–Civita tensor. The rotation matrix in the Euler parameterization is obtained by substituting the matrices in the right-hand side of equation (13.2.2.3) by the corresponding expressions given by equation (13.2.2.4).

13.2.2.1. *The metric of the rotation group*

The idea of distance between rotations is necessary for a correct formulation of the problem of sampling and for plotting functions of rotations (Burdina, 1971; Lattman, 1972). It can be demonstrated that the quantity

$$\mathrm{d}s^2 = \mathrm{Tr}(\mathrm{d}\mathbf{R}\ \mathrm{d}\mathbf{R}^{+}) = \sum_{i,j=1}^{3}(\mathrm{d}R_{ij})^2 \qquad (13.2.2.6)$$

defines a metric on the rotation group, unique up to a multiplicative

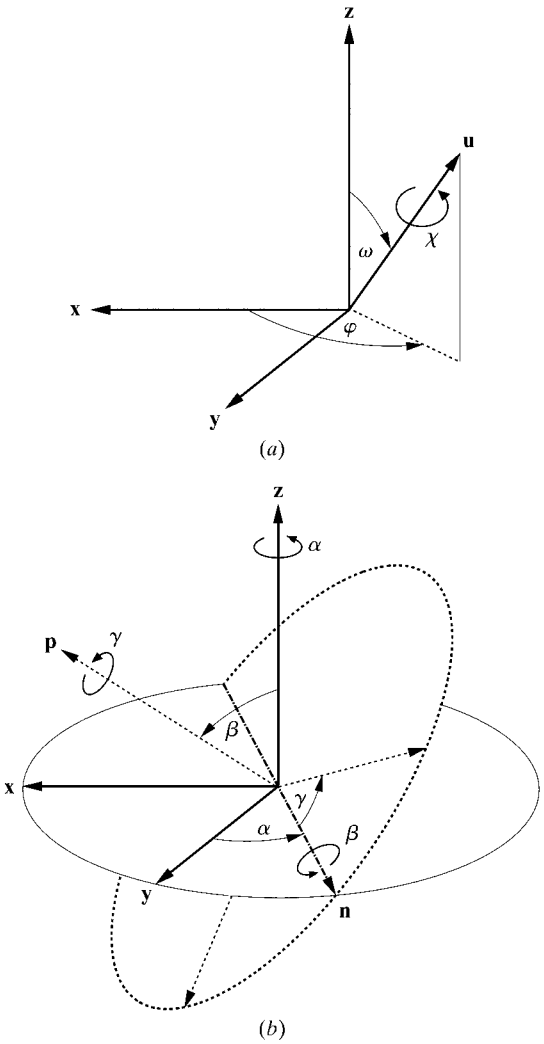


Fig. 13.2.2.1. Illustration of rotations defined by (*a*) the spherical polar angles (χ, ω, φ); (*b*) the Euler angles (α, β, γ).

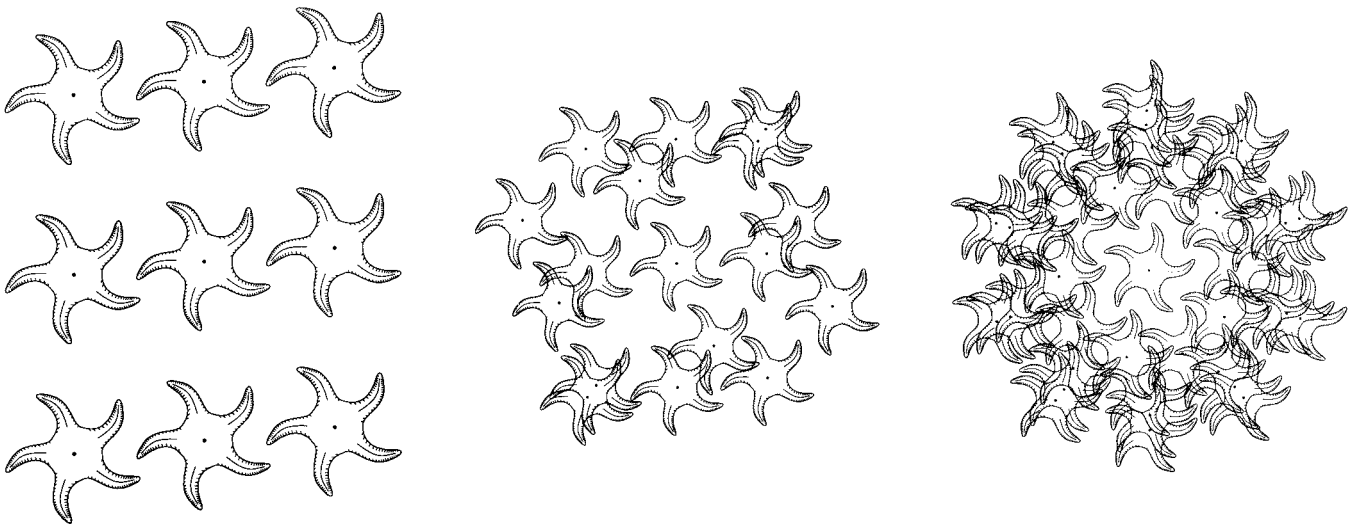


Fig. 13.4.2.2. (*a*) NCS in a triclinic cell. (*b*) Superposition of the pattern in (*a*) on itself after operation with the noncrystallographic fivefold axis. (*c*) Superposition of the pattern in (*a*) on itself after a rotation of one-fifth, two-fifths, three-fifths and four-fifths. Note that the sum or product of periodic patterns is aperiodic and in (*c*) has the point symmetry of the noncrystallographic operation. [Reprinted with permission from Rossmann (1990). Copyright (1990) International Union of Crystallography.]

their orthogonalized *a*, *b* and *c* axes parallel) must equally be an improper rotation.

The position in space of a noncrystallographic rotation symmetry operator can be arbitrarily assigned. The rotation operation will orient the two molecules similarly. A subsequent translation, whose magnitude depends upon the location of the NCS operator, will always be able to superimpose the molecules (Fig. 13.4.2.3). Nevertheless, it is possible to select the position of the NCS axis such that the translation is a minimum, and that will occur when the translation is entirely parallel to the noncrystallographic rotation axis.

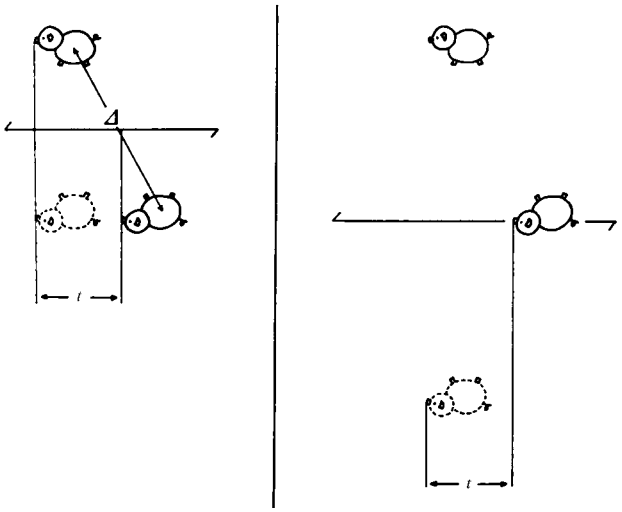


Fig. 13.4.2.3. The position of the twofold rotation axis which relates the two piglets is completely arbitrary. The diagram on the left shows the situation when the translation is parallel to the rotation axis. The diagram on the right has an additional component of translation perpendicular to the rotation axis, but the component parallel to the axis remains unchanged. [Reprinted with permission from Rossmann *et al.* (1964). Copyright (1964) International Union of Crystallography.]

The position of an NCS axis, like everything else in the unit cell, must be defined with respect to a selected origin. Consider the noncrystallographic rotation defined by the 3×3 matrix [C]. Then, if the point $\mathbf{x}$ is rotated to $\mathbf{x}'$ (both defined with respect to the selected origin and axial system),

$$\mathbf{x}' = [\mathrm{C}]\mathbf{x} + \mathbf{d},$$

where $\mathbf{d}$ is a three-dimensional vector which expresses the translational component of the NCS operation. The magnitude of the components of $\mathbf{d}$ is quite arbitrary unless the position of the rotation axis is defined. If the rotation axis represents a proper NCS element, there will exist a point $\mathbf{x}$ on the rotation axis, when positioned to eliminate translation, such that it is rotated onto $\mathbf{x}'$. It follows that for such a point

$$\mathbf{x} = [\mathrm{C}]\mathbf{x} + \mathbf{d},$$

from which $\mathbf{d}$ can be determined if the position of the molecular centre is known. Note that $\mathbf{d} = 0$ if, and only if, the noncrystallographic rotation axis passes through the crystallographic origin.

The presence of proper NCS in a crystal can help phase determination considerably. Consider, for example, a tetramer with 222 symmetry. It is not necessary to define the chemical limits of any one polypeptide chain as the NCS is true everywhere within the molecular envelope and the boundaries of the polypeptide chain are irrelevant to the geometrical considerations. The electron density at every point within the molecular envelope (which itself must have 222 symmetry) can be averaged among all four 222-related points without any chemical knowledge of the configuration of the monomer polypeptide. On the other hand, if there is only improper NCS, then the envelope must define the limits of one noncrystallographic asymmetric unit, although the crystallographic asymmetric unit contains two or more such units.

13.4.3. Phase determination using NCS

The molecular replacement method [*cf.* Rossmann & Blow (1962); Rossmann (1972, 1990); Argos & Rossmann (1980); Rossmann & Arnold (2001)] is dependent upon the presence of NCS, whether it

14. ANOMALOUS DISPERSION

14.1. Heavy-atom location and phase determination with single-wavelength diffraction data

BY B. W. MATTHEWS

14.1.1. Introduction

As is well known, the successful introduction of the method of isomorphous replacement by Green *et al.* (1954) was the turning point in the subsequent development of protein crystallography as we now know it.

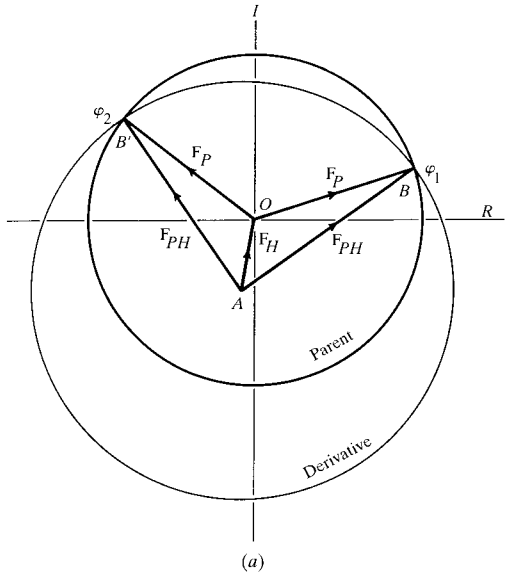
The idea that the phases of X-ray reflections from a protein crystal could be obtained by the introduction of heavy atoms into the crystal was not new, having been suggested by J. D. Bernal in 1939 (Bernal, 1939). The isomorphous-replacement method was used as early as 1927 by Cork (1927) in studying the alums. Bokhoven *et al.* (1951) subsequently extended the method to the study of a noncentrosymmetric projection of strychnine sulfate, using what would now be termed the method of single isomorphous replacement. They also suggested that by using a double isomorphous replacement, a unique phase determination could be obtained, even for noncentrosymmetric reflections. The details of the double (or multiple) isomorphous-replacement method were worked out by Harker (1956), who introduced the very useful concept of phase circles. Another contribution which was of great practical value, and which will provide the basis for much of the subsequent discussion, is the method introduced by Blow & Crick (1959) for the treatment of errors in the isomorphous-replacement method. In addition to the determination of protein phases by the method of substitution with heavy atoms, it is now routine to supplement this information by utilizing the anomalous scattering of the substituted atoms. The underlying principles trace back to articles by Bijvoet (1954), Ramachandran & Raman (1956), and Okaya & Pepinsky (1960). The first application of the anomalous-scattering method to protein crystallography was by Blow (1958), who used the anomalous scattering of the iron atoms to determine phase information for a noncentrosymmetric projection of horse oxyhaemoglobin.

In the following discussion, we first review the classical method of phase determination by isomorphous replacement, then discuss the inclusion of single-wavelength anomalous-scattering data, and conclude by discussing the use of such data for heavy-atom location. Part of the review is based on Matthews (1970).

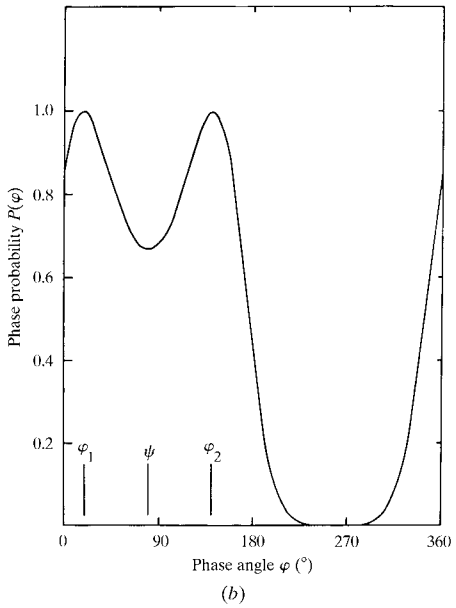
14.1.2. The isomorphous-replacement method

Consider a protein crystal with an isomorphous heavy-atom derivative, *i.e.* a modified crystal in which heavy atoms occupy specific sites throughout the crystal, but which is in all other respects identical to the unsubstituted 'parent' crystal. Let the structure factors of the protein crystal be $\mathbf{F}_P(\mathbf{h})$, of the isomorph be $\mathbf{F}_{PH}(\mathbf{h})$, and of the heavy atoms $\mathbf{F}_H(\mathbf{h})$. (Note: Structure *amplitudes* are indicated by italic type, *e.g.* F_P, and *vectors* by bold-face type, *e.g.* $\mathbf{F}_P$.) In practice, one can measure the structure amplitudes F_P and F_{PH}, and it is desired to obtain from these observable quantities the value of the phase angle of $\mathbf{F}_P(\mathbf{h})$ so that a Fourier synthesis showing the electron density of the protein structure may be calculated. It will be assumed, for the moment, that the positions and occupancy of the sites of heavy-atom binding have been determined as accurately as possible.

From the heavy-atom parameters, the corresponding structure factor $\mathbf{F}_H(\mathbf{h})$ is calculated. To determine φ, the phase of $\mathbf{F}_P(\mathbf{h})$, we construct a set of phase circles, as proposed by Harker (1956). From a chosen origin O (Fig. 14.1.2.1*a*), the vector OA is drawn equal to $-\mathbf{F}_H$. Circles of radius F_P and F_{PH} are then drawn about O and A, respectively. The intersections of the phase circles at B and B' define two possible phase angles for F_P. Note that the angles are



(*a*)



(*b*)

Fig. 14.1.2.1. (*a*) Harker construction for a single isomorphous replacement. φ_1 and φ_2 are the 'most probable' phases for $\mathbf{F}_P$. (*b*) Phase probability distribution for a single isomorphous replacement. This and subsequent probabilities are unnormalized. [All figures in this chapter are reproduced with permission from Matthews (1970). Copyright (1970) International Union of Crystallography.]

14.2. MAD and MIR

BY J. L. SMITH, W. A. HENDRICKSON, T. C. TERWILLIGER AND J. BERENDZEN

14.2.1. Multiwavelength anomalous diffraction

(J. L. SMITH AND W. A. HENDRICKSON)

Anomalous-scattering effects measured at several X-ray wavelengths can provide a direct solution to the crystallographic phase problem. For many years this was appreciated as a hypothetical possibility (Okaya & Pepinsky, 1956), but, until tunable synchrotron radiation became available, experimental investigation with the weakly diffracting crystals of biological macromolecules was limited to one heroic experiment (Hoppe & Jakubowski, 1975). Multiwavelength anomalous diffraction (MAD) became a dominant phasing method in macromolecular crystallography with the advent of reliable, brilliant synchrotron-radiation sources, the adoption of cryopreservation techniques for crystals of macromolecules, and the development of general anomalous-scatterer labels for proteins and nucleic acids.

Anomalous scattering, first recognized as a source of phase information by Bijvoet (1949), has been employed since the early days of macromolecular crystallography (Blow, 1958). It has been used to locate positions of anomalous scatterers (Rossmann, 1961), to supplement phase information from isomorphous replacement (North, 1965; Matthews, 1966*a*) and to identify the enantiomorph of the heavy-atom partial structure in multiple isomorphous replacement (MIR) phasing (Matthews, 1966*b*). Anomalous scattering at a single wavelength was the sole source of phase information in the structure determination of crambin (Hendrickson & Teeter, 1981), an important precursor to development of MAD. MAD differs from these other applications in using anomalous scattering at several wavelengths for complete phase determination without approximations or simplifying assumptions.

14.2.1.1. *Anomalous scattering factors*

The scattering of X-rays by an isolated atom is described by the atomic scattering factor, f^0, based on the assumption that the electrons in the atom oscillate as free electrons in response to X-ray stimulation. The magnitude of f^0 is normalized to the scattering by a single electron. Thus the 'normal' scattering factor f^0 is a real number, equal to the Fourier transform of the electron-density distribution of the atom. At zero scattering angle ($s = \sin\theta/\lambda = 0$), f^0 equals Z, the atomic number. f^0 falls off rapidly with increasing scattering angle due to weak scattering by the diffuse parts of the electron-density distribution. In reality, electrons in an atom do not oscillate freely because they are bound in atomic orbitals. Deviation from the free-electron model of atomic scattering is known as anomalous scattering. Using a classical mechanical model (James, 1948), an atom scatters as a set of damped oscillators with resonant frequencies matched to the absorption frequencies of the electronic shells. The total atomic scattering factor, f, is thus a complex number. f is denoted as a sum of 'normal' and 'anomalous' components, where the anomalous components are corrections to the free-electron model:

$$f = f^0 + f' + if''. \qquad (14.2.1.1)$$

f' and f'' are expressed in electron units, as is f^0. The real component of anomalous scattering, f', is in phase with the normal scattering, f^0, whilst the imaginary component, f'', is out of phase by $\pi/2$.

The imaginary component of anomalous scattering, f'', is proportional to the atomic absorption coefficient of the atom, μ_a, at X-ray energy E:

$$f''(E) = (mc/4\pi e^2\hbar)E\mu_a(E), \qquad (14.2.1.2)$$

where m is the electronic mass, c is the speed of light, e is the electronic charge and h ($= 2\pi\hbar$) is Planck's constant. Thus, f'' can be determined experimentally by measurement of the atomic absorption coefficient. The relationship between f'' and f' is known as the Kramers–Kronig dispersion relation (James, 1948; Als-Nielsen & McMorrow, 2001):

$$f'(E) = \left(\frac{2}{\pi}\right) P \int_0^\infty \frac{E' f''(E')}{E^2 - E'^2}\, \mathrm{d}E', \qquad (14.2.1.3)$$

where P represents the Cauchy principal value of the integral such that integration over E' is performed from 0 to $(E - \varepsilon)$ and from $(E + \varepsilon)$ to ∞, and then the limit $\varepsilon \to 0$ is taken. The principal value of the integral can be evaluated numerically from limited spectral data that have been scaled to theoretical f'' scattering factors (or μ_a absorption coefficients) at points remote from the absorption edge.

Anomalous scattering is present for all atomic types at all X-ray energies. However, the magnitudes of f' and f'' are negligible at X-ray energies far removed from the resonant frequencies of the atom. This includes all light atoms (H, C, N, O) of biological macromolecules at all X-ray energies commonly used for crystallography. f' and f'' are rather insensitive to scattering angle, unlike f^0, because the electronic resonant frequencies pertain to inner electron shells, which have radii much smaller than the X-ray wavelengths used for anomalous-scattering experiments. The magnitudes of f' and f'' are greatest at X-ray energies very near resonant frequencies, and are also highly energy-dependent (Fig. 14.2.1.1). This property of anomalous scattering is exploited in MAD.

Three means are available for evaluating anomalous scattering factors, f' and f''. Calculations from first principles on isolated elemental atoms are accurate for energies remote from resonant frequencies (Cromer & Liberman, 1970*a*,*b*). However, these calculated values do not apply to the energies most critical in a MAD experiment. f' and f'' can also be estimated by fitting to diffraction data measured at different energies (Templeton *et al.*, 1982). Finally, f'' can be obtained from X-ray absorption spectra by the equation above, and f' from f'' by the Kramers–Kronig transform [equation (14.2.1.3); Hendrickson *et al.*, 1988; Smith, 1998]. Both the precise position of a resonant frequency and the values of f' and f'' near resonance generally depend on transitions to unoccupied molecular orbitals, and are quite sensitive to the electronic environment surrounding the atom. Complexities in the X-ray absorption edge, particularly so-called 'white lines', can enhance the anomalous scattering considerably (Fig. 14.2.1.1). Thus, experimental measurements are needed to select wavelengths for optimal signals, and the values of f' and f'' should be determined either from an absorption spectrum or by refinement against the diffraction data.

X-ray spectra near absorption edges of anomalous scatterers depend on the orientation of the local chemical environment in the X-ray beam, which is polarized for synchrotron radiation. The anisotropy of anomalous scattering may affect both the edge position and the magnitude of absorption. In such cases, f' and f'' for individual atoms are also dependent on orientation. Orientational averaging due to multiple anomalous scatterer sites or crystallographic symmetry may prevent macroscopic detection of polarization effects in crystals. A formalism to describe anisotropic anomalous scattering in which f' and f'' are tensors has been

15.1.3. Reciprocal-space interpretation of density modification

Density modification, although mostly performed in real space for ease of application, can be understood in terms of reciprocal-space constraints on structure-factor amplitudes and phases.

Main & Rossmann (1966) showed that the NCS-averaging operation in real space can be expressed in reciprocal space as the convolution of the structure factors and the Fourier transform of the molecular envelope and the NCS matrices. Similarly, the solvent-flattening operation can be considered a multiplication of the map by some mask, $g_{\rm sf}(\mathbf{x})$, where $g_{\rm sf}(\mathbf{x}) = 1$ in the protein region and $g_{\rm sf}(\mathbf{x}) = 0$ in the solvent region. Thus

$$\rho_{\rm mod}(\mathbf{x}) = g_{\rm sf}(\mathbf{x}) \times \rho(\mathbf{x}). \qquad (15.1.3.1)$$

This assumes that the solvent level is zero, which can be achieved by suitable adjustment of the $F(000)$ term.

If we transform this equation to reciprocal space, then the product becomes a convolution; thus

$$F_{\rm mod}(\mathbf{h}) = (1/V)\sum_{\mathbf{k}} G_{\rm sf}(\mathbf{k})F(\mathbf{h}-\mathbf{k}), \qquad (15.1.3.2)$$

where $G_{\rm sf}(\mathbf{k})$ is the Fourier transform of the mask $g_{\rm sf}(\mathbf{x})$. The solvent mask $g_{\rm sf}(\mathbf{x})$ shows the outline of the molecule with no internal detail, so must be a low-resolution image. Therefore, all but the lowest-resolution terms of $G_{\rm sf}$ will be negligible.

The convolution expresses the relationship between phases in reciprocal space from the constraint of solvent flatness in real space. Since only the terms near the origin of $G_{\rm sf}$ are nonzero, the convolution can only relate phases that are local to each other in reciprocal space. Thus, it can only provide phase information for structure factors near the current phasing resolution limit.

This reasoning may also be applied to other density modifications. Histogram matching applies a nonlinear rescaling to the current density in the protein region. The equivalent multiplier, $g_{\rm hm}(\mathbf{x})$, shows variations of about 1.0 that are related to the features in the initial map. The function $G_{\rm hm}(\mathbf{h})$ for histogram matching is, therefore, dominated by its origin term, but shows significant features to the same resolution as the current map or further, as the density rescaling becomes more nonlinear. Histogram matching can therefore give phase indications to twice the resolution of the initial map or beyond, although phase indications will be weak and contain errors related to the level of error in the initial map.

$$\rho_{\rm mod}(\mathbf{x}) = g_{\rm ncs}(\mathbf{x})(1/N_{\rm ncs})\sum_i \rho_i(\mathbf{x}). \qquad (15.1.3.3)$$

Averaging may be described as the summation of a number of reoriented copies of the electron density within the region of the averaging mask (Main & Rossmann, 1966), *i.e.* where $\rho_i(\mathbf{x})$ is the initial density, $\rho(\mathbf{x})$, transformed by the ith NCS operator and $g_{\rm ncs}(\mathbf{x})$ is the mask of the molecule to be averaged. This summation is repeated for each copy of the molecule in the whole unit cell. The reciprocal-space averaging function, $G_{\rm ncs}(\mathbf{h})$, is the Fourier transform of a mask, as for solvent flattening, but since the mask covers only a single molecule, rather than the molecular density in the whole unit cell, the extent of $G_{\rm ncs}(\mathbf{h})$ in reciprocal space is greater.

Sayre's equation is already expressed as a convolution, although in this case the function $G(\mathbf{h})$ is given by the structure factors $F(\mathbf{h})$ themselves. It is, therefore, the most powerful method for phase extension. However, as resolution decreases, more of the reflections required to form the convolution are missing, and the error increases.

The functions $g(\mathbf{x})$ and $G(\mathbf{h})$ for these density modifications are illustrated in Fig. 15.1.3.1 for a simple one-dimensional structure.

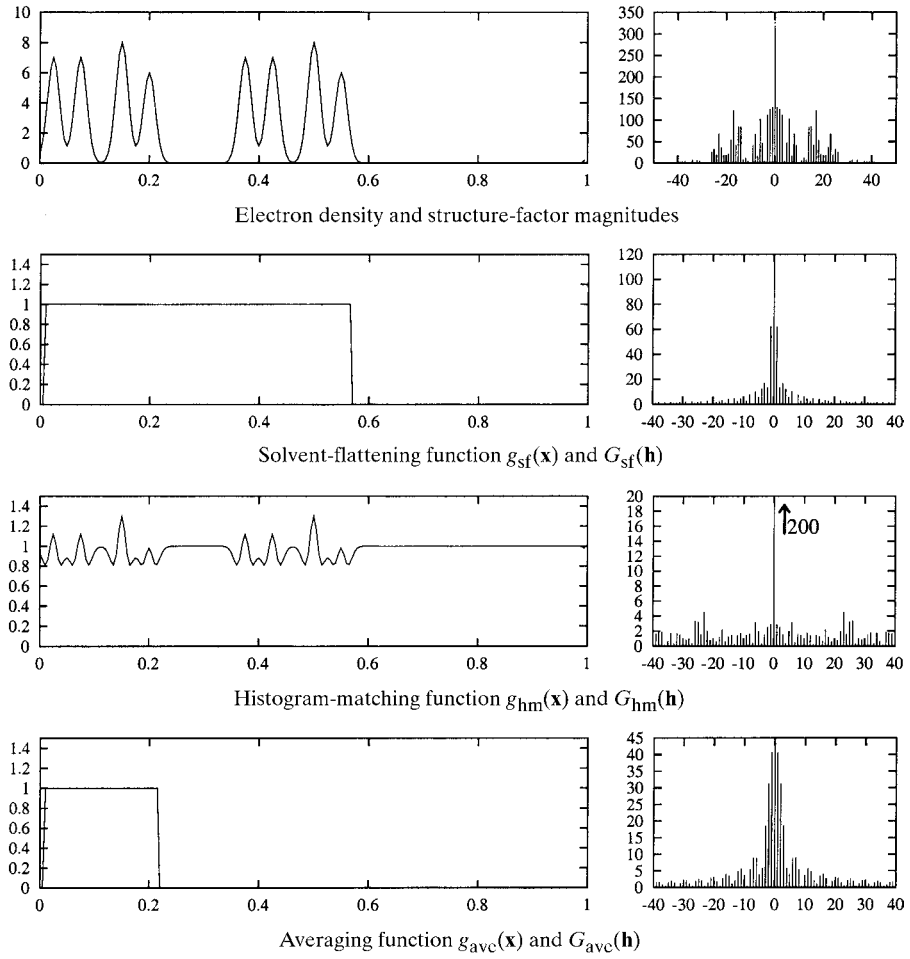


Fig. 15.1.3.1. The functions $g(\mathbf{x})$ and $G(\mathbf{h})$ for solvent flattening, histogram matching and averaging.

15.1.4. Phase combination

Phase combination is used to filter the noise in the modified phases and eliminate the incorrect component of the modified phases through a statistical process. The observed structure-factor amplitudes are used to estimate the reliability of the phases after density modification. The estimated probability of the modified phases is combined with the probability of observed phases to produce a more reliable phase estimate,

$$P_{\rm new}[\varphi(\mathbf{h})] = P_{\rm obs}[\varphi(\mathbf{h})]P_{\rm mod}[\varphi(\mathbf{h})]. \qquad (15.1.4.1)$$

Once a modified map has been obtained, modified phases and amplitudes may be derived from an inverse Fourier transform. The modified phases are normally combined with the initial phases by multiplication of their probability distributions. The probability distribution for the experimentally observed phases is usually de-

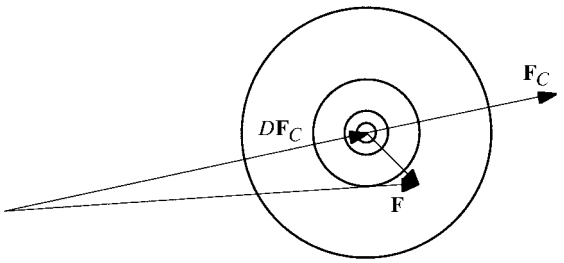


Fig. 15.2.3.2. Schematic illustration of the general structure-factor distribution, relevant in the case of any set of independent random errors in the atomic model.

in the acentric case, where $\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$, ε is the expected intensity factor and Σ_P is the Wilson distribution parameter for the model.

For centric reflections, the scattering differences are distributed along a line, so the probability distribution is a one-dimensional Gaussian.

$$p(\mathbf{F};\, \mathbf{F}_C) = [1/(2\pi\varepsilon\sigma_\Delta^2)^{1/2}] \exp\Big(-|\mathbf{F} - D\mathbf{F}_C|^2/2\varepsilon\sigma_\Delta^2\Big).$$

15.2.3.4. *Estimating* σ_A

Srinivasan (1966) showed that the Sim and Luzzati distributions could be combined into a single distribution that had a particularly elegant form when expressed in terms of normalized structure factors, or E values. This functional form still applies to the general distribution that reflects a variety of sources of error; the only difference is the interpretation placed on the parameters (Read, 1990). If $\mathbf{F}$ and $\mathbf{F}_C$ are replaced by the corresponding E values, a parameter σ_A plays the role of D, and σ_Δ^2 reduces to $(1 - \sigma_A^2)$. [The parameter σ_A is equivalent to D after correction for model completeness; $\sigma_A = D(\Sigma_P/\Sigma_N)^{1/2}$.] When the structure factors are normalized, overall scale and B-factor effects are also eliminated. The parameter σ_A that characterizes this probability distribution varies as a function of resolution. It must be deduced from the amplitudes $|\mathbf{F}_O|$ and $|\mathbf{F}_C|$, since the phase (thus the phase difference) is unknown.

A general approach to estimating parameters for probability distributions is to maximize a likelihood function. The likelihood function is the overall joint probability of making the entire set of observations, which is a function of the desired parameters. The parameters that maximize the probability of making the set of observations are the most consistent with the data. The idea of using maximum likelihood to estimate model phase errors was introduced by Lunin & Urzhumtsev (1984), who gave a treatment that was valid for space group $P1$. In a more general treatment that applies to higher-symmetry space groups, allowance is made for the statistical effects of crystal symmetry (centric zones and differing expected intensity factors) (Read, 1986).

The σ_A values are estimated by maximizing the joint probability of making the set of observations of $|\mathbf{F}_O|$. If the structure factors are all assumed to be independent, the joint probability distribution is the product of all the individual distributions. The assumption of independence is not completely justified in theory, but the results are fairly accurate in practice.

$$L = \prod_{\mathbf{h}} p(|\mathbf{F}_O|;\, |\mathbf{F}_C|).$$

The required probability distribution, $p(|\mathbf{F}_O|;\, |\mathbf{F}_C|)$, is derived from $p(\mathbf{F};\, \mathbf{F}_C)$ by integrating over all possible phase differences and neglecting the errors in $|\mathbf{F}_O|$ as a measure of $|\mathbf{F}|$. The form of this distribution, which is given in other publications (Read, 1986, 1990), differs for centric and acentric reflections. (It is important to note that although the distributions for structure factors are Gaussian, the distributions for amplitudes obtained by integrating out the phase are not.) It is more convenient to deal with a sum than a product, so the log likelihood function is maximized instead. In the program *SIGMAA*, reciprocal space is divided into spherical shells, and a value of the parameter σ_A is refined for each resolution shell. Details of the algorithm are given elsewhere (Read, 1986).

The resolution shells must be thick enough to contain several hundred to a thousand reflections each, in order to provide σ_A estimates with a sufficiently small statistical error. A larger number of shells (fewer reflections per shell) can be used for refined structures, since estimates of σ_A become more precise as the true value approaches 1. If there are sufficient reflections per shell, the estimates will vary smoothly with resolution. As discussed below, the smooth variation with resolution can also be exploited through a restraint that allows σ_A values to be estimated from fewer reflections.

15.2.4. Figure-of-merit weighting for model phases

Blow & Crick (1959) and Sim (1959) showed that the electron-density map with the least r.m.s. error is calculated from centroid structure factors. This conclusion follows from Parseval's theorem, because the centroid structure factor (its probability-weighted average value or expected value) minimizes the r.m.s. error of the structure factor. Since the structure-factor distribution $p(\mathbf{F};\, \mathbf{F}_C)$ is symmetrical about $\mathbf{F}_C$, the expected value of $\mathbf{F}$ will have the same phase as $\mathbf{F}_C$, but the averaging around the phase circle will reduce its magnitude if there is any uncertainty in the phase value (Fig. 15.2.4.1). We treat the reduction in magnitude by applying a weighting factor called the figure of merit, m, which is equivalent to the expected value of the cosine of the phase error.

15.2.5. Map coefficients to reduce model bias

15.2.5.1. *Model bias in figure-of-merit weighted maps*

A figure-of-merit weighted map, calculated with coefficients $m|\mathbf{F}_O| \exp(i\alpha_C)$, has the least r.m.s. error from the true map. According to the normal statistical (minimum variance) criteria, then, it is the best map. However, such a map will suffer from model bias; if its purpose is to allow the detection and repair of errors in the model, this is a serious qualitative defect. Fortunately, it is possible to predict the systematic errors leading to model bias and to make some correction for them.

Main (1979) dealt with this problem in the case of a perfect partial structure. Since the relationships among structure factors are the same in the general case of a partial structure with various errors, once $D\mathbf{F}_C$ is substituted for $\mathbf{F}_C$, all that is required to apply Main's results more generally is a change of variables (Read, 1986, 1990).

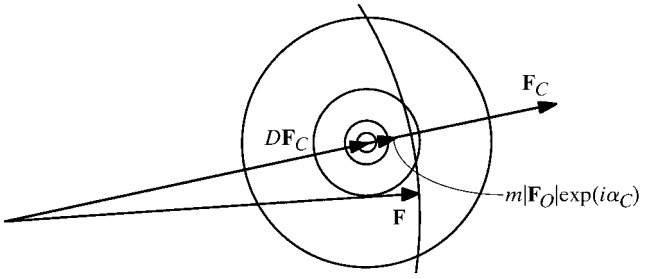


Fig. 15.2.4.1. Figure-of-merit weighted model-phased structure factor, obtained as the probability-weighted average over all possible phases.

abrupt increase in correct peaks occurs when Fourier refinement is started.

Since the correlation coefficient is a relatively absolute figure of merit (given atomic resolution, values greater than 65% almost invariably correspond to correct solutions), it is usually clear when *SHELXD* has solved a structure. The current version of *SHELXD* includes an option for calculating it using the full data every 10 or 20 internal loop cycles, and jumping to the external loop if the value is high enough. Recalculating it every cycle would be computationally less efficient overall.

16.1.8. Applying dual-space programs successfully

The solution of the (known) structure of triclinic lysozyme by *SHELXD* and shortly afterwards by *SnB* (Deacon *et al.*, 1998) finally broke the 1000-atom barrier for direct methods (there happen to be 1001 protein atoms in this structure!). Both programs have also solved a large number of previously unsolved structures that had defeated conventional direct methods; some examples are listed in Table 16.1.8.1. The overall quality of solutions is generally very good, especially if appropriate action is taken during the Fourier-

Table 16.1.8.1. *Some large structures solved by the Shake-and-Bake method*

Previously known test data sets are indicated by an asterisk (*). When two numbers are given in the resolution column, the second indicates the lowest resolution at which truncated data have yielded a solution. The program codes are *SnB* (S) and *SHELXD* (D).

(*a*) Full structures (>300 atoms).

Compound	Space group	N_u (molecule)	N_u + solvent	N_u (heavy)	Resolution (Å)	Program	Reference
Vancomycin	$P4_32_12$	202	258	8Cl	0.9–1.4	S	[1]
			312	6Cl	1.09	D	[2]
Actinomycin X2	$P1$	273	305	—	0.90	D	[3]
Actinomycin Z3	$P2_12_12_1$	186	307	2Cl	0.96	D	[4]
Actinomycin D	$P1$	270	314	—	0.94	D	[4]
Gramicidin A*	$P2_12_12_1$	272	317	—	0.86–1.1	S, D	[5]
DMSO d6 peptide	$P1$	320	326	—	1.20	S	[6]
Er-1 pheromone	$C2$	303	328	7S	1.00	S	[7]
Ristocetin A	$P2_1$	294	420	—	1.03	D	[8]
Crambin*	$P2_1$	327	423	6S	0.83–1.2	S, D	[9], [10]
Hirustasin	$P4_32_12$	402	467	10S	1.2–1.55	D	[11]
Cyclodextrin derivative	$P2_1$	448	467	—	0.88	D	[12]
Alpha-1 peptide	$P1$	408	471	Cl	0.92	S	[13]
Rubredoxin*	$P2_1$	395	497	Fe, 6S	1.0–1.1	S, D	[14]
Vancomycin	$P1$	404	547	12Cl	0.97	S	[15]
BPTI*	$P2_12_12_1$	453	561	7S	1.08	D	[16]
Cyclodextrin derivative	$P2_1$	504	562	28S	1.00	D	[17]
Balhimycin*	$P2_1$	408	598	8Cl	0.96	D	[18]
Mg-complex*	$P1$	576	608	8Mg	0.87	D	[19]
Scorpion toxin II*	$P2_12_12_1$	508	624	8S	0.96–1.2	S	[20]
Amylose-CA26	$P1$	624	771	—	1.10	D	[21]
Mersacidin	$P3_2$	750	826	24S	1.04	D	[22]
Cv HiPIP H42Q*	$P2_12_12_1$	631	837	4Fe	0.93	D	[23]
HEW lysozyme*	$P1$	1001	1295	10S	0.85	S, D	[24], [25]
rc-WT Cv HiPIP	$P2_12_12_1$	1264	1599	8Fe	1.20	D	[23]
Cytochrome c3	$P3_1$	2024	2208	8Fe	1.20	D	[26]

(*b*) Se substructures (> 25 Se) solved using peak-wavelength anomalous-difference data.

Protein	Space group	Molecular weight (kDa)	Se located	Se total	Resolution (Å)	Program	Reference
SAM decarboxylase	$P2_1$	77	20	26	2.25	S	[27]
AIR synthetase	$P2_12_12_1$	147	28	28	3.0	S	[28]
FTHFS	$R32$	200	28	28	2.5	D	[29]
AdoHcy hydrolase	$C222$	95	30	30	2.8–5.0	S	[30]
Epimerase	$P2_1$	370	64	70	3.0	S	[31]

References: [1] Loll *et al.* (1997); [2] Schäfer *et al.* (1996); [3] Schäfer (1998); [4] Schäfer, Sheldrick, Bahner & Lackner (1998); [5] Langs (1988); [6] Drouin (1998); [7] Anderson *et al.* (1996); [8] Schäfer & Prange (1998); [9] Stec *et al.* (1995); [10] Weeks *et al.* (1995); [11] Usón *et al.* (1999); [12] Aree *et al.* (1999); [13] Prive *et al.* (1999); [14] Dauter *et al.* (1992); [15] Loll *et al.* (1998); [16] Schneider (1998); [17] Reibenspiess (1998); [18] Schäfer, Sheldrick, Schneider & Vértesy (1998); [19] Teichert (1998); [20] Smith *et al.* (1997); [21] Gessler *et al.* (1999); [22] Schneider *et al.* (2000); [23] Parisini *et al.* (1999); [24] Deacon *et al.* (1998); [25] Walsh *et al.* (1998); [26] Frazão *et al.* (1999); [27] Ekstrom *et al.* (1999); [28] Li *et al.* (1999); [29] Radfar *et al.* (2000); [30] Turner *et al.* (1998); [31] Deacon & Ealick (1999).

16.2. The maximum-entropy method

BY G. BRICOGNE

16.2.1. Introduction

The modern concept of entropy originated in the field of statistical thermodynamics, in connection with the study of large material systems in which the number of internal degrees of freedom is much greater than the number of externally controllable degrees of freedom. This concept played a central role in the process of building a quantitative picture of the multiplicity of microscopic states compatible with given macroscopic constraints, as a measure of how much remains unknown about the detailed fine structure of a system when only macroscopic quantities attached to that system are known. The collection of all such microscopic states was introduced by Gibbs under the name 'ensemble', and he deduced his entire formalism for statistical mechanics from the single premise that the equilibrium picture of a material system under given macroscopic constraints is dominated by that configuration which can be realized with the greatest combinatorial multiplicity (*i.e.* which has maximum entropy) while obeying these constraints.

The notions of ensemble and the central role of entropy remained confined to statistical mechanics for some time, then were adopted in new fields in the late 1940s. Norbert Wiener studied Brownian motion, and subsequently time series of random events, by similar methods, considering in the latter an ensemble of messages, *i.e.* 'a repertory of possible messages, and over that repertory a measure determining the probability of these messages' (Wiener, 1949). At about the same time, Shannon created information theory and formulated his fundamental theorem relating the entropy of a source of random symbols to the capacity of the channel required to transmit the ensemble of messages generated by that source with an arbitrarily small error rate (Shannon & Weaver, 1949). Finally, Jaynes (1957, 1968, 1983) realized that the scope of the principle of maximum entropy could be extended far beyond the confines of statistical mechanics or communications engineering, and could provide the basis for a general theory (and philosophy) of statistical inference and 'data processing'.

The relevance of Jaynes' ideas to probabilistic direct methods was investigated by the author (Bricogne, 1984). It was shown that there is an intimate connection between the maximum-entropy method and an enhancement of the probabilistic techniques of conventional direct methods known as the 'saddlepoint method', some aspects of which have already been dealt with in Section 1.3.4.5.2 in Chapter 1.3 of *IT* B (Bricogne, 2001).

16.2.2. The maximum-entropy principle in a general context

16.2.2.1. *Sources of random symbols and the notion of source entropy*

Statistical communication theory uses as its basic modelling device a discrete source of random symbols, which at discrete times $t = 1, 2, \ldots$, randomly emits a 'symbol' taken out of a finite alphabet $\mathcal{A} = \{s_i | i = 1, \ldots, n\}$. Sequences of such randomly produced symbols are called 'messages'.

An important numerical quantity associated with such a discrete source is its *entropy per symbol* H, which gives a measure of the amount of uncertainty involved in the choice of a symbol. Suppose that successive symbols are independent and that symbol i has probability q_i. Then the general requirements that H should be a continuous function of the q_i, should increase with increasing uncertainty, and should be additive for independent sources of uncertainty, suffice to define H uniquely as

$$H(q_1, \ldots, q_n) = -k \sum_{i=1}^{n} q_i \log q_i, \qquad (16.2.2.1)$$

where k is an arbitrary positive constant [Shannon & Weaver (1949), Appendix 2] whose value depends on the unit of entropy chosen. In the following we use a unit such that $k = 1$.

These definitions may be extended to the case where the alphabet $\mathcal{A}$ is a continuous space endowed with a uniform measure μ: in this case the entropy per symbol is defined as

$$H(q) = -\int_{\mathcal{A}} q(\mathbf{s}) \log q(\mathbf{s}) \, \mathrm{d}\mu(\mathbf{s}), \qquad (16.2.2.2)$$

where q is the probability density of the distribution of symbols with respect to measure μ.

16.2.2.2. *The meaning of entropy: Shannon's theorems*

Two important theorems [Shannon & Weaver (1949), Appendix 3] provide a more intuitive grasp of the meaning and importance of entropy:

(1) H is approximately the logarithm of the reciprocal probability of a typical long message, divided by the number of symbols in the message; and

(2) H gives the rate of growth, with increasing message length, of the logarithm of the number of reasonably probable messages, regardless of the precise meaning given to the criterion of being 'reasonably probable'.

The entropy H of a source is thus a direct measure of the strength of the restrictions placed on the permissible messages by the distribution of probabilities over the symbols, lower entropy being synonymous with greater restrictions. In the two cases above, the maximum values of the entropy $H_{\max} = \log n$ and $H_{\max} = \log \mu(\mathcal{A})$ are reached when all the symbols are equally probable, *i.e.* when q is a uniform probability distribution over the symbols. When this distribution is not uniform, the usage of the different symbols is biased away from this maximum freedom, and the entropy of the source is lower; by Shannon's theorem (2), the number of 'reasonably probable' messages of a given length emanating from the source decreases accordingly.

The quantity that measures most directly the strength of the restrictions introduced by the non-uniformity of q is the difference $H(q) - H_{\max}$, since the proportion of N-atom random structures which remain 'reasonably probable' in the ensemble of the corresponding source is $\exp\{N[H(q) - H_{\max}]\}$. This difference may be written (using continuous rather than discrete distributions)

$$H(q) - H_{\max} = -\int_{\mathcal{A}} q(\mathbf{s}) \log[q(\mathbf{s})/m(\mathbf{s})] \, \mathrm{d}\mu(\mathbf{s}), \qquad (16.2.2.3)$$

where $m(\mathbf{s})$ is the uniform distribution which is such that $H(m) = H_{\max} = \log \mu(\mathcal{A})$.

16.2.2.3. *Jaynes' maximum-entropy principle*

From the fundamental theorems just stated, which may be recognized as Gibbs' argument in a different guise, Jaynes' own maximum-entropy argument proceeds with striking lucidity and constructive simplicity, along the following lines:

(1) experimental observation of, or 'data acquisition' on, a given system enables us to progress from an initial state of uncertainty to a state of lesser uncertainty about that system;

(2) uncertainty reflects the existence of numerous possibilities of accounting for the available data, viewed as constraints, in terms of a physical model of the internal degrees of freedom of the system;

film writer (Max, 1983) and Olson had used an early colour vector display from Evans and Sutherland to produce an eight-minute animation depicting the structure of tomato bushy stunt virus (Olson, 1981). By the early 1980s, animation projects became more ambitious. Olson produced large-screen OmniMax DNA and virus animation segments for Disney's EPCOT center in 1983. Max produced a red–blue stereo OmniMax film for Fujitsu entitled *We Are Born of Stars*, which included a continuous scene depicting the hierarchical packaging of DNA from atoms to chromosomes, based on the best current model of the time.

Computer-graphics animation has presented both great potential and significant challenges to the molecular scientist wishing to communicate the results of structural research. Animation can not only enhance the depiction of three-dimensional structure through motion stereopsis, it can show relationships through time, and demonstrate mechanism and change. The use of pans, zooms, cuts and other film techniques can effectively lead the viewer through a complex scene and focus attention on specific structures or processes. The vocabulary of film, video and animation is familiar to all, but can be a difficult language to master. While short animations showing simple rotations or transitions between molecular states, or dynamics trajectories, are now routinely made for video or web viewing, extended animations showing molecular structure and function in depth are still relatively rare. The time, tools and expertise that are required are not generally available to structural researchers.

17.2.4.3. *The return of physical models*

While the use of physical models of molecules has largely been replaced by computer graphics, new computer-driven rapid-prototyping technologies which originated in the manufacturing sector have begun to be utilized in the display of molecular structure. A number of 'three-dimensional printing' methods have been developed to build up a physical model directly from a computational surface representation of an object (Burns, 1954). One of the earliest methods, stereolithography, uses a resin which is polymerized when exposed to laser light of a given wavelength. The laser is passed through a vat of the liquid resin and is lowered, layer-by-layer as it plots out the shape of the object (Fig. 17.2.4.2). Other approaches build up layers of paper or plastic through lamination or deposition. These methods have been used by a number of scientists to produce various representations of molecular structure (Bailey *et al.*, 1998). The ability to hold an accurate representation of a molecular surface in one's hand and feel its shape can give great insight, not only to people with visual impairments, but to anyone. Moreover, when one is dealing with processes such as docking and assembly, these physical models can add a haptic and manipulative appreciation of the nature of the problem. While at this point colour has not been implemented in these technologies, there remains the promise that such automated production of molecular models will enhance the communication and appreciation of molecular structure.

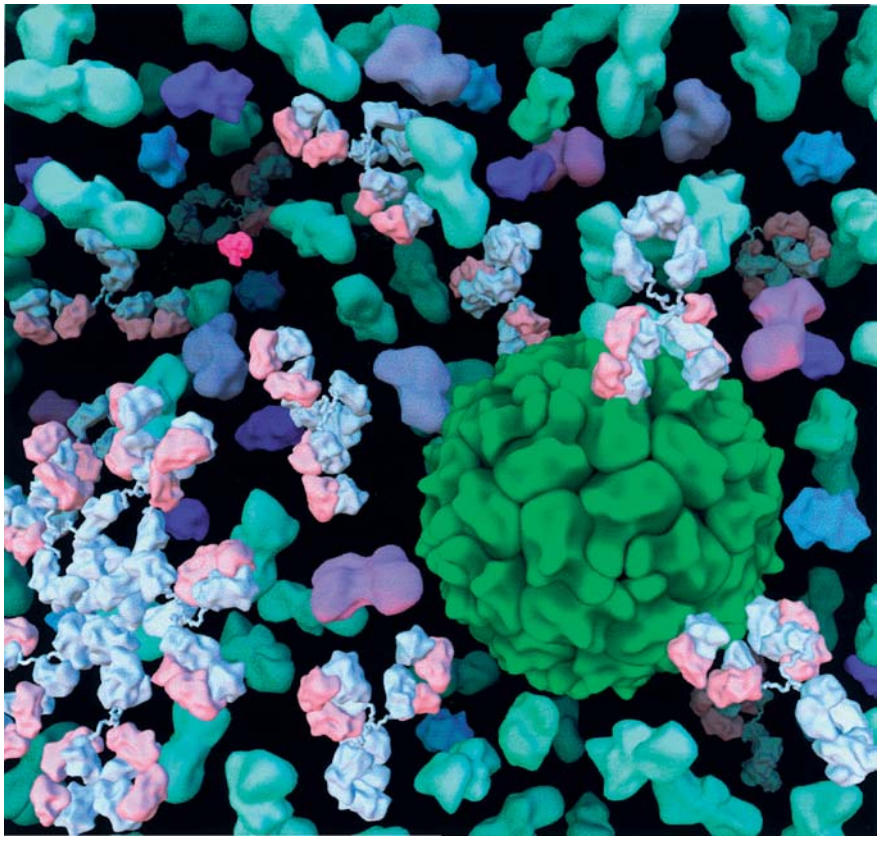


Fig. 17.2.5.1. This image represents a volume of blood plasma 750 Å on a side. Within the three-dimensional model, antibodies (Y- and T-shaped molecules in light blue and pink) are binding to a virus (the large green spherical assembly on the right), labelling it for destruction. It shows all macromolecules present in the blood plasma at a magnification of about 10 000 000 times. This model is composed of over 450 individual protein domains, ranging in size from the 60 protomers making up the poliovirus to a single tiny insulin molecule (in magenta). The model was constructed using atomic level descriptions for each molecule, for a total of roughly 1.5 million atoms. Detailed surfaces were computed for each type of protein using *MSMS* by Michel Sanner and then smoothed to a lower resolution using the *HARMONY* spherical-harmonic surfaces developed by Bruce Duncan. The model geometry contains over 1.5 million triangles.

17.2.5. Looking ahead

Moore's law has already delivered on the promise of three-dimensional graphics capability for the desktop and laptop. The internet and World Wide Web have made molecular structure data and display software available to the masses. Have molecular graphics reached a stage of maturity beyond which only small incremental changes will be made?

The Human Genome Initiative and high-throughput structure determination are beginning to change the scope of the questions asked of molecular modelling. Prediction of function, interactions, and large-scale assembly and mechanism will become the dominant domain of molecular graphics and modelling. These tasks will challenge the capabilities of the hardware, software and, particularly, the user interface. New modes of interacting with data and models are coming from the computer-graphics community. Molecular docking and protein manipulation using force-feedback devices have been demonstrated at the University of North Carolina (Brooks *et al.*, 1990). The same team has developed a 'nanomanipulator' which couples a scanning atomic force microscope with stereoscopic display and force-feedback manipulation to control and sense the positioning and interactions of the probe

commonly placed in the model in plausible positions according to molecular geometry, but this can be misleading to people using the coordinate set. If the atoms are included in the model, the atomic displacement parameters generally become very large, and this may be an acceptable flag for dynamic disorder. The hazard with this procedure is that including these atoms in the model provides additional parameters to conceal any error signal in the data that might relate to problems elsewhere in the model.

At high resolution, it is sometimes possible to model the correlated motion of atoms in rigid groups by a single tensor that describes translation, libration and screw. This is rarely done for macromolecules at present, but may be an extremely accurate way to model the behaviour of the molecules. The recent development of efficient anisotropic refinement methods for macromolecules by Murshudov *et al.* (1999) will undoubtedly produce a great deal more information about the modelling of dynamic disorder and anisotropy in macromolecular structures.

Macromolecular crystals contain between 30 and 70% solvent, mostly amorphous. The diffraction is not accurately modelled unless this solvent is included (Tronrud, 1997). The bulk solvent is generally modelled as a continuum of electron density with a high atomic displacement parameter. The high displacement parameter blurs the edges, so that the contribution of the bulk solvent to the scattering is primarily at low resolution. Nevertheless, it is important to include this in the model for two reasons. First, unless the bulk solvent is modelled, the low-resolution structure factors cannot be used in the refinement. This has the unfortunate effect of rendering the refinement of *all* of the atomic displacement parameters ill-determined. Second, omission or inaccurate phasing of the low-resolution reflections tends to produce long-wavelength variations in the electron-density maps, rendering them more difficult to interpret. In some regions, the maps can become overconnected, and in others they can become fragmented.

18.1.8. Optimization methods

Optimization methods for small molecules are straightforward, but macromolecules present special problems due to their sheer size. The large number of parameters vastly increases the volume of the parameter space that must be searched for feasible solutions and also increases the storage requirements for the optimization process. The combination of a large number of parameters and a large number of observations means that the computations at each cycle of the optimization process are expensive.

Optimization methods can be roughly classified according to the order of derivative information used in the algorithm. Methods that use no derivatives find an optimum through a search strategy; examples are Monte Carlo methods and some forms of simulated annealing. First-order methods compute gradients, and hence can always move in a direction that should reduce the objective function. Second-order methods compute curvature, which allows them to predict not only which direction will reduce the objective function, but how that direction will change as the optimization proceeds. The zero-order methods are generally very slow in high-dimensional spaces because the volume that must be searched becomes huge. First-order methods can be fast and compact, but cannot determine whether or not the solution is a true minimum. Second-order methods can detect null subspaces and singularities in the solution, but the computational cost grows as the cube of the number of parameters (or worse), and the storage requirements grow as the square of the number of parameters – undesirable properties where the number of parameters is of the order of 10^4.

Historically, the most successful optimization methods for macromolecular structures have been first-order methods. This is beginning to change as multi-gigabyte memories are becoming more common on computers and processor speeds are in the gigahertz range. At this time, there are no widely used refinement programs that run effectively on multiprocessor systems, although there are no theoretical barriers to writing such a program.

18.1.8.1. *Solving the refinement equations*

Methods for solving the refinement equations are described in *IT* C Chapters 8.1 to 8.5 and in many texts. Prince (1994) provides an excellent starting point. There are two commonly used approaches to finding the set of parameters that minimizes equation (18.1.4.1). The first is to treat each observation separately and rewrite each term of (18.1.4.1) as

$$w_i[y_i - f_i(\mathbf{x})] = w_i \sum_{j=1}^{N} \left(\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right)(x_j^0 - x_j), \qquad (18.1.8.1)$$

where the summation is over the N parameters of the model. This is simply the first-order expansion of $f_i(\mathbf{x})$ and expresses the hypothesis that the calculated values should match the observed values. The system of simultaneous *observational equations* can be solved for the parameter shifts provided that there are at least as many observations as there are parameters to be determined. When the number of observational equations exceeds the number of parameters, the least-squares solution is that which minimizes (18.1.4.1). This is the method generally used for refining small-molecule crystal structures, and increasingly for macromolecular structures at atomic resolution.

18.1.8.2. *Normal equations*

In matrix form, the observational equations are written as

$$\mathbf{A}\Delta = \mathbf{r},$$

where $\mathbf{A}$ is the M by N matrix of derivatives, Δ is the parameter shifts and $\mathbf{r}$ is the vector of residuals given on the left-hand sides of equation (18.1.8.1). The *normal equations* are formed by multiplying both sides of the equation by $\mathbf{A}^T$. This produces an N by N square system, the solution to which is the desired least-squares solution for the parameter shifts.

$$\mathbf{A}^T\mathbf{A}\Delta = \mathbf{A}^T\mathbf{r} \text{ or } \mathbf{M}\Delta = \mathbf{b},$$

$$m_{ij} = \sum_{k=1}^{M} w_k \left(\frac{\partial f_k(\mathbf{x})}{\partial x_i}\right)\left(\frac{\partial f_k(\mathbf{x})}{\partial x_j}\right),$$

$$b_i = \sum_{k=1}^{M} w_k[y_k - f_k(\mathbf{x})]\left(\frac{\partial f_k(\mathbf{x})}{\partial x_i}\right).$$

Similar equations are obtained by expanding (18.1.4.1) as a second-order Taylor series about the minimum $\mathbf{x}_0$ and differentiating.

$$\Phi(\mathbf{x} - \mathbf{x}_0) \approx \Phi(\mathbf{x}_0) + \left\langle \left(\frac{\partial \Phi}{\partial x_i}\right)_{\mathbf{x}_0} \middle| (\mathbf{x} - \mathbf{x}_0) \right\rangle + \frac{1}{2}\left\langle (\mathbf{x} - \mathbf{x}_0) \middle| \left(\frac{\partial^2 \Phi}{\partial x_i \partial x_j}\right)_{\mathbf{x}_0} \middle| (\mathbf{x} - \mathbf{x}_0) \right\rangle,$$

$$\left\langle \left(\frac{\partial \Phi}{\partial \mathbf{x}}\right) \right\rangle \approx \left\langle \left(\frac{\partial^2 \Phi}{\partial x_i \partial x_j}\right)_{\mathbf{x}_0} \middle| (\mathbf{x} - \mathbf{x}_0) \right\rangle.$$

The second-order approximation is equivalent to assuming that the matrix of second derivatives does not change and hence can be computed at $\mathbf{x}$ instead of at $\mathbf{x}_0$.

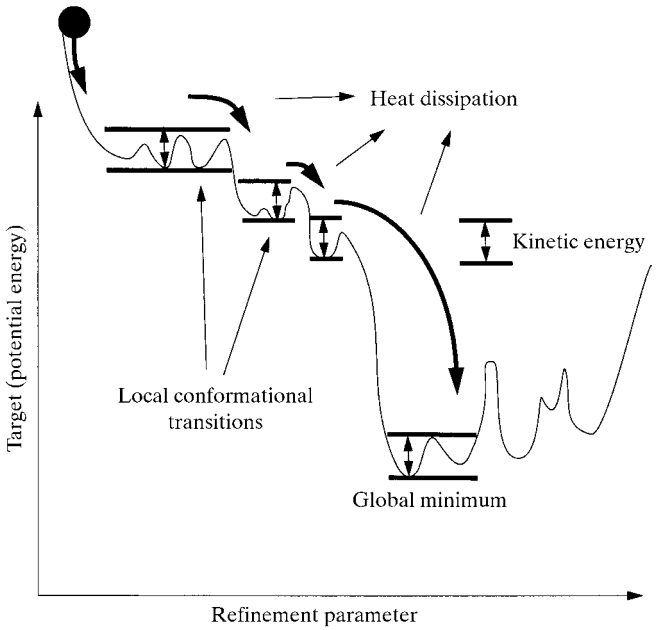


Fig. 18.2.4.1. Illustration of simulated annealing for minimization of a one-dimensional function. The kinetic energy of the system (a 'ball' rolling on the one-dimensional surface) allows local conformational transitions with barriers smaller than the kinetic energy. If a larger drop in energy is encountered, the excess kinetic energy is dissipated. It is thus unlikely that the system can climb out of the global minimum once it has reached it.

The simulated-annealing algorithm requires a mechanism to create a Boltzmann distribution at a given temperature, T, and an annealing schedule, that is, a sequence of temperatures $T_1 \geq T_2 \geq \ldots \geq T_l$ at which the Boltzmann distribution is computed. Implementations differ in the way they generate a transition, or move, from one set of parameters to another that is consistent with the Boltzmann distribution at a given temperature. The two most widely used methods are Metropolis Monte Carlo (Metropolis *et al.*, 1953) and molecular dynamics (Verlet, 1967) simulations. For X-ray crystallographic refinement, molecular dynamics has proven extremely successful (Brünger *et al.*, 1987) because it limits the search to physically reasonable 'moves'.

18.2.4.1. *Molecular dynamics*

A suitably chosen set of atomic parameters can be viewed as generalized coordinates that are propagated in time by the classical equations of motion (Goldstein, 1980). If the generalized coordinates represent the x, y, z positions of the atoms of a molecule, the classical equations of motion reduce to the familiar Newton's second law:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = -\nabla_i E. \qquad (18.2.4.1)$$

The quantities m_i and $\mathbf{r}_i$ are, respectively, the mass and coordinates of atom i, and E is given by equation (18.2.3.1). The solution of the partial differential equations (18.2.4.1) can be achieved numerically using finite-difference methods (Verlet, 1967; Abramowitz & Stegun, 1968). This approach is referred to as molecular dynamics.

Initial velocities for the integration of equation (18.2.4.1) are usually assigned randomly from a Maxwell distribution at the appropriate temperature. Assignment of different initial velocities will generally produce a somewhat different structure after simulated annealing. By performing several refinements with different initial velocities, one can therefore improve the chances of success of simulated-annealing refinement. Furthermore, this improved sampling can be used to study discrete disorder and conformational variability, especially when using torsion-angle molecular dynamics (see below).

Although Cartesian (*i.e.* flexible bond lengths and bond angles) molecular dynamics places restraints on bond lengths and bond angles [through $E_{\rm chem}$, equation (18.2.3.1)], one might want to implement these restrictions as constraints, *i.e.*, fixed bond lengths and bond angles (Diamond, 1971). This is supported by the observation that the deviations from ideal bond lengths and bond angles are usually small in macromolecular X-ray crystal structures. Indeed, fixed-length constraints have been applied to crystallographic refinement by least-squares minimization (Diamond, 1971). It is only recently, however, that efficient and robust algorithms have become available for molecular dynamics in torsion-angle space (Bae & Haug, 1987, 1988; Jain *et al.*, 1993; Rice & Brünger, 1994). We chose an approach that retains the Cartesian-coordinate formulation of the target function and its derivatives with respect to atomic coordinates, so that the calculation remains relatively straightforward and can be applied to any macromolecule or their complexes (Rice & Brünger, 1994). In this formulation, the expression for the acceleration becomes a function of positions and velocities. Iterative equations of motion for constrained dynamics in this formulation can be derived and solved by finite-difference methods (Abramowitz & Stegun, 1968). This method is numerically very robust and has a significantly increased radius of convergence in crystallographic refinement compared to Cartesian molecular dynamics (Rice & Brünger, 1994).

18.2.4.2. *Temperature control*

Simulated annealing requires the control of the temperature during molecular dynamics. The current temperature of the simulation ($T_{\rm curr}$) is computed from the kinetic energy

$$E_{\rm kin} = \sum_i^n \tfrac{1}{2} m_i \left(\frac{\partial r_i}{\partial t}\right)^2 \qquad (18.2.4.2)$$

of the molecular-dynamics simulation,

$$T_{\rm curr} = 2E_{\rm kin}/3nk_B. \qquad (18.2.4.3)$$

Here, n is the number of atoms, m_i is the mass of the atom and k_B is Boltzmann's constant. One commonly used approach to control the temperature of the simulation consists of coupling the equations of motion to a heat bath through a 'friction' term (Berendsen *et al.*, 1984). Another approach is to rescale periodically the velocities in order to match $T_{\rm curr}$ with the target temperature.

18.2.4.3. *Annealing schedules*

The simulated-annealing temperature needs to be high enough to allow conformational transitions, but not so high that the model moves too far away from the correct structure. The optimal temperature for a given starting structure is a matter of trial and error. Starting temperatures that work for the average case have been determined for a variety of simulated-annealing protocols (Brünger, 1988; Adams *et al.*, 1997). However, it might be worth trying a different temperature if a particularly difficult refinement problem is encountered. In particular, significantly higher temperatures are attainable using torsion-angle molecular dynamics. Note that each simulated-annealing refinement is subject to 'chance' by using a random-number generator to generate the initial velocities. Thus, multiple simulated annealing runs can be carried out in order to increase the success rate of the refinement. The best structure(s) (as determined by the free R value) among a set of refinements using

Table 18.3.2.1. *Bond lengths of standard amino-acid side chains*

EH denotes the values of Engh & Huber (1991), which were clustered according to atom type. The EH99 values are taken from recent Cambridge Structural Database releases with clustering of parameters only in the choice of fragments, based on amino acids. Parameters marked with an asterisk involving CA—CB bonds were taken from peptide fragment geometries. Two asterisks mark long-chain aliphatic parameters taken from arginine statistics. The number of fragments and the number of structures containing these fragments are noted after the amino-acid name. The fragments used for generating the statistics are described after the amino-acid name: incomplete valences indicate unspecified substituents with, however, specified orbital hybridization.

Alanine, 163/268, CO—NH—CH(CH_3)—CO—NH

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.521	0.033	1.520	0.021

Arginine, 71/98, CH—$(CH_2)_3$—NH—$C(NH_2)_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.521	0.027
CG—CD	1.520	0.030	1.515	0.025
CD—NE	1.460	0.018	1.460	0.017
NE—CZ	1.329	0.014	1.326	0.013
CZ—NH(1,2)	1.326	0.018	1.326	0.013

Asparagine, 145/247, —C—CH_2—CO—NH_2

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.527	0.026
CB—CG	1.516	0.025	1.506	0.023
CG—OD1	1.231	0.020	1.235	0.022
CG—ND2	1.328	0.021	1.324	0.025

Aspartate, 265/404, C—CH_2—CO_2

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.516	0.025	1.513	0.021
CG—OD(1,2)	1.249	0.019	1.249	0.023

Cysteine, 10/17, N—CH(CO)—CH_2—SH

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.526	0.013
CB—SG	1.808	0.033	1.812	0.016

Disulfides, 53/68, C—CH_2—S—S—CH_2—C

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—SG	1.808	0.033	1.818	0.017
SG—SG	2.030	0.008	2.033	0.016

Glutamate, 74/88, C—CH_2—CH_2—CO_2

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.517	0.019
CG—CD	1.516	0.025	1.515	0.015
CD—OE(1,2)	1.249	0.019	1.252	0.011

Glutamine, 145/247, —C—CH_2—CO—NH_2

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.521**	0.027**
CG—CD	1.516	0.025	1.506	0.023
CD—OE1	1.231	0.020	1.235	0.022
CD—NE2	1.328	0.021	1.324	0.025

Glycine: see peptide parameters, Table 18.3.2.3

Histidine (HISE), 35/37, C—CH_2—imidazole; NE protonated

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.497	0.014	1.496	0.018
CG—ND1	1.371	0.017	1.383	0.022
CG—CD2	1.356	0.011	1.353	0.014
ND1—CE1	1.319	0.013	1.323	0.015
CD2—NE2	1.374	0.021	1.375	0.022
CE1—NE2	1.345	0.020	1.333	0.019

Histidine (HISD), 10/12, C—CH_2—imidazole; ND protonated

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.497	0.014	1.492	0.016
CG—ND1	1.378	0.011	1.369	0.015
CG—CD2	1.356	0.011	1.353	0.017
ND1—CE1	1.345	0.020	1.343	0.025
CD2—NE2	1.382	0.030	1.415	0.021
CE1—NE2	1.319	0.013	1.322	0.023

Histidine (HISH), 50/54, C—CH_2—imidazole; NE, ND protonated

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.497	0.014	1.492	0.010
CG—ND1	1.378	0.011	1.380	0.010
CG—CD2	1.354	0.011	1.354	0.009
ND1—CE1	1.321	0.010	1.326	0.010
CD2—NE2	1.374	0.011	1.373	0.011
CE1—NE2	1.321	0.010	1.317	0.011

Isoleucine, 54/80, NH—CH(CO)—CH(CH_3)—CH_2—CH_3

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.540	0.027	1.544	0.023
CB—CG1	1.530	0.020	1.536	0.028
CB—CG2	1.521	0.033	1.524	0.031
CG1—CD1	1.513	0.039	1.500	0.069*

18.5. Coordinate uncertainty

BY D. W. J. CRUICKSHANK

18.5.1. Introduction

18.5.1.1. *Background*

Even in 1967 when the first few protein structures had been solved, it would have been hard to imagine a time when the best protein structures would be determined with a precision approaching that of small molecules. That time was reached during the 1990s. Consequently, the methods for the assessment of the precision of small molecules can be extended to good-quality protein structures.

The key idea is simply stated. At the conclusion and full convergence of a least-squares or equivalent refinement, *the estimated variances and covariances of the parameters may be obtained through the inversion of the least-squares full matrix.*

The inversion of the full matrix for a large protein is a gigantic computational task, but it is being accomplished in a rising number of cases. Alternatively, approximations may be sought. Often these can be no more than rough order-of-magnitude estimates. Some of these approximations are considered below.

Caveat. Quite apart from their large numbers of atoms, protein structures show features differing from those of well ordered small-molecule structures. Protein crystals contain large amounts of solvent, much of it not well ordered. Parts of the protein chain may be floppy or disordered. All natural protein crystals are noncentrosymmetric, hence the simplifications of error assessment for centrosymmetric structures are inapplicable. The effects of incomplete modelling of disorder on phase angles, and thus on parameter errors, are not addressed explicitly in the following analysis. Nor does this analysis address the quite different problem of possible gross errors or misplacements in a structure, other than by their indication through high B values or high coordinate standard uncertainties. These various difficulties are, of course, reflected in the values of $\Delta|F|$ used in the precision estimates.

On the problems of structure validation see Part 21 of this volume and Dodson (1998).

Some structure determinations do make a first-order correction for the effects of disordered solvent on phase angles by application of Babinet's principle of complementarity (Langridge *et al.*, 1960; Moews & Kretsinger, 1975; Tronrud, 1997). Babinet's principle follows from the fact that if $\rho(\mathbf{x})$ is constant throughout the cell, then $F(\mathbf{h}) = 0$, except for $F(\mathbf{0})$. Consequently, if the cell is divided into two regions C and D, $F_C(\mathbf{h}) = -F_D(\mathbf{h})$. Thus if D is a region of disordered solvent, $F_D(\mathbf{h})$ can be estimated from $-F_C(\mathbf{h})$. A first approximation to a disordered model may be obtained by placing negative point-atoms with very high Debye B values at all the ordered sites in region C. This procedure provides some correction for very low resolution planes. Alternatively, corrections are sometimes made by a mask bulk solvent model (Jiang & Brünger, 1994).

The application of restraints in protein refinement does not affect the key idea about the method of error estimation. A simple model for restrained refinement is analysed in Section 18.5.3, and the effect of restraints is discussed in Section 18.5.4 and later.

Much of the material in this chapter is drawn from a Topical Review published in *Acta Crystallographica*, Section D (Cruickshank, 1999).

Protein structures exhibiting noncrystallographic symmetry are not considered in this chapter.

18.5.1.2. *Accuracy and precision*

A distinction should be made between the terms *accuracy* and *precision.* A single measurement of the magnitude of a quantity differs by error from its unknown true value λ. In statistical theory (Cruickshank, 1959), the fundamental supposition made about errors is that, for a given experimental procedure, the possible results of an experiment define the probability density function $f(x)$ of a *random variable.* Both the true value λ and the probability density $f(x)$ are unknown. The problem of assessing the accuracy of a measurement is thus the double problem of estimating $f(x)$ and of assuming a relation between $f(x)$ and λ.

Precision relates to the function $f(x)$ and its spread.

The problem of what relationship to assume between $f(x)$ and the true value λ is more subtle, involving particularly the question of *systematic errors.* The usual procedure, after correcting for known systematic errors, is to suppose that some typical property of $f(x)$, often the mean, is the value of λ. No repetition of the same experiment will ever reveal the systematic errors, so statistical estimates of precision take into account only random errors. Empirically, systematic errors can be detected only by remeasuring the quantity with a different technique.

Care is needed in reading older papers. The word accuracy was sometimes intended to cover both random and systematic errors, or it may cover only random errors in the above sense of precision (known systematic errors having been corrected).

In recent years, the well established term *estimated standard deviation* (e.s.d.) has been replaced by the term *standard uncertainty* (s.u.). (See Section 18.5.2.3 on statistical descriptors.)

18.5.1.3. *Effect of atomic displacement parameters (or 'temperature factors')*

It is useful to begin with a reminder that the Debye $B = 8\pi^2\langle u^2\rangle$, where u is the atomic displacement parameter. If $B = 80$ Å^2, the r.m.s. amplitude is 1.01 Å. The centroid of an atom with such a B is unlikely to be precisely determined. For $B = 40$ Å^2, the 0.71 Å r.m.s. amplitude of an atom is approximately half a C—N bond length. For $B = 20$ Å^2, the amplitude is 0.50 Å. Even for $B = 5$ Å^2, the amplitude is 0.25 Å. The size of the atomic displacement amplitudes should always be borne in mind when considering the precision of the position of the centroid of an atom.

Scattering power depends on $\exp[-2B(\sin\theta/\lambda)^2] = \exp[-B/(2d^2)]$. For $B = 20$ Å^2 and $d = 4$, 2 or 1 Å, this factor is 0.54, 0.08 or 0.0001. For $d = 2$ Å and $B = 5$, 20 or 80 Å^2, the factor is again 0.54, 0.08 or 0.0001. The scattering power of an atom thus depends very strongly on B and on the resolution $d = 1/s = \lambda/2\sin\theta$. Scattering at high resolution (low d) is dominated by atoms with low B.

An immediate consequence of the strong dependence of scattering power on B is that the standard uncertainties of atomic coordinates also depend very strongly on B, especially between atoms of different B within the same structure.

[An IUCr Subcommittee on Atomic Displacement Parameter Nomenclature (Trueblood *et al.*, 1996) has recommended that the phrase 'temperature factor', though widely used in the past, should be avoided on account of several ambiguities in its meaning and usage. The Subcommittee also discourages the use of B and the anisotropic tensor $\mathbf{B}$ in favour of $\langle u^2\rangle$ and $\mathbf{U}$, on the grounds that the latter have a more direct physical significance. The present author concurs (Cruickshank, 1956, 1965). However, as the use of B or B_{eq} is currently so widespread in biomolecular crystallography, this chapter has been written in terms of B.]

19.2.4. Data processing

19.2.4.1. *Data sampling*

The principle of three-dimensional reconstruction is based on the central section theorem, which states that the experimental or computed projected diffraction pattern of a three-dimensional object is a plane that intersects the centre of the three-dimensional Fourier space in the direction normal to the direction of the projection (DeRosier & Klug, 1968). Because of the crystallographic symmetry inherent in a protein crystal, only a portion of the entire three-dimensional Fourier space, equivalent to an asymmetric unit of the crystal unit cell, is needed for the reconstruction. The structure factors of a three-dimensional crystal are localized in the three-dimensional reciprocal lattice, whereas the structure factors of a two-dimensional crystal are distributed continuously along the lattice lines, each of which passes through the reciprocal lattice in the zero projection plane (Fig. 19.2.4.1) (Henderson & Unwin, 1975). The assignment of z^* for each observation (h, k, z^*) along the lattice line is determined from the tilt angle and direction of the tilt axis for each image (Shaw & Hills, 1981). In general, the three-dimensional data set is initially built up from low-angle data and is gradually extended to the high-angle data. The angular parameters for each observed reflection are iteratively refined among one another within the whole data set. The

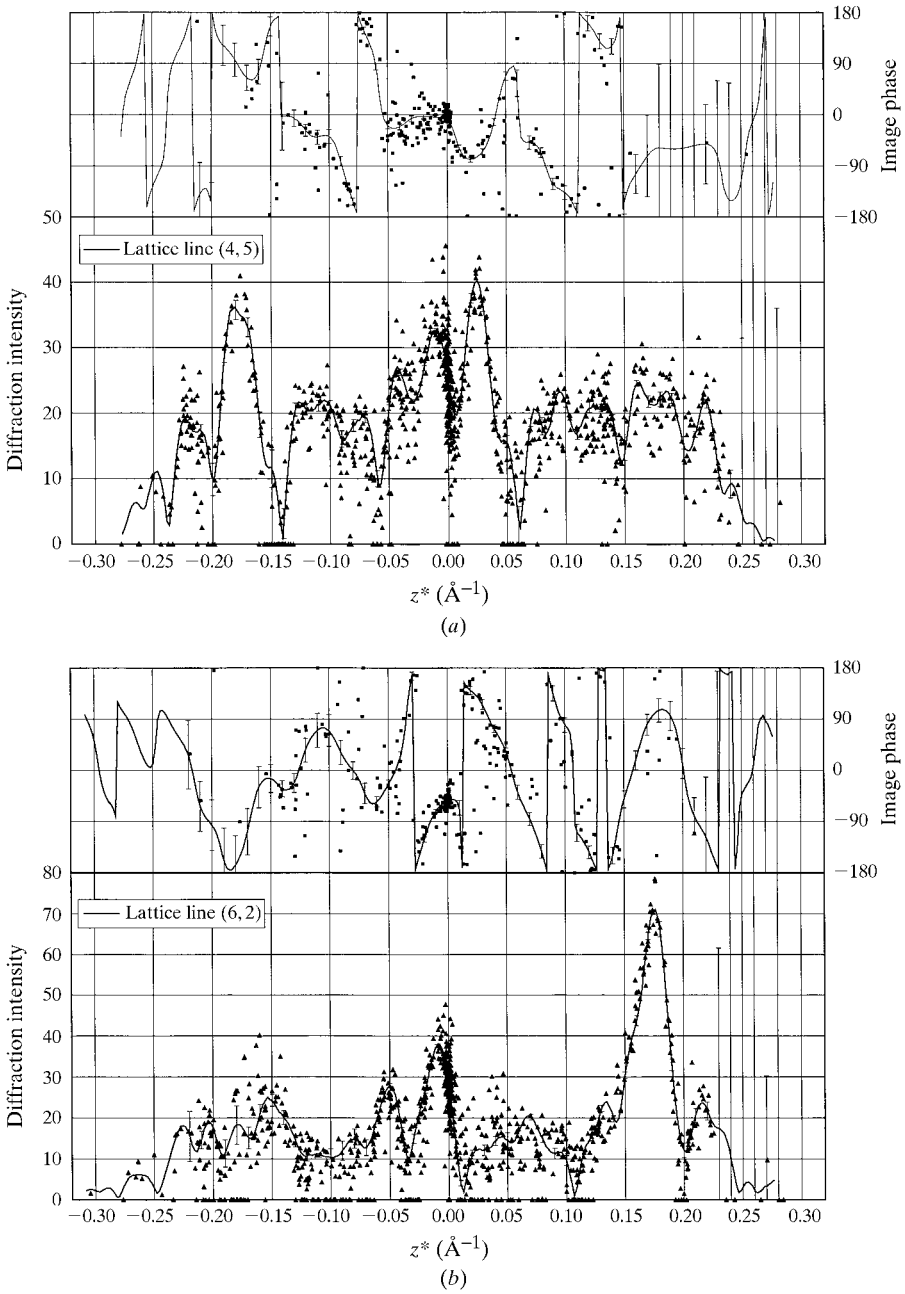


Fig. 19.2.4.2. Experimental intensities from electron diffraction patterns and phases from images of bacteriorhodopsin, recorded from tilted crystals in an electron cryomicroscope. Fitted curves for two representative lattice lines are shown: (*a*) (4, 5, z^*) and (*b*) (6, 2, z^*) (Courtesy of Drs Terushisa Hirai and Yoshinori Fujiyoshi at Kyoto University.)

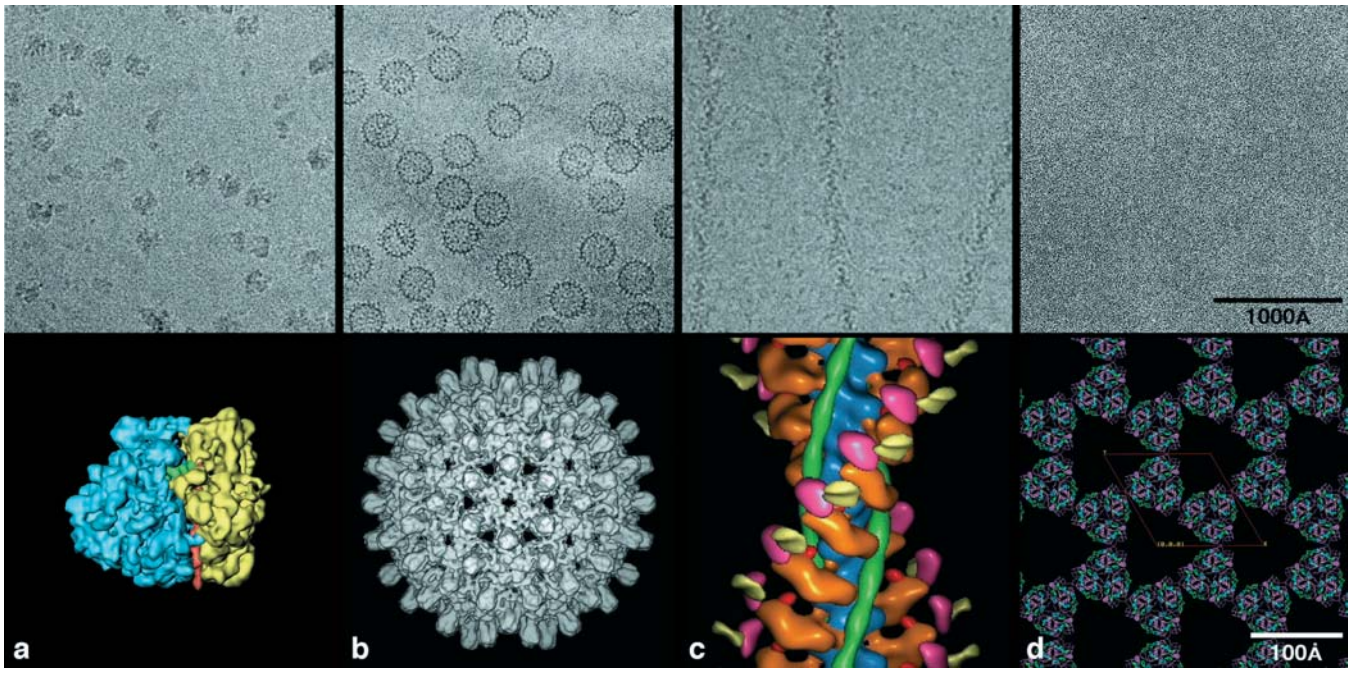


Fig. 19.6.5.1. Examples of macromolecules studied by cryo EM and 3D image reconstruction and the resulting 3D structures (bottom row) after cryo EM analysis. All micrographs (top row) are displayed at ∼170 000× magnification and all models at ∼1 200 000× magnification. (*a*) A single particle without symmetry. The micrograph shows 70S *E. coli* ribosomes complexed with mRNA and fMet-tRNA. The surface-shaded density map, made by averaging 73 000 ribosome images from 287 micrographs, has a resolution of 11.5 Å. The 50S and 30S subunits and the tRNA are coloured blue, yellow and green, respectively. The identity of many of the protein and RNA components is known and some RNA double helices are clearly recognizable by their major and minor grooves (*e.g.* helix 44 is shown in red). Courtesy of J. Frank (SUNY, Albany), using unpublished data from I. Gabashvili, R. Agrawal, C. Spahn, R. Grassucci, J. Frank & P. Penczek. (*b*) A single particle with symmetry. The micrograph shows hepatitis B virus cores. The 3D reconstruction, at a resolution of 7.4 Å, was computed from 6384 particle images taken from 34 micrographs. From Böttcher, Wynne & Crowther (1997). (*c*) A helical filament. The micrograph shows actin filaments decorated with myosin S1 heads containing the essential light chain. The 3D reconstruction, at a resolution of 30–35 Å, is a composite in which the differently coloured parts are derived from a series of difference maps that were superimposed on F-actin. The components include: F-actin (blue), myosin heavy-chain motor domain (orange), essential light chain (purple), regulatory light chain (white), tropomyosin (green) and myosin motor domain N-terminal beta-barrel (red). Courtesy of A. Lin, M. Whittaker & R. Milligan (Scripps Research Institute, La Jolla). (*d*) A 2D crystal: light-harvesting complex LHCII at 3.4 Å resolution. The model shows the protein backbone and the arrangement of chromophores in a number of trimeric subunits in the crystal lattice. In this example, image contrast is too low to see any hint of the structure without image processing (see also Fig. 19.6.4.2). Courtesy of W. Kühlbrandt (Max-Planck-Institute for Biophysics, Frankfurt).

(Pebay-Peyroula *et al.*, 1997; Essen *et al.*, 1998; Luecke *et al.*, 1998). (2) Density maps of the 50S ribosomal subunits from two species obtained by cryo EM (Frank *et al.*, 1995; Ban *et al.*, 1998) were used to help solve the X-ray crystal structure of the *Haloarcula marismortui* 50S subunit (Ban *et al.*, 1998).

19.6.5. Recent trends

The new generation of intermediate-voltage (∼300 kV) FEG microscopes becoming available is now making it much easier to obtain higher-resolution images which, by use of larger defocus values, have good image contrast at both very low and very high resolution. The greater contrast at low resolution greatly facilitates particle-alignment procedures, and the increased contrast resulting from the high-coherence illumination helps to increase the signal-to-noise ratio for the structure at high resolution. Cold stages are constantly being improved, with several liquid-helium stages now in operation (*e.g.* Fujiyoshi *et al.*, 1991; Zemlin *et al.*, 1996). Two of these are commercially available from JEOL and FEI/Philips/Gatan. The microscope vacuums are improving so that the bugbear of ice contamination in the microscope, which prevents prolonged work on the same grid, is likely to disappear soon. The improved drift and vibration performance of the cold stage means longer (and therefore more coherently illuminated) exposures at higher resolution can be recorded more easily. Hopefully, the first atomic structure of a single-particle macromolecular assembly solved by electron microscopy will soon become a reality.

Finally, three additional likely trends include: (1) increased automation, including the recording of micrographs, and the use of spot-scan procedures in remote microscope operation (Kisseberth *et al.*, 1997; Hadida-Hassan *et al.*, 1999) and in every aspect of image processing; (2) production of better electronic cameras (*e.g.* CCD or pixel detectors); and (3) increased use of dose-fractionated, tomographic tilt series to extend EM studies to the domain of larger supramolecular and cellular structures (McEwen *et al.*, 1995; Baumeister *et al.*, 1999).

Acknowledgements

We are greatly indebted to all our colleagues at Purdue and Cambridge for their insightful comments and suggestions, to B. Böttcher, R. Crowther, J. Frank, W. Kühlbrandt and R. Milligan for supplying images used in Fig. 19.6.5.1, which gives some examples of the best work done recently, and J. Brightwell for editorial assistance. TSB was supported in part by grant GM33050 from the National Institutes of Health.

Table 21.2.3.1. *Parameters computed for the analysis of the structure-factor data*

The first column lists the parameter, the second column gives the formula or definition of the parameter and the third column contains a short description of the meaning of the parameters when warranted.

Parameter	Formula/definition	Meaning
Completeness (%)	Percentage of the expected number of reflections for the given crystal space group and resolution	
B_overall (Patterson)	$8\pi^2\sigma_{\rm Patt}/(2)^{1/2}$ *	Overall B factor
R_stand(F)	$\langle\sigma(F)\rangle/\langle F\rangle$ †	Uncertainty of the structure-factor amplitudes
Optical resolution	$(\sigma^2_{\rm Patt}+\sigma^2_{\rm sph})^{1/2}$ *‡	Expected minimum distance between two resolved atomic peaks
Expected optical resolution	Optical resolution computed considering all reflections	
CC_F	$\dfrac{\langle F_{\rm obs}F_{\rm calc}\rangle-\langle F_{\rm obs}\rangle\langle F_{\rm calc}\rangle}{\left[(\langle F^2_{\rm obs}\rangle-\langle F_{\rm obs}\rangle^2)(\langle F^2_{\rm calc}\rangle-\langle F_{\rm calc}\rangle^2)\right]^{1/2}}$	Correlation coefficient between the observed and calculated structure-factor amplitudes
S	$\left\{\dfrac{\sum(F_{\rm obs}f_{\rm cutoff})^2}{\sum\left[F_{\rm calc}\exp(-B^{\rm overall}_{\rm diff}s^2)f_{\rm cutoff}\right]^2}\right\}^{1/2}$ §	Factor applied to scale $F_{\rm calc}$ to $F_{\rm obs}$
$f_{\rm cutoff}$	$1-\exp(-B_{\rm off}s^2)$ ¶	Function applied to obtain a smooth cutoff for low-resolution data

* $\sigma_{\rm Patt}$ is the standard deviation of the Gaussian fitted to the Patterson origin peak.
† F is the structure-factor amplitude, and $\sigma(F)$ is the structure-factor standard deviation. The brackets denote averages.
‡ $\sigma_{\rm sph}$ is the standard deviation of the spherical interference function, which is the Fourier transform of a sphere of radius $1/d_{\rm min}$, with $d_{\rm min}$ being the minimum d spacing.
§ $B^{\rm overall}_{\rm diff}=B^{\rm overall}_{\rm obs}-B^{\rm overall}_{\rm calc}$ is added to the calculated overall B factor, $B_{\rm overall}$, so as to make the width of the calculated Patterson origin peak equal to the observed one; s is the magnitude of reciprocal-lattice vector.
¶ $B_{\rm off}=4d^2_{\rm max}$, where s and $d_{\rm max}$, respectively, are the magnitude of the reciprocal-lattice vector and the maximum d spacing.

approach of Sheriff & Hendrickson (1987), and applies the anisotropic scaling after the Patterson scaling is performed (Murshudov *et al.*, 1998).

To assess the quality of the structure-factor data, the program computes four additional quantities (see Table 21.2.3.1 for details): the completeness of the data, the uncertainty of the structure-factor amplitudes, the optical resolution and the expected optical resolution. The latter two quantities represent the expected minimum distance between two resolved atomic peaks in the electron-density map when the latter is computed with the set of reflections specified by the authors and with all the reflections, respectively.

21.2.3.1.1.2. *Global agreement between the model and experimental data*

To evaluate the global agreement between the atomic model and the experimental data, the program computes three classical quality indicators: the R factor, $R_{\rm free}$ (Brünger, 1992*b*) and the correlation coefficient CC_F between the calculated and observed structure-factor amplitudes (Table 21.2.3.1). The R factor is computed using all the reflections considered (except those approximated by their average value in the corresponding resolution shell) and applying the same resolution and σ cutoff as those reported by the authors. $R_{\rm free}$ is computed using the subset of reflections specified by the authors. In addition, the R factor is evaluated using the 'non-free' subset of reflections (those not used to compute $R_{\rm free}$). The correlation coefficient is computed using all reflections from the reported high-resolution limit, applying the smooth low-resolution cutoff (see Table 21.2.3.1) but no σ cutoff.

21.2.3.1.1.3. *Estimations of errors in atomic positions*

The errors associated with the atomic positions are expressed as standard deviations (σ) of these positions. *SFCHECK* computes three different error measures. One is the original error measure of Cruickshank (1949). The second is a modified version of this error measure, in which the difference between the observed and calculated structure factors is replaced by the error in the experimental structure factors. The first two error measures are the expected maximal and minimal errors, respectively, and the third measure is the diffraction-component precision indicator (DPI). The mathematical expressions for these error measures are given in Table 21.2.3.2, and further details can be found in Vaguine *et al.* (1999).

21.2.3.1.1.4. *Local agreement between the model and the experimental data*

In addition to the global structure quality measures, *SFCHECK* also determines the quality of the model in specific regions. Several quality estimators can be calculated for each residue in the macromolecule and, whenever appropriate, for solvent molecules and groups of atoms in ligand molecules. These estimators are the normalized atomic displacement (Shift), the correlation coefficient between the calculated and observed electron densities (Density correlation), the local electron-density level (Density index), the average B factor (B-factor) and the connectivity index (Connect), which measures the local electron-density level along the molecular backbone. These quantities are computed for individual atoms and averaged over those composing each residue or group of atoms [see Table 21.2.3.3 and Vaguine *et al.* (1999) for details].

21.2.3.1.2. *Evaluation of individual structures*

Figs. 21.2.3.1–21.2.3.3 summarize the analysis carried out by *SFCHECK* on the protein rusticyanin from *Thiobacillus ferrooxidans* (1RCY) (Walter *et al.*, 1996). Fig. 21.2.3.1 displays the numerical results from the analysis of the structure-factor data and from the evaluation of the global agreement between the model and the data. The R-factor and $R_{\rm free}$ values, computed by *SFCHECK*

Table 22.1.1.4. *Standard atomic volumes*

Tsai *et al.* (1999) and Tsai *et al.* (2001) clustered all the atoms in proteins into the 18 basic types shown below. Most of these have a simple chemical definition, *e.g.* '=O' are carbonyl carbons. However, some of the basic chemical types, such as the aromatic CH group ('≥CH'), need to be split into two subclusters (bigger and smaller), as is indicated by the column labelled 'Cluster'. Volume statistics were accumulated for each of the 18 types based on averaging over 87 high-resolution crystal structures (in the same fashion as described for the residue volumes in Table 22.1.1.3). No. is the number of atoms averaged over. The final column ('Symbol') gives the standardized symbol used to describe the atom in Tsai *et al.* (1999). The atom volumes shown here are part of the *ProtOr* parameter set (also known as the BL+ set) in Tsai *et al.* (1999).

Atom type	Cluster	Description	Average volume (Å^3)	Standard deviation (Å^3)	No.	Symbol
>C=	Bigger	Trigonal (unbranched), aromatics	9.7	0.7	4184	C3H0b
>C=	Smaller	Trigonal (branched)	8.7	0.6	11876	C3H0s
≥CH	Bigger	Aromatic, CH (facing away from main chain)	21.3	1.9	2063	C3H1b
≥CH	Smaller	Aromatic, CH (facing towards main chain)	20.4	1.7	1742	C3H1s
>CH–	Bigger	Aliphatic, CH (unbranched)	14.4	1.3	3642	C4H1b
>CH–	Smaller	Aliphatic, CH (branched)	13.2	1.0	7028	C4H1s
$-CH_2-$	Bigger	Aliphatic, methyl	24.3	2.1	1065	C4H2b
$-CH_2-$	Smaller	Aliphatic, methyl	23.2	2.3	4228	C4H2s
$-CH_3$		Aliphatic, methyl	36.7	3.2	3497	C4H3u
>N–		Pro N	8.7	0.6	581	N3H0u
>NH	Bigger	Side chain NH	15.7	1.5	446	N3H1b
>NH	Smaller	Peptide	13.6	1.0	10016	N3H1s
$-NH_2$		Amino or amide	22.7	2.1	250	N3H2u
$-NH_3^+$		Amino, protonated	21.4	1.2	8	N4H3u
=O		Carbonyl oxygen	15.9	1.3	7872	O1H0u
–OH		Alcoholic hydroxyl	18.0	1.7	559	O2H1u
–S–		Thioether or –S–S–	29.2	2.6	263	S2H0u
–SH		Sulfhydryl	36.7	4.2	48	S2H1u

Calculations based on crystal structures and simulations have shown that the protein surface has intermediate packing, being packed less tightly than the core but not as loosely as liquid water (Gerstein & Chothia, 1996; Gerstein *et al.*, 1995). One can understand the looser packing at the surface than in the core in terms of a simple trade-off between hydrogen bonding and close packing, and this can be explicitly visualized in simulations of the packing in simple toy systems (Gerstein & Lynden-Bell, 1993*a*,*b*).

22.1.2. Molecular surfaces: calculations, uses and representations

(M. S. CHAPMAN AND M. L. CONNOLLY)

22.1.2.1. *Introduction*

22.1.2.1.1. *Uses of surface-area calculations*

Interactions between molecules are most likely to be mediated by the properties of residues at their surfaces. Surfaces have figured prominently in functional interpretations of macromolecular structure. Which residues are most likely to interact with other molecules? What are their properties: charged, polar, or hydrophobic? What would be the estimated energy of interaction? How do the shapes and properties complement one another? Which surfaces are most conserved among a homologous family? At the centre of these questions that are often asked at the start of a structural interpretation lies the calculation of the molecular and/or accessible surfaces.

Surface-area calculations are used in two ways. Graphical surface representations help to obtain a quick intuitive understanding of potential molecular functions and interactions through visualization of the shape, charge distribution, polarity, or sequence conservation on the molecular surface (for example). Quantitative calculations of surface area are used *en route* to approximations of the free energy of interactions in binding complexes.

Part of this subject area was the topic of an excellent review by Richards (1985), to which the reader is referred for greater coverage of many of the methods of calculation. This review will attempt to incorporate more recent developments, particularly in the use of graphics, both realistic and schematic.

22.1.2.1.2. *Molecular, solvent-accessible and occluded surface areas*

The concept of molecular surface derives from the behaviour of non-bonded atoms as they approach each other. As indicated by the Lennard–Jones potential, strong unfavourable interactions of overlapping non-bonding electron orbitals increase sharply according to $1/r^{12}$, and atoms behave almost as if they were hard spheres with *van der Waals* radii that are characteristic for each atom type and nearly independent of chemical context. Of course, when orbitals combine in a covalent bond, atoms approach much more closely. Lower-energy attractions between atoms, such as hydrogen bonds or aromatic ring stacking, lead to modest reductions in the distance of closest approach. The van der Waals surface is the area of a volume formed by placing van der Waals spheres at the centre of each atom in a molecule.

Non-bonded atoms of the same molecule contact each other over (at most) a very small proportion of their van der Waals surface. The surface is complicated with gaps and crevices. Much of this surface is inaccessible to other atoms or molecules, because there is insufficient space to place an atom without resulting in forbidden overlap of non-bonded van der Waals spheres (Fig. 22.1.2.1). These crevices are excluded in the *molecular surface area*. The molecular

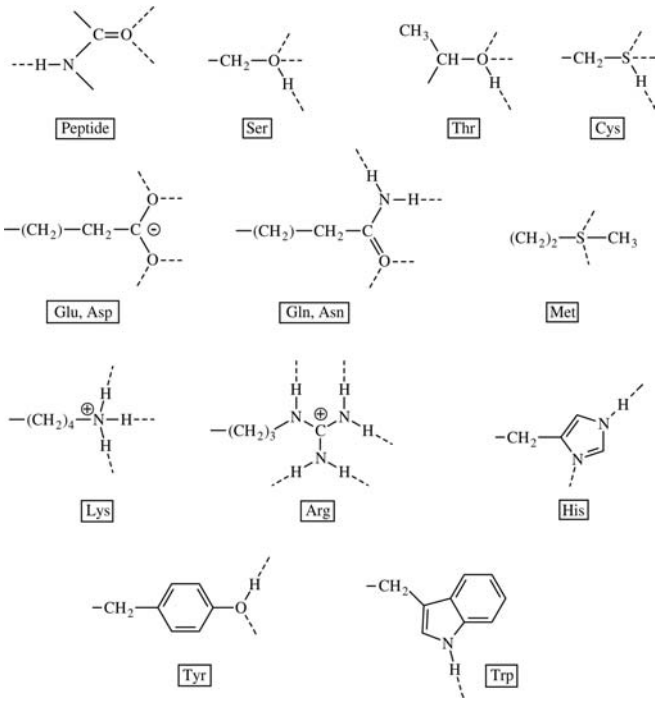


Fig. 22.2.3.1. Hydrogen-bonding potential of protein functional groups. Potential hydrogen bonds are shown with broken lines. Arg, Lys, Asp and Glu side chains are shown in their ionized forms.

recipients of hydrogen bonds from protein side chains in protein–DNA complexes. The sugar residues of RNA have a 2′-OH which can act as both hydrogen-bond donor and acceptor, and the 4′-O of both ribose and deoxyribose can potentially accept two hydrogen bonds.

It is the bases of DNA and RNA that have the greatest hydrogen-bonding potential, however, with a variety of hydrogen-bond donor or acceptor sites. Although each of the bases could theoretically occur in several tautomeric forms, only the canonical forms shown in Fig. 22.2.3.2 are actually observed in nucleic acids. This leads to clearly defined hydrogen-bonding patterns which are critical to both base pairing and protein–nucleic acid recognition. The —NH_2 and >NH groups act only as hydrogen-bond donors, and C=O only as acceptors, whereas the >N— centres are normally acceptors but at low pH can be protonated and act as hydrogen-bond donors.

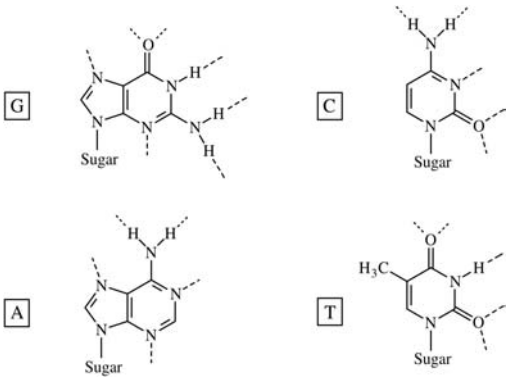


Fig. 22.2.3.2. Hydrogen-bonding potential of nucleic acid bases guanine (G), adenine (A), cytosine (C) and thymine (T) in their normal canonical forms.

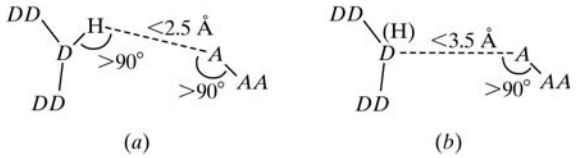


Fig. 22.2.4.1. Suggested criteria for identifying likely hydrogen bonds. *DD* and *AA* represent atoms covalently bonded to the donor atom, *D*, and acceptor atom, *A*, respectively. Here, (*a*) represents the criteria when the donor H atom can be placed, and (*b*) when it cannot be placed. Additional criteria based on the angle *DD*—*D*···*A* could be incorporated with (*b*). Adapted from Baker & Hubbard (1984) and McDonald & Thornton (1994*a*).

22.2.4. Identification of hydrogen bonds: geometrical considerations

Because hydrogen bonds are electrostatic interactions for which the attractive energy falls off rather slowly (Hagler *et al.*, 1974), it is not possible to choose an exact cutoff for hydrogen-bonding distances. Rather, both distances and angles must be considered together; the latter are particularly important because of the directionality of hydrogen bonding. Inferences drawn from distances alone can be highly misleading. An approach with an N—H···O angle of 90° and an H···O distance of 2.5 Å would be very unfavourable for hydrogen bonding, yet it translates to a N···O distance of 2.7 Å. This could (wrongly) be taken as evidence of a strong hydrogen bond.

For macromolecular structures determined by X-ray crystallography, problems also arise from the imprecision of atomic positions and the fact that H atoms cannot usually be seen. Thus, the geometric criteria must be relatively liberal. H atoms should also be added in calculated positions where this is possible; this can be done reliably for most NH groups (peptide NH, side chains of Trp, Asn, Gln, Arg, His, and all >NH and NH_2 groups in nucleic acid bases).

The hydrogen-bond criteria used by Baker & Hubbard (1984) are shown in Fig. 22.2.4.1. Very similar criteria are used in the program *HBPLUS* (McDonald & Thornton, 1994*a*), which also adds H atoms in their calculated positions if they are not already present in the coordinate file. In general, hydrogen bonds may be inferred if an interatomic contact obeys *all* of the following criteria:

(1) The distance H···*A* is less than 2.5 Å (or *D*···*A* less than 3.5 Å if the donor is an —OH or —NH_3^+ group or a water molecule).

(2) The angle at the H atom, *D*—H···*A*, is greater than 90°.

(3) The angle at the acceptor, *AA*—*A*···H (or *AA*—*A*···*D* if the H-atom position is unreliable), is greater than 90°.

Other criteria can be applied, for example taking into account the hybridization state of the atoms involved and the degree to which any approach lies in the plane of the lone pair(s). In all analyses of hydrogen bonding, however, it is clear that a combination of distance and angle criteria is effective in excluding unlikely hydrogen bonds.

22.2.5. Hydrogen bonding in proteins

22.2.5.1. *Contribution to protein folding and stability*

The net contribution of hydrogen bonding to protein folding and stability has been the subject of much debate over the years. The current view is that although the hydrophobic effect provides the driving force for protein folding (Kauzmann, 1959), many polar groups, notably peptide NH and C=O groups, inevitably become buried during this process, and failure of these groups to find hydrogen-bonding partners in the folded protein would be strongly destabilizing. This, therefore, favours the formation of secondary

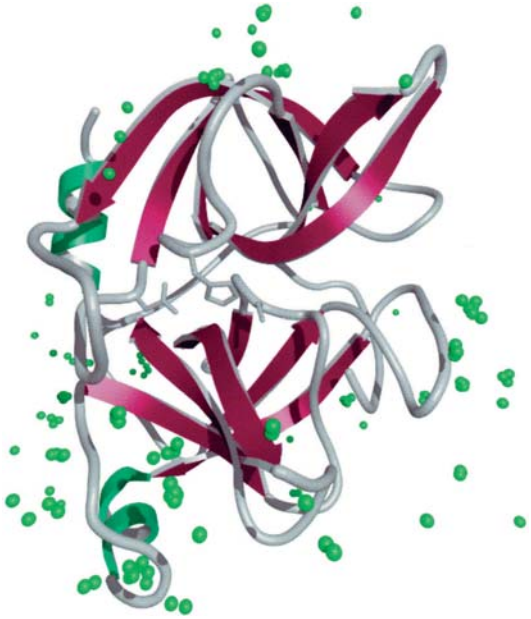


Fig. 23.4.4.6. Elastase structure represented as in Fig. 23.4.4.4. The crystallographic water molecules involved in crystal contacts in 11 superimposed elastase structures solved in a variety of solvents are shown in green.

isopropylanilide (Mattos *et al.*, 1994) or the trifluoroacetyl-Lys-Pro-*p*-trifluoromethylanilide (Mattos *et al.*, 1995) inhibitors in the structures of their complexes with elastase. These inhibitors span a large area of the active site, including an exosite not occupied by substrate analogue inhibitors (Mattos *et al.*, 1994, 1995). The water-binding sites in the active site are not very well conserved, with most sites represented in only two to four of the 11 structures. When all of the structures are superimposed, there is at least one water molecule in each of the subsites in the elastase active site. These water molecules are displaced either by inhibitors or by organic solvent molecules in the various structures. It is not surprising that in elastase, a protein with relatively broad substrate specificity, the active site in the uncomplexed native protein is populated by many displaceable surface water molecules. With the exception of a water molecule present in the oxyanion hole, these water molecules tend to make a single hydrogen bond with the protein. This hydrogen-bonding interaction is not generally conserved between different structures where a given site is occupied in multiple structures. The displacement of these water molecules upon ligand binding is entropically favourable, as they are released into bulk solvent, without too much enthalpic cost. This relatively small enthalpic cost can be compensated by the protein–ligand interactions.

Fig. 23.4.4.8 shows all of the 1661 water molecules colour-coded by the various classifications described above. Clearly, the entire surface of the protein is well hydrated. Notice how the yellow channel waters are often followed by a red buried water molecule. In addition, there is often no obvious spatial distinction between molecules categorized as crystal contacts (green) and those categorized as surface (blue).

23.4.4.2.2. *T4 lysozyme*

Over 150 mutants of T4 lysozyme have been studied to date, and, for the majority of these, the crystal structures are available. Although most of the mutant structures crystallize isomorphously to the wild type, many of them provide a view of the molecule in different crystal environments. This collection of structures leads to

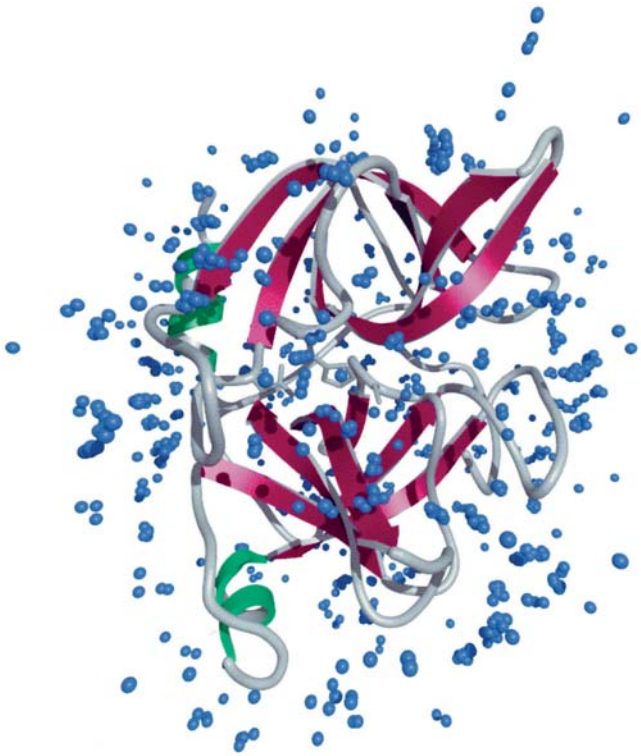


Fig. 23.4.4.7. Elastase structure represented as in Fig. 23.4.4.4. The surface crystallographic water molecules found in 11 superimposed elastase structures solved in a variety of solvents are shown in blue.

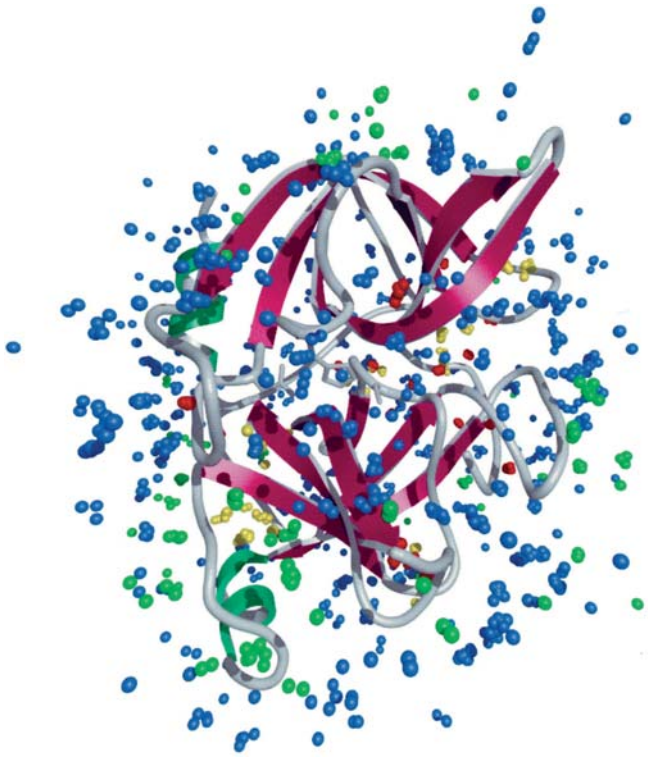


Fig. 23.4.4.8. Elastase structure represented as in Fig. 23.4.4.4. The 1661 water molecules found in 11 superimposed elastase structures of elastase are colour-coded as in Figs. 23.4.4.4–23.4.4.7.

26. A HISTORICAL PERSPECTIVE

26.1. How the structure of lysozyme was actually determined

BY C. C. F. BLAKE, R. H. FENN, L. N. JOHNSON, D. F. KOENIG, G. A. MAIR, A. C. T. NORTH, J. W. H. OLDHAM, D. C. PHILLIPS, R. J. POLJAK, V. R. SARMA AND C. A. VERNON

26.1.1. Introduction

For protein crystallographers, the year 1960 was the spring of hope. The determination of the three-dimensional structure of sperm-whale myoglobin at 2 Å resolution (Kendrew *et al.*, 1960) had shown that such analyses were possible, and the parallel study of horse haemoglobin at 5.5 Å resolution (Perutz *et al.*, 1960) had shown that even low-resolution studies could, under favourable circumstances, reveal important biological information. All seemed set for a dramatic expansion in protein studies.

At the Royal Institution in London, two of us (CCFB and DCP) had used the laboratory-prototype linear diffractometer (Arndt & Phillips, 1961) to extend the myoglobin measurements to 1.4 Å resolution for use in refinement of the structure (Watson *et al.*, 1963), and we had begun a detailed study of irradiation damage in the myoglobin crystals (Blake & Phillips, 1962). Meanwhile, David Green, an early contributor to the haemoglobin work (Green *et al.*, 1954), and ACTN had initiated a study of β-lactoglobulin (Green *et al.*, 1956) and worked together on oxyhaemoglobin before Green went to the Massachusetts Institute of Technology (MIT) in 1959 on leave for a year. At roughly this time, many of the participants in the myoglobin and haemoglobin work at Cambridge went off to other laboratories to initiate or reinforce other studies. Thus, Dick Dickerson went with Larry Steinrauf to the University of Illinois, Urbana, to start a study of the triclinic crystals of hen egg-white lysozyme.

RJP went to MIT from the Argentine as a post-doctoral fellow in 1958 and worked initially with Martin Buerger. In 1959 he transferred to Alex Rich's laboratory and there he soon came into contact with a number of veterans of the myoglobin and haemoglobin work. In addition to David Green were Howard Dintzis, who had discovered a number of the important heavy-atom derivatives of myoglobin (Bluhm *et al.*, 1958) and was now on the staff at MIT, and David Blow, who had first used multiple isomorphous replacement and anomalous scattering to determine haemoglobin phases (Blow, 1958) and was on leave from Cambridge. The influence of these people, combined with lectures by John Kendrew and then by Max Perutz on visits to MIT, soon convinced RJP that working on the three-dimensional structures of proteins was the most challenging and fruitful research that a crystallographer could undertake. Dintzis, in particular, persuaded him that preparing heavy-atom derivatives was no great problem, and Blow urged him to look for commercially available proteins that were known to crystallize. This soon focused his attention also on hen egg-white lysozyme (Fleming, 1922), but in the tetragonal rather than the triclinic crystal form. He quickly learned to grow crystals by the method described by Alderton *et al.* (1945) and then found that precession photographs of crystals soaked in uranyl nitrate showed intensities that differed significantly from those given by the native crystals. Encouraged by these results, he asked Max Perutz whether he could join the Cambridge Laboratory, but Max, having no room in Cambridge, suggested that he write to Sir Lawrence Bragg about going to the Royal Institution. Bragg replied with an offer of a place to work on β-lactoglobulin with David Green, who had by then returned to London. RJP accepted the offer and left for London late in 1960 – after first discussing what was going on at the Royal Institution with ACTN, who had just arrived at MIT for a year's leave with Alex Rich.

Early in 1961, RJP showed Bragg his precession photographs of potential lysozyme derivatives, and Bragg enthusiastically encouraged him to continue the work, at the same time urging DCP to arrange as much support as possible. This was a characteristic response by Bragg, who was well aware that at least two other groups were already working on lysozyme, Dickerson and Steinrauf at Urbana and Pauling and Corey at Cal Tech (Corey *et al.*, 1952): competition with Pauling was a common feature of his career. In describing his reaction to Bragg's encouragement, RJP recalled Metchnikoff's view of Pasteur. 'He transferred his enthusiasm and energy to his colleagues. He never discouraged anyone by the air of scepticism so common among scientists who had attained the height of their success . . . He combined with genius a vibrant soul, a profound goodness of heart.'

26.1.2. Structure analysis at 6 Å resolution

26.1.2.1. *Technical facilities*

In 1961, the Davy Faraday Laboratory was well equipped with X-ray generators. They included both conventional X-ray tubes, operating at 40 kV and 20 mA to produce copper $K\alpha$ radiation, and high-powered rotating-anode tubes that had been built in the laboratory to the design of D. A. G. Broad (patent 1956) under the direction of U. W. Arndt. We had a number of Buerger precession cameras and a Joyce–Loebl scanning densitometer, which had been used in the analysis of myoglobin (Kendrew *et al.*, 1960). In addition, we had a laboratory prototype linear diffractometer (Arndt & Phillips, 1961), which had been made in the laboratory workshop by T. H. Faulkner, and the manually operated three-circle diffractometer that had been used to make some of the measurements in the 6 Å studies of myoglobin (Kendrew *et al.*, 1958) and haemoglobin (Cullis *et al.*, 1961). The diffractometers were used with sealed X-ray tubes, since the rotating anodes were not considered to be reliable or stable enough for this purpose.

At this stage, most of the computations were done by hand, but we did have access to the University of London Ferranti MERCURY computer, usually in the middle of the night. This machine was programmed in MERCURY Autocode. The development of the early computers, their control systems and compilers mentioned in this article have been described by Lavington (1980).

26.1.2.2. *Lysozyme crystallization*

Tetragonal lysozyme crystals were first reported by Abraham & Robinson (1937) and the standard method of preparation was developed by Alderton *et al.* (1945); RJP used this method. Lyophilized lysozyme was obtained commercially and dissolved in distilled water at concentrations ranging from 50 to 100 mg ml^{-1}. To a volume of the lysozyme solution, an equal volume of 10% (*w*/*v*) NaCl in 0.1 *M* sodium acetate (pH 4.7) was added. About 1 to 2 ml aliquots of this mixture were pipetted into glass vials and tightly capped. Large crystals, frequently with