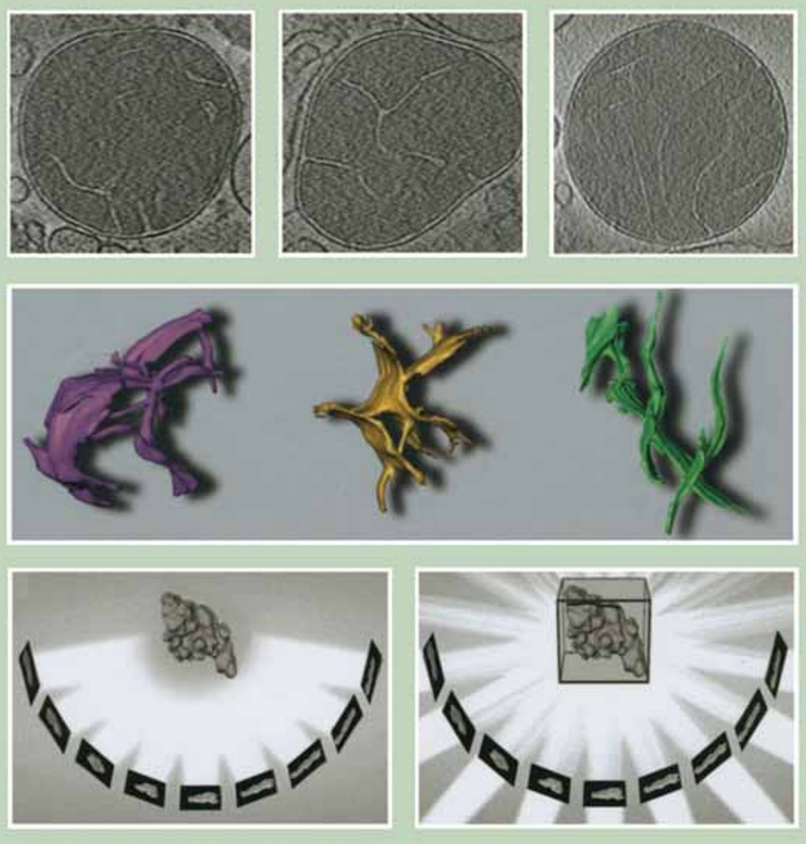
ELECTRON TOMOGRAPHY

Methods for Three-Dimensional Visualization of Structures in the Cell

SECOND EDITION



JOACHIM FRANK
EDITOR

Springer

Electron Tomography

Second Edition

Electron Tomography

Methods for Three-Dimensional Visualization of Structures in the Cell

Second Edition

Joachim Frank
Editor

Springer

Joachim Frank
Wadsworth Center and
State University of New York
Albany, NY 10201
USA

Cover Illustration

Upper two panels: various topologies exhibited by the inner membrane of rat liver mitochondria. Chyong Ere Hsieh, Christian Renken, and Michael Marko, Resource for the Visualization of Biological Complexity, Wadsworth Center, Albany, New York. (Design: Michael Watters.)
Lower panel: principle of three-dimensional reconstruction from a tilt series. After Baumeister *et al*., Trends Cell Biol. 9 (1999) 81–85. (Design: Paxton Provitera and Michael Watters.)

Library of Congress Control Number: 200620791

ISBN-10: 0-387-31234-X Printed on acid-free paper
ISBN-13: 978-0387-31234-7

9 8 7 6 5 4 3 2 1

springer.com

Preface to the First Edition (1992)

Some physicists may be drawn to biology because of the challenge that lies in the vast complexity of biological matter; what attracted me initially was the curious paradox that makes electron microscopy of macromolecules possible—*phase contrast*, the contrast that arises not *despite* but *because* of, the imperfections of the objective lens. It is the capricious nature of such details that carries the promise of future problems finding totally unexpected (and sometimes surprisingly simple) solutions. Once engaged in electron microscopy, as a student I was in awe of the wide range of forms in which living matter is organized, but I was also frustrated by the central limitation of the instrument—that it renders these structures only in the confusing, highly ambiguous form of projections.

Three-dimensional information about an object is usually obtained in a cumbersome way, by a process that does not leave the object intact, namely by cutting and slicing, and by stacking or geometrically relating the resulting images. Consider the origins of anatomy, which set itself the task of making a three-dimensional image of the body with all its organs. It started as a heretical undertaking because it required dissection, intrusion into the body, violating its sancity which was being upheld by the Roman Church. Because of the need for dissection, the teaching of anatomy in the Middle Ages was a clandestine operation performed by candlelight in a windowless hall, with the corpse lying on a table that was specially designed to hide it rapidly, in case the authorities stormed the premises. Perspective anatomical drawings and three-dimensional models emerged as the result of an intense visual, tactile and visceral effort on the part of the scholar. Centuries after this type of three-dimensional imaging with the scalpel was begun, computerized axial tomography (CAT) was invented, a miraculous tool to look inside a living body without a single cut.

This book deals with a similar revolution (albeit on a different time scale) in the study of the cell's ultrastructure, brought about by the application of tomographic techniques to electron microscopy. For a long time, structural information about cell components had to be inferred from images of thin sections, the thickness being limited by the path length of 100-kV electrons in biological matter. The limitations of sectioning are well

known: it produces distortions and material loss, and additional errors arise in the attempt to stack the section images to form a three-dimensional representation. Organelles of complex shape have proved difficult or impossible to study in this way. The problem is solved by increasing the voltage to the range of 400 to 1000 kV, thereby increasing the penetration thickness, and using a *series of views* rather than a single one to generate a 'true' three-dimensional image. Again, an inside look is obtained into the structure, which remains intact during the investigation.

Similar techniques have been developed for macromolecular assemblies that are in a much smaller size range and require no increase in voltage. Thus, electron tomography has filled a large gap: for the first time, all hierarchies of structural organization, ranging from the level of atomic structure (explored by X-ray crystallography) to the architecture of the cell (explored by confocal scanning light microscopy) can now be studied by quantitative three-dimensional imaging techniques that require no symmetry or order. Although this book deals only with the mid-level of structural organization in this vast logarithmic range, the challenges posed by the explosive increase in the amount of data, and the need to make them accessible in some 'nested' way are becoming evident. Clearly, the revolution in the biology of the cell will not be complete until a system of data storage, retrieval and visualization is found that is capable of mapping out the intrinsic complexity of the cell's components—the cell as a walk-in world, one of the momentous challenges of computational biology.

This book emerged as the result of a long and sometimes tedious interaction with the contributors. I was lucky to find authors that were not only experts in their fields but also enthusiastic to cooperate and share my vision. I am very grateful for their patience and endurance. Special thanks go to Michael Radermacher and Bruce McEwen, who discussed with me the concept of the book. I also wish to acknowledge valuable suggestions by Pawel Penczek and Terry Wagenknecht, who helped me read and reconcile the contributions. Finally, I thank Amelia McNamara of Plenum for initiating an endeavor that allowed me to illuminate this stimulating topic from many directions.

Preface to the Second Edition

Electron tomography has come of age. The technique, which had long led an existence as a more or less esoteric exercise of a few determined groups, has largely become main-stream. Packaged software that can be bought with the electron microscope has alleviated the need for special training and has made electron tomography accessible to scientists with diverse backgrounds, including those with little or no background in mathematics, physics or computer science. High-visibility papers and reviews have appeared with stunning three-dimensional images depicting the organization of the cell or a particular organelle. In some cases, such as the mitochondrion, long-standing ideas about the architecture have turned out to be utterly false. As a result of this development, today's cell biologists confronted with vexing problems of spatial organization are more likely to consider an investment in 3D imaging. Depending on temperament, extent of funding and determination, this investment can take the form of collaboration with one of the existing NCRR/NIH-supported Biotechnology Centers, collaboration with a colleague in the same institution or an effort to install an electron microscope equipped with an automated tomography kit in their own laboratories.

The first edition of this book brought together a group of experts in the fundamental and practical aspects of the technique. While the material in the mathematically oriented chapters is still relevant, new ideas have emerged on how to optimize the results, and a literature has sprung up around the applications of the different approaches. Updated chapters by the original contributors will therefore be useful at this point. Additional mathematical/computational tools have gained importance, namely those that aid in the interpretation of the reconstructed volumes. Among these are techniques for denoising, segmentation, docking and fitting. I am grateful to all contributors for the great investment of time and effort they have put in this endeavor, not only in drafting their chapters, but also in helping me review all the material for consistency and accuracy.

Joachim Frank,
December 14, 2005

Contributors

Montserrat Bárcena • Universiteit Utrecht, Department of Molecular Cell Biology, Krupt Building, Room West 511, Padualaan 8, 3584 CH Utrecht, The Netherlands. m.barcena@bio.uu.nl

Sami S. Brandt • Helsinki University of Technology, Laboratory of Computational Engineering, PO Box 9203, FI-02015, Finland. Sami.Brandt@tkk.fi

Jose-Maria Carazo • Centro Nacional de Biotecnologia (CSIC), Universidad Autónoma, 28049 Cantoblanco, Madrid, Spain. carazo@cnb.uam.es

Achilleas S. Frangakis • EMBL, European Molecular Biology Laboratory, Meyerhofstr. 1, D-69117 Heidelberg, Germany. frangak@embl.de

Joachim Frank • Howard Hughes Medical Institute, Health Research, Inc. at the Wadsworth Center, Department of Biomedical Sciences, State University of New York at Albany, Empire State Plaza, Albany, NY 12201-0509, USA. Joachim@wadsworth.org

Peter W. Hawkes • CEMES-CNR, BP 94347, F-31055 Toulouse Cedex 4, France. hawkes@wanadoo.fr

Reiner Hegerl • Max Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany. hegerl@ biochem.mpg.de

Gabor T. Herman • Department of Computer Sciences, The Graduate Center, City University of New York, New York, NY 10016, USA. gherman@gc.cuny.edu

Chyong-Ere Hsieh • Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, PO Box 509, Albany, NY 12201 USA. hsiehc@wadsworth.org

Qiang Ji • Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA qji@ecse.rpi.edu

Ming Jiang • Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA. qji@ecse.rpi.edu

Abraham J. Koster • Leiden University Medical Center, Molecular Cell Biology, Eindhovenweg 20, Leiden, The Netherlands. a.j.koster@lumc.nl

Jun Liu • Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306-4380, USA. winker@bio.fsu.edu

Pradeep K. Luther • Imperial College, Biomedical Sciences Division, Sir Alexander Fleming Building, Exhibition Road, London SW7 2AZ, UK. p.luther@imperial.ac.uk

Carmen Mannella • Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, PO Box 509, Albany, NY 12201, USA. Carmen@wadsworth.org

Roberto Marabini • Escuela Politécnica Superior, Universidad Autónoma, 28049 Cantoblanco, Madrid, Spain. Roberto@cnb.uam.es

Michael Marko • Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, PO Box 509, Albany, NY 12201, USA. marko@wadsworth.org

David N. Mastronarde • The Boulder Laboratory for 3D Electron Microscopy of Cells, Department of Molecular, Cellular, and Developmental Biology, Campus Box 347, University of Colorado, Boulder CO 80309, USA. mast@colorado.edu

Bruce F. McEwen • Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, PO Box 509, Albany, NY 12201-0509, USA. bruce@wadsworth.org

Pawel A. Penczek • The University of Texas–Houston Medical School, Department of Biochemistry and Molecular Biology, 6431 Fannin, MSB 6.218, Houston, TX 77030, USA. Pawel.A.Penczek@uth.tmc.edu

Michael Radermacher • University of Vermont College of Medicine, Department of Molecular Physiology & Biophysics, HSRF Building, Rm 120, Burlington, VT 05405, USA. mraderma@physiology.med.uvm.edu

Bimal K. Rath • Wadsworth Center, Empire State Plaza, PO Box 509, Albany, NY 12201-0509, USA. bimal@wadsworth.org

Carlos O. S. Sorzano • Escuela Politécnica Superior, Universidad San Pablo–CEU, 28668 Boadilla del Monte, Madrid, Spain. Coss.eps@ceu.es

Kenneth A. Taylor • Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306-4380, USA. taylor@bio.fsu.edu

Xun Wang • Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA. qji@ecse.rpi.edu

Hanspeter Winkler • Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306-4380, USA. winkler@bio.fsu.edu

Elmar Zeitler • Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4–6, D-14195 Berlin, Germany. zr@fhi-berlin.mpg.de

Contents

Introduction: Principles of Electron Tomography

Joachim Frank

1. WHAT IS ELECTRON TOMOGRAPHY?

Tomography is a method for reconstructing the interior of an object from its projections. The word *tomography* literally means the visualization of slices, and is applicable, in the strict sense of the word, only in the narrow context of the single-axis tilt geometry: for instance, in medical computerized axial tomography (CAT-scan imaging), the detector–source arrangement is tilted relative to the patient around a single axis (Fig. 1a). In electron microscopy, where the beam direction is fixed, the specimen holder is tilted around a single axis (Fig. 1b). However, the usage of this term has recently become more liberal, encompassing arbitrary geometries, *provided that the specimen is actively tilted into multiple angles*. In line with this relaxed convention, we will use the term *electron tomography* for any technique that employs the transmission electron microscope to collect projections of an object that is tilted in multiple directions and uses these projections to reconstruct the object in its entirety. Excluded from this definition are 'single-particle' techniques that make use of multiple occurrences of the

Joachim Frank • Howard Hughes Medical Institute, Health Research, Inc. at the Wadsworth Center, Department of Biomedical Sciences, State University of New York at Albany, Empire State Plaza, Albany, NY 12201-0509, USA

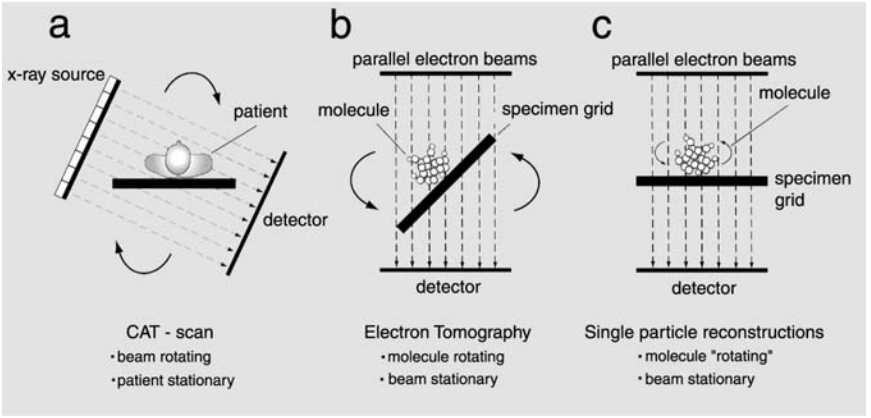


FIGURE 1. Three popular data collection geometries in 3D construction. (a) CAT-scan geometry, with the patient being stationary and a rigid source–detector arrangement tilted by equal increments; (b) equivalent single-axis tilt geometry in the transmission electron microscope, with the source–detector arrangement being stationary and the specimen tilted by equal increments; (c) as (b), but tilting replaced by the multiple incidence of molecules found in different random orientations.

object in different orientations, with or without the additional aid of symmetry (Fig. 1c). These techniques are covered elsewhere (non-symmetric: Frank, 1996, 2006; symmetric: Glaeser *et al*., 2007).

The terms '3D *imaging*' and '3D *electron microscopy*' have come into use as general terms to denote the capabilities of the instrument combined with the necessary computational tools to obtain a 3D image of an object's interior. For instance, a new series of Gordon Conferences was started in 1985 under the title 'Three-dimensional Electron Microscopy of Macromolecules', with the intention of providing a forum for scientists approaching the study of biological structure with both crystallographic and non-crystallographic techniques. (The term 3D *electron microscopy* may actually sound misleading since it conjectures an instrument with true 3D imaging performance. Such an instrument was actually conceived (Hoppe, 1972; Typke *et al*., 1976) but never advanced beyond the blueprint stage.)

2. A HISTORICAL PERSPECTIVE

3D imaging techniques are now commonplace in many areas of science, and it is difficult to recall that they have emerged only within the past 30 years; before that time, computers were simply too slow to be useful in processing 3D data on a routine basis, although much of the mathematical theory was well developed.

We may consider Plato's simile of the cave as a precursor to the reconstruction problem: here our ignorance of the essence of reality is depicted

by the situation of a man in a cave who watches shadows on the walls of his domicile; the shadows are all he sees of the world outside, and, because of the scantness of the information he receives, his comprehension of reality is severely limited. Similarly, a single projection, sometimes actually called a 'shadowgraph', of an object, is totally insufficient to establish its 3D shape. If we were prevented from changing the angle of view, we would be in a similar situation to the man in the cave, although without the dire existential ramifications.

The history of tomography (see also the brief account by Herman and Lewitt, 1979) is a history of intellectual challenges in a number of unrelated fields of science. As Elmar Zeitler recounts in Chapter 9, the same mathematical solution to the reconstruction problem that was found by Radon (1917) has had to be rediscovered numerous times. Two Nobel Prizes are directly related to 3D reconstruction: one that was shared by A. Cormack and G. N. Hounsfield in 1979 for the development of computerized axial tomography, and one in 1982 to Aaron Klug, in part for his pioneering work in the 3D reconstruction of molecular structures from their electron micrographs.

Klug traces the origins of 3D reconstruction in electron microscopy in his Nobel lecture (Klug, 1983). His laboratory, the Molecular Biology Laboratory of the Medical Research Council (MRC), is the appropriate starting point for a brief history of 3D imaging in electron microscopy. The predisposition of this institute for initiating quantitative structure research with the electron microscope is obvious, considering its historic role in the development of protein crystallography under Max Perutz's leadership.

DeRosier and Klug (1968) considered the problem of reconstructing the helical structure of the T4 phage tail from its projection (Fig. 2). To put their contribution into perspective, we must skip ahead and give a basic outline of the principle underlying 3D reconstruction. According to a fundamental mathematical theorem, the measurement of a projection yields a single central plane of the object's 3D Fourier transform. The Fourier transform, an alternative representation of the object, is a breakdown of the object's density distribution into sine waves. The Fourier transform constitutes a complete description of the object in the sense that, if we know the strengths (amplitudes) and phase shifts of all sine waves traveling in all possible directions and having wavelengths down to $d/2$, then the object is completely known to a resolution of d. The projection theorem thus suggests a recipe for reconstructing the object from its projections: by tilting the object through a range of $\pm 90^\circ$, we effectively sample its Fourier transform on a bundle of planes all intersecting one another on a single axis. It is clear that the angular spacing must be close enough to prevent information loss; in particular far away from the axis where the planes are maximally spaced and where the information on sine waves with the smallest wavelengths is situated.

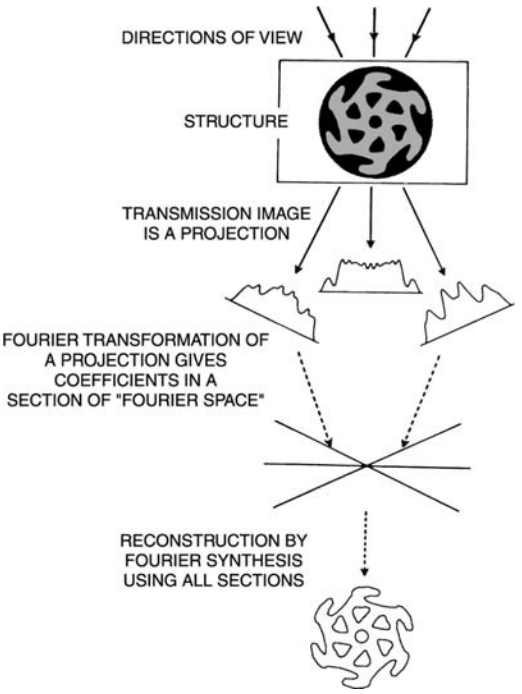


FIGURE 2. Principle of 3D reconstruction: the projections of the object furnish different central sections of the object's Fourier transform. If the number of projections is sufficient (making use of symmetries where possible), then the complete Fourier transform can be regenerated by interpolation, and from this the original object can be retrieved by inverse Fourier transformation. (Reproduced from DeRosier and Klug (1968), by permission of Macmillan Journals, Ltd.)

The application of this method to electron microscopy poses a problem because the tilt range is normally restricted for several reasons, the most important of which is the need to support the specimen on some type of grid that obstructs the electron path at high angles. Therefore, the angular range in commercial instruments does not usually exceed ±60°. Special tilt stages have been designed that push the range to ±85° (Chalcroft and Davey, 1984). However, when the object is contained in a thick plastic section, the increased path length of electrons traversing the sections at high angles also become a serious problem. One way to overcome this restriction is the development of tilt stages for cylindrical mounting of objects with 360° rotation capability. For instance, Barnard *et al.* (1992) placed a test object (spores) at the edge of an ultrathin glass capillary. Apart from these special cases, the experimental restriction to a range of about ±60° applies, which means that in the general case of an object without symmetry, a significant portion of the Fourier transform simply cannot be measured.

In contrast, when an object does possess symmetries, then the measurement of any projection yields other symmetry-related projections simultaneously. Another way of saying this is that, in this case, only part of the Fourier transform needs to be known for the entire Fourier transform to be generated. Among symmetric objects, those with helical symmetry, such as the T4 phage tail studied by DeRosier and Klug (1968), have a special position in that a single projection may be sufficient to generate the entire Fourier transform.

As early as 1970, Crowther and co-workers at the MRC formulated the approach to be used for reconstructing objects with or without symmetry with great clarity, and they also derived a general formula linking resolution, object size and number of projections. The first particle with icosahedral symmetry was reconstructed in 1970 (Crowther *et al.*, 1970b). Subsequently, Henderson and Unwin (1975) developed the reconstruction of single-layer, 'two-dimensional' crystals in the general crystallographic framework (see Amos *et al.*, 1982).

It is now necessary to illuminate the substantial contributions to the field by another group closely linked to crystallography: the group of Walter Hoppe at the Max Planck Institute in Munich (later relocated to Martinsried). Hoppe envisaged the prospect of 3D reconstruction in electron microscopy in imaging objects not amenable to crystallographic techniques. Consequently, he pursued almost exclusively the development of methods aimed at reconstructing objects lacking symmetry or crystalline order. Progress in this direction was initially slow because many tools of data processing had yet to be developed or adopted from other fields. The reconstruction of the fatty acid synthetase molecule in 1974 (Hoppe *et al.*, 1974) represented a significant achievement, which marked the beginning of electron tomography in the proper sense of the term. At that time, essentially all important tools were in place: the use of correlation functions for the alignment of projections, the Smith–Cormack scheme of 3D reconstruction (Cormack, 1964; Smith *et al.*, 1973) and the first sophisticated image-processing software system of modular design dedicated to electron microscopy applications (see Hegerl and Altbauer, 1982).

However, work in several other laboratories during that same period pointed to the deleterious effects of radiation damage, which made the quantitative interpretation of images taken with the standard imaging conditions questionable, and cast serious doubts on the significance of 3D information obtained by multiple exposure of the same object. According to Unwin and Henderson (1975), high-resolution information (at least to 7 Å) is preserved when the total dose is kept below 1 e/Å^2. Thus, it became apparent that 3D reconstruction would produce biologically significant results only under two rather narrowly defined circumstances: (i) when applied to macromolecular structures, only those data collection schemes are acceptable that make use of multiple occurrences of the same molecules, by extracting different projections from different 'repeats' of the molecule; and

(ii) when applied to cell components in an entirely different size range where resolution requirements are normally more modest (50–100 Å), and specialized higher voltage microscopes must be used for increased penetration, much higher accumulated radiation doses may be acceptable. In fact, these types of objects rarely exist in 'copies' with identical structure, thus excluding any approach that uses averaging implicitly or explicitly.

With hindsight, it must be seen as unfortunate that Hoppe's leading laboratory in 3D reconstruction of non-crystalline objects invested its main efforts in an area that does not fall in either category, namely tomography of single macromolecules (or complex assemblies such as the ribosome) from a tilt series, in the course of which the molecule receives a radiation dose that exceeds the limit found by Unwin and Henderson (1975) by a large factor. (The arguments put forward by Hoppe (1981) attempting to justify 3D electron microscopy of individual macromolecules receiving high doses of radiation are not convincing.)

Meanwhile, the general theory of 3D reconstruction was advanced by a number of studies; among these, the works of Bates's group (Lewitt and Bates, 1978a,b; Lewitt *et al*., 1978), Zwick and Zeitler (1973), Colsher (1977) and Gilbert (1972) should be mentioned for their relevance to our subject matter. 3D reconstruction in all areas of science proceeded at such a rapid rate that, in 1975, the Optical Society of America decided to organize a topical meeting on 3D reconstruction in Stanford, California. This meeting brought together contributors from a wide range of fields, such as geology, radioastronomy, radiology and electron microscopy. An overview of various implementations and applications presented at that meeting was compiled by Herman (1979).

At that time point, 3D reconstruction of general, asymmetric biological objects in electron microscopy took different paths, distinguished by the presence of redundancies or lack thereof, and the applicability of averaging techniques (Frank, 1975; Saxton and Frank, 1977). Particles, such as macromolecular assemblies, that exist in abundance with identical structure can be reconstructed from their 'single-particle' projections, i.e. from projections of particles that are dispersed and randomly oriented (Radermacher *et al*., 1987a,b; see Frank, 1996, 2006). Methods to align, classify and orient such projections, as well as 3D reconstruction from data with general geometries, constituted the main directions of algorithm development in this branch of 3D reconstruction. On the other hand, there are the kinds of specimen that lack redundancy altogether, and for such specimens—typically organelles and other subcellular structures—electron tomography is the only approach to 3D visualization available.

In the development of electron tomography of subcellular structures, progress hinged on the availability, to the biologist, of high- or intermediate-voltage electron microscopes equipped with precision tilt stages and supported by sophisticated image-processing resources. Centers with this degree of organization and sophistication did not emerge until the begin-

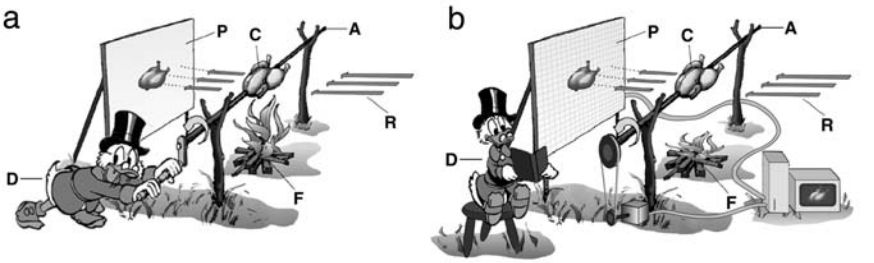


FIGURE 3. Electron tomography then (a) and now (b). (a) Adapted from Hoppe (1983); (b) adapted from B. Carragher, unpublished drawing.

ning of the 1980s when the National Institute of Health's Biotechnology program started to support three high-voltage microscopes dedicated to the biological sciences in the USA[1]. Thus, the pace of development of this technology was rather slow, especially considering the state of the art that already existed when Hoppe *et al*.'s fatty acid synthetase study (Hoppe *et al*., 1974) was published. However, perhaps the most important factor determining the pace with which 3D imaging with the electron microscope developed has been the speed and memory of computers. It must be realized that electron tomography posed computational problems of such magnitude that, until the beginning of the 1990s, only groups with access to mainframes were able to make significant progress. Other important factors were the slow progress toward automation of data collection and the need for image-processing software capable of handling the numerous combinations of operations that are encountered in the analysis of electron microscopic data.

Finally, the 1990s brought a breakthrough toward full automation, as affordable CCD cameras grew large enough to cope with the field sizes encountered in electron tomography, and electron microscopes were integrated with fast computer control. Here the work by Abraham Koster, one of the contributors to this volume (Chapter 4), deserves special mention (Koster *et al*., 1992). Nowadays, thanks to his and others' pioneering work, commercial instruments come equipped with the necessary gadgetry and software to perform low-dose data collection, as well as preliminary reconstruction, on the spot. Thus, with the new generation of powerful and smart electron microscopes, the drawing Walter Hoppe once used to illustrate both the potential (the capability of 3D imaging) and limitations (radiation damage) of electron tomography (Fig. 3a) has to be substantially revised (Fig. 3b).

[1] University of Colorado in Boulder, Colorado; University of Wisconsin in Madison; and New York State Department of Health in Albany, New York. Of these, only the one in Albany has remained in operation.

3. THE PRINCIPLE OF 3D RECONSTRUCTION

The principle of 3D reconstruction becomes clear from a formulation of the fundamental relationship between an object and its projections. An understanding of the basic concept of the Fourier transform is needed for this formulation. A brief introduction is provided in the following. For a more detailed introduction, the reader is referred to the specialized literature such as Bracewell (1999). A compilation of definitions and formulae for the case of discrete data is provided in the appendix of a book on 3D electron microscopy of macromolecular assemblies by the author (Frank, 2006).

The Fourier transform provides an alternative representation of an object by breaking it down into a series of trigonometric basis functions. For mathematical expediency, complex exponential waves of the form $\exp[2\pi i\mathbf{R}\mathbf{r}]$ are used instead of the more familiar sine and cosine functions. The argument vector describing a location in 3D space is $\mathbf{r} = (x, y, z)$, while $\mathbf{R} = (X, Y, Z)$ is a so-called spatial frequency vector, which gives both the direction of travel of a spatial wave and the number of full spatial oscillations per unit length. From such spatial waves, the object can be built up by linear superposition:

$$o(\mathbf{r}) = \sum_n c_n \exp[2\pi i\mathbf{R}_n\mathbf{r}] \tag{1}$$

with the complex coefficients c_n. The 3D Fourier transform may be visualized as a 3D scheme ('Fourier space') in which the coefficients c_n are arranged, on a regular grid, according to the position of the spatial frequency vector. Each coefficient c_n contains the information on the associated wave's amplitude (or strength),

$$a_n = |c_n| \tag{2}$$

and phase (or shift of the spatial wave in its travelling direction, with respect to the origin),

$$\phi_n = arctan\frac{Im\{c_n\}}{Re\{c_n\}} \tag{3}$$

The projection theorem offers a way to sample the Fourier transform of an object by measuring its projections. According to this theorem, *the 2D Fourier transform of a projection of the object is identical to a central section of the object's 3D Fourier transform.* Thus, by tilting the object into many orientations, one is, in principle, able to measure its entire Fourier transform. Obviously, the projections must be collected with a small angular increment and, ideally, over the full angular range. Then, after the Fourier summation in equation (1) is performed, the object can be retrieved. As always, the devil is in the details, as evidenced by the lengths and depths of the treatises by specialists found in this book.

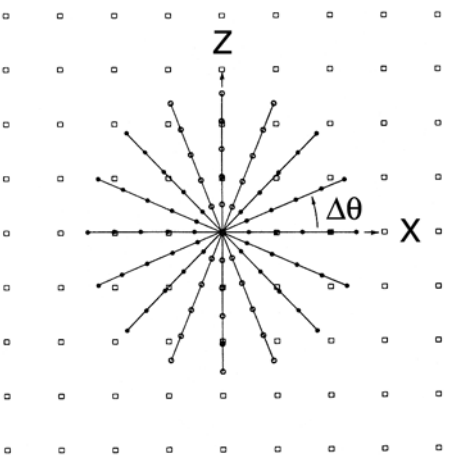


FIGURE 4. Sampling in Fourier space for single-axis tilting with equal increments $\Delta\theta$. For explanation, see text. Adapted from Frank (1992).

The angular increment $\Delta\theta$ is evidently determined by two parameters (Fig. 4): (i) the mesh size of the Fourier space grid; and (ii) the size of the region, in Fourier space, that needs to be filled. These quantities are in turn determined by object diameter and resolution:

1. The mesh size must be smaller than $1/D$, where D is the object diameter.
2. The region in Fourier space for which data must be acquired is a sphere with radius $1/d$, where d is the resolution distance, i.e. the size of the smallest feature to be visualized in the reconstruction.

According to these essentially geometrical requirements, the minimum number of (equispaced) projections works out to be (Bracewell and Riddle, 1967; Crowther *et al.*, 1970a):

$$N = \frac{\pi D}{d} \tag{4}$$

Reconstruction methods may be classified according to the way in which projections are collected or, alternatively, according to the way in which the object is retrieved from its measured projections. The former relates to the experiment, while the latter relates to the mathematical and computational aspects of reconstruction as discussed in Chapters 6, 7 and 8. As to the data collection geometries, there are three that have gained practical importance in electron microscopy: single-axis, double-axis and conical tilting.

Single-axis tilting is simply achieved by rotation of a side-entry rod in the electron microscope, whereas double-axis tilting involves a second tilt capability around an axis perpendicular to the first (Fig. 5a), and conical tilting provides a rotation capability in the inclined plane defined by the

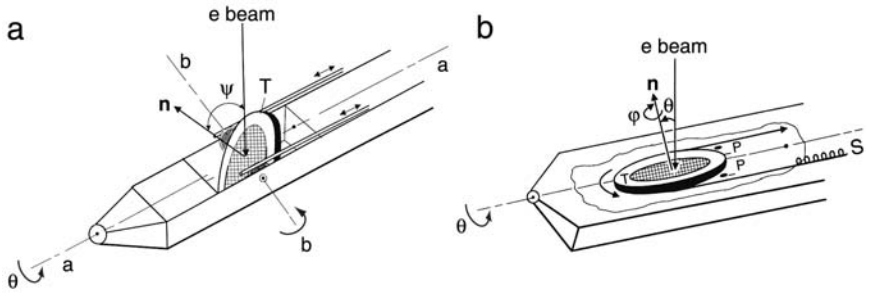


FIGURE 5. Schematic diagrams showing the principle of side-entry tilt stages with two degrees of freedom. (a) Double-tile stage. aa is the principal tilt axis, corresponding to the long axis of the rod. bb is the second tilt axis. θ and ψ are the corresponding tilt angles. **n** denotes the normal to the specimen plane. Tilting around the second axis is actuated by translation (indicated by arrows) of sliding rods which engage wheels attached to the turret T. (b) Tilt-rotation stage for conical geometry. Again **n** denotes the normal to the specimen plane. θ is the tilt angle, and φ the rotation angle in the plane of the circular turret T. Rotation is actuated by a cable pulled in the direction of the arrow, with return movement provided by a spring S. The turret is held in a stable position by retaining pins P. Adapted from Turner (1981).

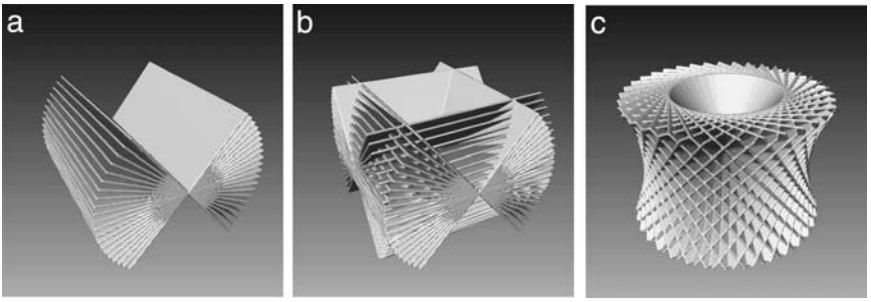


FIGURE 6. Coverage of 3D Fourier space in the case of three data collection geometries: (a) single-axis; (b) double-axis; and (c) conical. In each case, equal angular increments are depicted. From Lanzevecchia *et al.* (2005); reproduced with permission of Elsevier.

first tilt (Fig. 5b). It is easy to see, by invoking the projection theorem, that double-axis and conical tilting provide a much wider coverage of Fourier space if the maximum tilt angle is the same in all cases (Fig. 6a–c). However, the price to be paid for this information gain is a >2-fold increase in total dose (Frank and Radermacher, 1986; Radermacher and Hoppe, 1980). Because of this disadvantage, and the more elaborate experimental procedure, conical data collection has not been widely used in experimental protocols where a single biological structure is multiply exposed. (Conical data

collection, of course, has found widespread application in reconstructions of macromolecules in single-particle form from their projections (see Radermacher *et al*., 1987b).) Lately, the idea of using the conical tilt geometry in tomography has been revived by Lanzavecchia's group, with remarkable success (Lanzavechia *et al*., 2005; Zampighi *et al*., 2005).

4. HOW THIS BOOK IS ORGANIZED

The sequence of chapters in this book essentially follows the flow of information in electron tomography, proceeding from specimen preparation, to data collection in the instrument, and then to the techniques used for alignment, reconstruction and interpretation of the resulting tomographic volumes.

We start with the question of to what extent the object reconstructed from electron microscopic projections resembles the biological object. This question has three different aspects to it: one that has to do with the relationship between the native biological object and the specimen investigated in the electron microscope; the second with the relationship between that specimen and the images formed by the electron microscope; and a third with the relationship between the set of multiple projections and the final 3D image.

Two chapters deal specifically with the specimen preparation aspect and the question of fidelity to the original biological structure. The first, by Pradeep Luther (Chapter 1), examines the quality of specimen preparation in plastic sections, and the damage inflicted by the beam, with special attention to the problem of section shrinkage. Knowledge of the behavior of the specimen is of crucial importance in planning an experiment that requires multiple exposure of the same specimen. The other chapter, by Mike Marko, Chyongere Hsieh and Carmen Mannella (Chapter 2), describes experience gained with the new, promising technique of sectioning frozen-hydrated biological material prepared by high-pressure freezing for the purpose of electron tomography.

Peter Hawkes (Chapter 3) explores the conditions that must be satisfied in the imaging by electron microscopy for the observed projections to be regarded as simple line integrals of an object function. It is of course always possible to apply the reconstruction procedure 'blindly' to a set of experimental images and obtain a 3D density map, or 'object function'. The question is whether this reconstructed object has any meaning, or even a tractable relationship to the physical object the images originated from. By implication, the simple projection relationship and the image formation in bright field under the usual weak object assumptions yield a very elegant linear system description of the imaging and reconstruction process. As the specimen thickness increases, which is the case for cell sections investigated by electron tomography, multiple scattering increasingly interferes with the

linear system concept, but energy filtering (also covered in this chapter) can effectively increase the thickness range of validity.

Low-dose data collection, in the now common automated mode, is covered in a chapter by the pioneer of automated tomography, Abraham Koster, together with Montserrat Bárcena (Chapter 4). This chapter goes into all the necessary details regarding optimum settings, data collection protocols and the important considerations of dose fractionation.

Electron tomographic reconstruction requires that projections be aligned to a common frame of reference. The use of gold bead markers is now routine, and therefore it is justified that a chapter be devoted to the mathematical basis of marker-based alignment. David Mastronarde (Chapter 5), author of the well-known IMOD software, gives an expert introduction into this subject. However, the search for a reliable marker*less* alignment method continues, since the electron-opaque markers produce artifacts in the reconstruction volume that cannot be removed computationally. Sami Brandt (Chapter 6), one of the pioneers of markerless alignment techniques, has contributed an authoritative chapter on recent approaches to this problem.

Three chapters are devoted to the theory of reconstruction, addressing different issues that arise due to the peculiarities of data collection and numerical computation. We first present the chapter by Jose-Maria Carazo, Gabor Herman and co-workers (Chapter 7), which gives an overview on the approaches to the inverse problem presented by the reconstruction from a finite number of projections. This same chapter also introduces iterative algebraic methods, such as algebraic reconstruction techniques (ART) and related techniques. Next, Michael Radermacher (Chapter 8) goes into the details of weighted back-projection for general geometries, and formulates the algorithms underlying the computer programs now widely used in the field. Weighted back-projection methods have a special position in the practical implementation of 3D reconstruction, mainly because of their mathematical tractability and high computational efficiency. Radermacher summarizes the rationales and important formulae of weighted back-projection methods for regular and general geometries. Finally, in this section on the mathematics of reconstruction, Elmar Zeitler (Chapter 9) has contributed a chapter that presents an elegant general framework of reconstruction using special functions of mathematical physics, a chapter that brilliantly illuminates the inter-relationships of all approaches to reconstruction in use today.

Although the theoretical resolution obtainable in tomographic reconstructions is well known, the problem of how to measure the actual resolution achieved has been elusive. The chapter by Pawel Penczek and this author (Chapter 10) addresses this important issue with a new approach.

The remaining chapters in this book deal with different aspects of interpretation of the reconstruction. The first step is the removal of noise, either by simple Fourier filtration or more advanced "denoising" procedures. These procedures are described by Reiner Hegerl and Achilleas

Frangakis (Chapter 11). Segmentation is obviously the most important aspect as it is instrumental for the assignment of meaning to the different parts of a density map. Three chapters deal with segmentation, namely one authored by Achilleas Frangakis and Reiner Hegerl (Chapter 12), with segmentation based on local characteristics of the density distribution; the second, by Ming Jiang *et al*. (Chapter 13), on model-based segmentation making use of level set methods. The third chapter in this category addresses segmentation by rigid body motif search using cross-correlation. This chapter is co-authored by investigators who each have made separate contributions to this area of research: Achilleas Frangakis and Bimal Rath (Chapter 14). Another aspect of interpretation comes up when we try to characterize quasi-periodic structures, as presented by the complex organization of muscle trapped in the rigor state. We are then dealing with multiple versions of a 3D motif, which may allow the tracking of a physiological process evolving in time and space—a promising method of analysis described in the final chapter by Ken Taylor, Jun Liu and Hanspeter Winkler (Chapter 15).

ACKNOWLEDGEMENTS

This work was supported by HHMI and by Grant P41 RR01219 from the National Center for Research Resources (NCRR/NIH) and R37 GM 29169. I would like to thank Michael Watters for assistance with the illustrations.

REFERENCES

Amos, L. A., Henderson, R. and Unwin, P. N. T. (1982). Three-dimensional structure determination by electron microscopy of 2-dimensional crystals. *Prog. Biophys. Mol. Biol.* **39**: 183–231.

Andrews, H. C. (1970). *Computer Techniques in Image Processing*. Academic Press, New York.

Barnard, D. P., Turner, J.N., Frank, J. and McEwen, B.F. (1992). A 360° single-axis tilt stage for the high-voltage electron microscope. *J. Microsc.* **167**:39–48.

Bracewell, R. N. (1999). *The Fourier Transform and Its Applications*, 3rd edn. McGraw-Hill, New York.

Bracewell, R. N. and Riddle, A. C. (1967). Inversion of fan-beam scans in radio astronomy. *Astrophys. Soc*. **150**:427–434.

Chalcroft, J. P. and Davey, C. L. (1984). A simply constructed extreme-tilt holder for the Philips eucentric goniometer stage. *J. Microsc.* **134**:41–48.

Colsher, J. G. (1977). Iterative three-dimensional image reconstruction from tomographic projections. *Comput. Graphics Image Process.* **6**:513–537.

Cormack, A. M. (1964). Representation of a function by its line integrals, with some radiological applications. I. *J. Appl. Phys.* **35**:2908–2912.

Crowther, R. A., Amos, L. A., Finch, J. T. and Klug, A. (1970a). Three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs. *Nature* **226**: 421–425.

Crowther, R. A., DeRosier, D. J. and Klug, A. (1970b). The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. *Proc. R. Soc. B* **317**:319–340.

DeRosier, D. and Klug, A. (1968). Reconstruction of 3-dimensional structures from electron micrographs. *Nature* **217**:130–134.

Frank, J. (1975). Averaging of low exposure electron micrographs of nonperiodic objects. *Ultramicroscopy* **1**:159–162.

Frank, J. (1992). Introduction. In *Electron Tomography* (J. Frank, ed.) pp. 1–13. Plenum, New York.

Frank, J. (1996). *Three-dimensional Electron Microscopy of Macromolecules*. Academic Press, San Diego.

Frank, J. (2006). *Three-dimensional Electron Microscopy of Macromolecules*, 2nd edn. Oxford University Press, New York.

Frank, J. and Radermacher, M. (1986). Three-dimensional reconstruction of nonperiodic macromolecular assemblies from electron micrographs. In *Advanced Techniques in Electron Microscopy III* (J. K. Koehler, ed.). Springer-Verlag, New York, pp. 1–72.

Gilbert, P. F. C. (1972). The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. II. Direct methods. *Proc. R. Soc. B* **182**:89–117.

Glaeser, R. M., Downing, K. H., Chiu, W., Frank, J. and DeRosier, D. (2007). *Electron Crystallography of Biological Macromolecules*. Oxford University Press, New York.

Hegerl, R. and Altbauer, A. (1982). The 'EM' program system. *Ultramicroscopy* **9**:109–116.

Herman, G.T. (ed.) (1979). *Image Reconstruction from Projections*. Springer-Verlag, Berlin.

Henderson, R. and Unwin, P. N. T. (1975). Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* **257**:28–32.

Herman, G. T. and Lewitt, R. M. (1979). Overview of image reconstruction from projections, in *Image Reconstruction from Projections* (G. T. Herman, ed.). Springer-Verlag, Berlin, pp. 1–7.

Hoppe, W. (1972). Drei-dimensional abbildende Elektronenmikroskope. *Z. Naturforsch.* **27a**: 919–929.

Hoppe, W. (1981). Three-dimensional electron microscopy. *Annu. Rev. Biophys. Bioeng*. **10**: 563–592.

Hoppe, W. (1983). Elektronenbeugung mit dem Transmissions-Elektronenmikroskop als phasenbestimmendem Diffraktometer-von der Ortsfrequenzfilterung zur dreidimensionalen Strukturanalyse an Ribosomen. *Angew. Chem.* **95**:465–494.

Hoppe, W., Gassmann, J., Hunsmann, N., Schramm, H. J. and Sturm, M. (1974). Three-dimensional reconstruction of individual negatively stained fatty-acid synthetase molecules from tilt series in the electron microscope. *Hoppe-Seyler's Z. Physiol. Chem*. **355**:1483–1487.

Klug, A. (1983). From macromolecules to biological assemblies (Nobel lecture). *Angew. Chem.* **22**:565–636.

Koster, A. J., Chen, J. W., Sedat, J. W. and Agard, D. A. (1992). Automated microscopy for electron tomography. *Ultramicroscopy* **46**:207–227.

Lanzavecchia, S., Cantele, F., Bellon, P. L., Zampighi, L., Kreman, M., Wright, E., and Zampighi, G. A. (2005). Conical tomography of freeze-fracture replicas: a method for the study of integral membrane proteins inserted in phospholipids bilayers. *J. Struct. Biol*. **149**:87–98.

Lewitt, R. M. and Bates, R. H. T. (1978a). Image reconstruction from projections. I: General theoretical considerations. *Optik (Stuttgart)* **50**:19–33.

Lewitt, R. M. and Bates, R. H .T. (1978b). Image reconstruction from projections. III: Projection completion methods (theory). *Optik (Stuttgart)* **50**:189–204.

Lewitt, R. M., Bates, R. H. T. and Peters, T. M. (1978). Image reconstruction from projections. II: Modified back-projection methods. *Optik (Stuttgart)* **50**:85–109.

Radermacher, M. and Hoppe, W. (1980). Properties of 3D reconstruction from projections by conical tilting compared to single axis tilting. In *Proceedings of the 7th European Congress on Electron Microscopy*, Den Haag, Vol. I, pp. 132–133.

Radermacher, M., Wagenknechet, T., Verschoor, A. and Frank, J. (1987a). Three-dimensional structure of the large ribosomal subunit from *Escherichia coli*. *EMBO J*. **6**:1107–1114.

Radermacher, M., Wagenknecht, T., Verschoor, A. and Frank, J. (1987b). Three-dimensional reconstruction from single-exposure random conical tilt series applied to the 50S ribosomal subunit of *Escherichia coli*. *J. Microsc*. **146**:113–136.

Radon, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. *Math. Phys. Klasse* **69**:262–277.

Saxton, W. O. and Frank, J. (1977). Motif detection in quantum noise-limited electron micrographs by cross-correlation. *Ultramicroscopy* **2**:219–227.

Smith, P. R., Peters, T. M. and Bates, R. H. T. (1973). Image reconstruction from a finite number of projections. *J. Phys. A* **6**:361–382.

Turner, J. N. (1981). Stages and stereo pair recording. *Methods Cell Biol.* **23**:33–51.

Typke, D., Hoppe, W., Sessier, W. and Burger, M. (1976). Conception of a 3-D imaging electron microscopy. In *Proceedings of the 6th European Congress on Electron Microscopy* (D. G. Brandon, ed.), Vol. 1, Tal International, Israel, pp. 334–335.

Unwin, P. N. T. and Henderson, R. (1975). Molecular structure determination by electron microscopy of unstained crystalline specimens. *J. Mol. Biol*. **94**:425–440.

Zampighi, G., Zampighi, L., Fain., N., Wright, E. M., Cantele, F. and Lanzavecchia, S. (2005). Conical tomography II: a method for the study of cellular organelles in thin sections. *J. Struct. Biol*. **151**:263–274.

Zwick, M. and Zeitler, E. (1973). Image reconstruction from projections. *Optik* **38**:550–565.

1

Sample Shrinkage and Radiation Damage of Plastic Sections

Pradeep K. Luther

Pradeep K. Luther • Imperial College London, Biomedical Sciences Division, Sir Alexander Fleming Building, Exhibition Road, London SW7 2AZ, UK

1. INTRODUCTION

Just as fossil insects embalmed in amber are extraordinarily preserved, so are biological samples that have been embedded in plastic for electron microscopy. The success of embedding samples in plastic lies in the astounding resilience of the sections in the electron microscope, albeit after initial changes. The electron microscope image results from projection of the sample density in the direction of the beam, i.e. through the depth of the section, and therefore is independent of physical changes in this direction. In contrast, the basis of electron tomography is the constancy of the physical state of the whole section during the time that different views at incremental tilt angle steps are recorded.

The shrinkage of a plastic section in each dimension, especially the depth, when viewed in the electron microscope, is now a well known phenomenon. Knowledge of the shrinkage behavior of a section of a sample embedded in a particular plastic is of crucial importance when embarking on the electron tomography of the sample. In the last 15 years, the most important advances in electron tomography have been the development of automated methods of recording tilt series and direct imaging onto CCD cameras (Koster *et al*., 1997; Koster and Barcena, Chapter 4 in this volume). These advances have enabled tremendous savings in labor but also in the total dose experienced by a sample. In this chapter, we review the studies carried out on shrinkage behavior of samples embedded in various resins and we review the protocols that have been followed by the leading proponents of electron tomography.

2. ON RADIATION DAMAGE

Several researchers have written reviews on the effects of the electron beam on biological samples (Egerton *et al*., 2004; Glaeser and Taylor, 1978; Grubb, 1974; Lamvik, 1991; Reimer, 1989; Stenn and Bahr, 1970). Electron microscope radiation has the primary effect of producing intense ionization in organic materials, which results in the formation of free radicals and ions. This causes bond scission and molecular fragments to be formed. These primary effects occur at all temperatures. At room temperature, the free radicals and molecular fragments can undergo diffusion, and produce cross-

linking or further chain scission. Damage to secondary structure occurs at an electron dose of <100 e/nm^2. Further exposure causes the tertiary structure to undergo dramatic reorganization following loss of specific groups and altered structural composition. The dominant effect finally is that of mass loss from the sample, which preferentially involves H and O in comparison with C and N. The mass loss is accompanied or followed closely by shrinkage of the sample normal to the beam.

We can reduce the radiation damage on an organic sample by cooling the sample to cryotemperatures and by using low-dose techniques (to be described later). One of the main effects of electron irradiation at conventional illumination levels is to cause specimen shrinkage normal to the plane of the sample. The effect of the shrinkage in reciprocal space is illustrated in Fig. 1. The example considered is a structure based on cubic

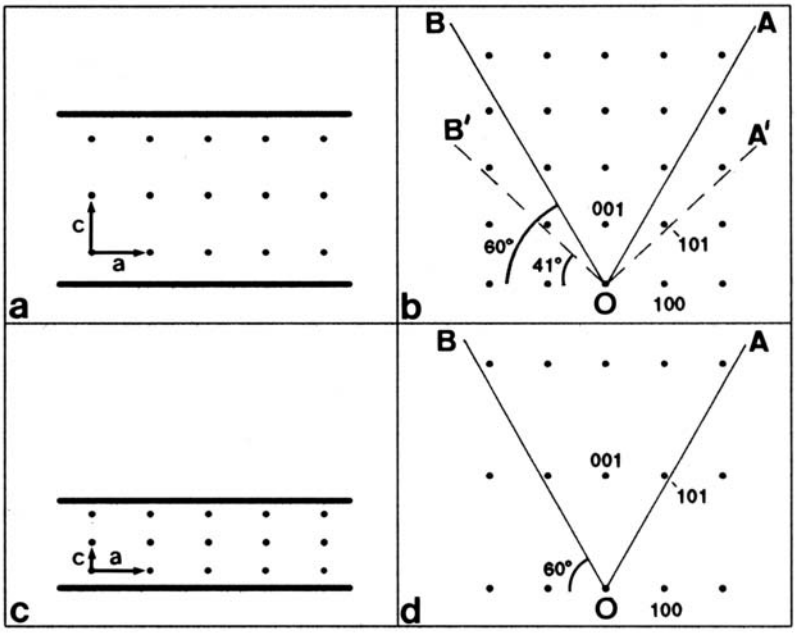


FIGURE 1. Illustration of the 'missing wedge' problem due to specimen shrinkage in the electron microscope. (a) Projection of a cubic crystal (for example) viewed edge-on with the *c* axis parallel to the electron beam. (b) Reciprocal lattice for the projection in (a). A conventional tilt holder in the electron microscope can be tilted in the range in the range of –60° to +60°. When a tilt series is recorded about a single tilt axis, and the 3D transform calculated by combining the individual transforms, data will be missing from the 3D transform inside the 'wedge' *AOB* of angle 60°. If the sample thickness reduces by 50% as in (c), then the corresponding reciprocal lattice shown in (d) is stretched 100% along *c**. The projection for the 101 diffraction spot, comfortably included in the tilt series for the unshrunk sample (a and b), now lies within the missing wedge. In relation to the original reciprocal lattice, the missing wedge is now described in (a) by *A′OB′* and has an angle of 98°. For 50% shrinkage, the tilt holder effectively covers only the range –41° to +41° in relation to the original sample.

symmetry (Fig. 1a), which collapses in thickness by 50% (Fig. 1c). With a conventional tilt holder, a series of views about a single tilt axis are recorded in the range –60° to +60°. The 3D transform in (Fig. 1b), obtained by combining the transforms of the individual views, has missing from it data in the 60° wedge *AOB*. In the case of the 50% collapsed sample (Fig. 1c), the reciprocal space is stretched by 100% in the corresponding direction (Fig. 1d). The volume of the 3D transform that can be sampled is now much reduced, e.g. the spot 101 present in (Fig. 1b) is missing from the transform (Fig. 1d). In relation to the original sample, the missing wedge (*A′OB′*) in the 3D transform has an angle of 98°. The effective tilt range is now only ±41°. Sample shrinkage therefore directly reduces the resolution normal to the sample plane in a tomogram of a plastic-embedded sample. Hence we must make every effort to curtail the shrinkage.

It is appropriate to describe the terminology of Amos *et al.* (1982) for the various imaging modes in the electron microscope and the electron dose involved for a single image in each case. A dose on the sample of 50–400 e/nm^2 is considered as a *very low dose*, which is appropriate for very high resolution studies of unstained crystalline specimens. A dose of ~1000–2000 e/nm^2 is termed a *minimal dose*, which is used for stained or non-crystalline specimens. Conventional microscopy for single images involves doses on the order of ~5000–50,000 e/nm^2 due to the time involved in searching and focusing. Minimal and low-dose methods require a search of suitable areas to be done at very low magnification, about ×2000, during which the dose should be extremely low (~2 e/nm^2). Modern electron microscopes provide low-dose imaging modes in which the focusing is done at high magnification on areas adjacent to the area of interest, followed by image recording at the desired magnification.

3. ON SAMPLE PREPARATION

3.1. Fast Freezing/Freeze Substitution

To produce plastic-embedded samples with the best possible preservation, rapid freezing followed by freeze substitution has been the most successful method. Various methods have been used for rapid freezing. Sosa *et al.* (1994) successfully froze muscle fibers by 'plunge-freezing', which involves rapid propulsion of the sample into a trough of cold cryogen, e.g. liquid ethane, at –180°C. 'Slam-freezing' involves rapidly propelling a sample onto a highly polished metal block (usually copper) cooled by liquid helium or liquid nitrogen. Spectacular fast-freeze deep-etch replica images of various samples were obtained in the pioneering studies by Heuser (Heuser, 1989; Heuser *et al.*, 1987). Details of the slam-freezing method applied to striated muscle fibres to capture different activity states have been described (Craig *et al.*, 1992; Hirose *et al.*, 1993; Liu *et al.*, 2004a; Padron *et al.*, 1988).

The currently preferred rapid freezing method for plastic-embedded specimens is high-pressure freezing. The two most commonly used instruments for this technique are the EMPACT made by Leica (www.leica-microsystems.com; Vienna, Austria) and the HPM 010 made by BAL-TEC-RMC (www.baltec-rmc.com). Excellent results have been obtained with these machines (Giddings *et al*., 2001). While the depth of good freeze is very shallow with plunge-freezing, ~5 μm or less, and is a little better with slam-freezing, ~10–30 μm, it is improved considerably, ~100 μm, with high-pressure freezing (Shimoni and Muller, 1998; Studer *et al*., 2001).

3.2. Conventional Processing

Considerable effort is often required to perfect the rapid freezing technique for some samples, and conventionally prepared samples may have to be used for the first investigations. The good news is that valuable insight can be obtained by processing a sample by conventional methods. This usually comprises fixation in glutaraldehyde followed by osmium tetroxide, dehydration in a solvent such as acetone or ethanol, infiltration and embedding in a selected plastic resin. The details of these standard protocols are available in various manuals of electron microscopy techniques (Glauert, 1998; Hayat, 2000). In a recent tomographic study, Harlow *et al.* (2001) used these methods to study frog neuromuscular junction. They prepared 50 nm thick sections of the Epon-embedded sample and recorded a 1° and 2° step tilt series while the section was cooled to cryogenic temperatures. From the tomograms, Harlow *et al.* produced a detailed account of active zone material that showed the arrangement and associations of structural components of the organelle.

4. METHODS OF MEASURING SAMPLE THICKNESS

Knowledge of the starting thickness of a sample is essential for carrying out 3D reconstructions and their interpretation. Knowledge of the changes that occur to the sample dimensions during electron microscopy is also just as important. Methods to measure the initial thickness of a sample and changes that occur following the microscopy are described in this section.

4.1. Methods Not Involving the Electron Microscope

4.1.1. Ultramicrotome Advance Setting

For sections, the advance setting on the ultramicrotome is potentially a good estimate of their thickness. The microtome setting may apply especially to sections belonging to a well-cut ribbon in which the sections have uniform interference color. However, past research has shown that there is

considerable variation in section thickness within a ribbon. Gunning and Hardham (1977) used a Reichert OMU3 ultramicrotome to cut ribbons of sections from a Spurr's resin block. By monitoring the thickness of the individual sections by interference microscopy (next section), they found a 33% variation in the thickness of the sections about the mean value. Ohno (1980) evaluated the section thickness in relation to the microtome setting using sections of Epon-embedded rat kidney. By cutting cross-sections (see later), he found a variation of 20% for sections of thickness <100 nm. For thicker sections ranging from 0.2 to 0.9 μm, the thickness was found to be closer to the microtome setting, with a relative variation of about 7%. It is likely that modern ultramicrotomes such as the Leica Ultracut UC6 (www.leica-microsystems.com) and the RMC PowerTome series (www.baltec-rmc.com), used in combination with good knives, can cut ribbons of uniform thickness sections.

4.1.2. Interference Color Scale

The thickness of sections can be estimated from the interference color of the sections as they float in the knife trough. The thickness scale of interference colors proposed by Peachey in 1958 is probably the most widely used. His measurements were done on methacrylate sections using an ellipsometer. Several groups have reported sightly differing scales (e.g. Williams and Meek, 1966). The author has found the thickness scale of Yang and Shea (1975), based on Epon–Araldite sections and measured by the resectioning technique (see later), to be more useful, especially for the important range <100 nm (Table 1).

In general, the interference color of a section is only used as a rough guide to its thickness. This is probably because, with current methods, thickness estimation from the observed color is subjective. It is the author's view that more quantitative estimates can be obtained from the interference colors. For this to be possible, two provisions are essential: (i) an accurately reproduced color scale; and (ii) a standard light source. Since different light

TABLE 1. The interference color scale of plastic sections as a measure of section thickness from the work of Peachey (1958) and Yang and Shea (1975)

Interference	Color Peachey (1958)	Thickness (nm) Yang and Shea (1975)
Gray	<60	<40
Silver	60–90	40–60
Yellowish-silver		60–67
Pale gold		78–90
Gold	90–150	90–100
Dark gold		100–110

sources have different spectral properties, thus giving different color biases, a standard light source used to illuminate the sections and the color scale simultaneously would provide a standard scale. In the photographic industry, color-corrected lamps are commonly used.

The interference color of a section becomes even more valuable when it is realized that the color of a section mounted on an uncoated grid, viewed subsequent to the microtomy, is in fact the same color as that observed for the section when it was floating on the trough during the microtomy. The physical basis for the color to be the same in the two situations is as follows. The color observed is due to the interference of the light reflected off the top and bottom surfaces of the section. When light reflects from a surface of higher refractive index, there is phase change of $\pi/2°$. This change occurs only for the reflection off the top surface of a section. Reflections off the bottom surface do not produce a phase change since the refractive index of a resin section (1.54) is greater than that of water (1.3) for the situation when the section is on the trough, and of air (1.0) when a section is viewed subsequent to the microtomy. There appears to be little effect of any absorbed water or stain. The above does not apply when plastic sections are mounted on grids coated with formvar or formvar/carbon, as the additional thickness causes a change in the interference color. Coated slot grids are required for electron tomography of serial sections.

4.1.3 Interference Microscopy

An important method of determining the thickness of thin transparent samples uses the transmitted light interference microscope. The technique has been used in several studies on the measurement of section thickness (Cosslett, 1960; Gillis and Wibo, 1971; Gunning and Hardham, 1977). Unfortunately, interference microscopes do not appear to be manufactured commercially any more. However, they are much cherished and can be tracked down in a few microscopy or physics research laboratories.

The use of an interference microscope fitted with a Jamin–Lebedeff interference system has been described by Gillis and Wibo (1971) and Spencer (1982). In this microscope (Fig. 2a), a polarized beam of light is divided into two. The measuring beam passes through the transparent sample while the reference beam passes through the reference medium. A small path difference in the measuring beam is introduced by the sample, which can be measured from the shift in the interference fringes obtained on combining the two beams (Fig. 2b). The thickness t is calculated from the fractional shift $\Delta x/x$ of the fringes and the wavelength of the light λ employed:

$$(\mu - n)t = \Delta x/x.\lambda$$

where μ is the refractive index of the section, about 1.54 for epoxy resins, and n is the refractive index of the reference medium (= 1 for air).

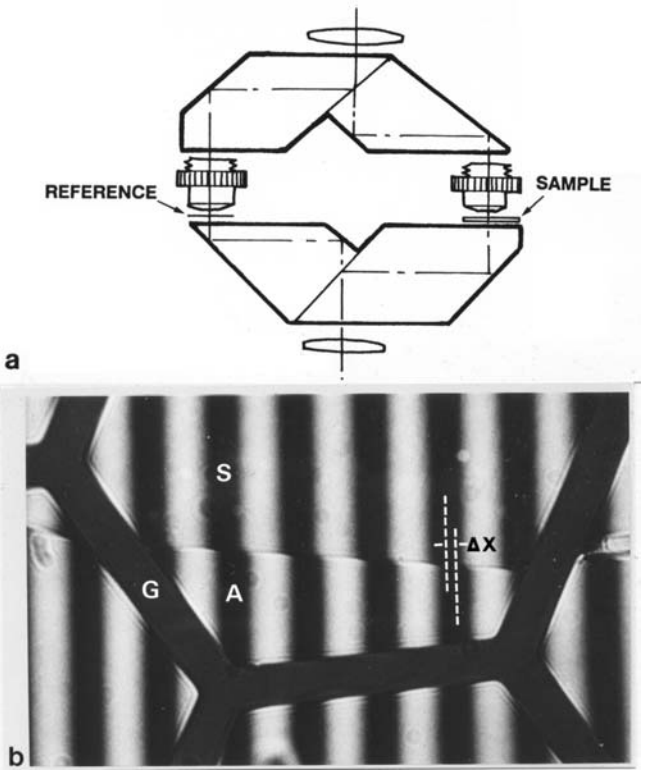


FIGURE 2. (a) Schematic representation of the Mach–Zender interference microscope. In this microscope, the incident light is divided into two beams which pass through the test sample and the reference. (b) Fringe pattern across a section and grid in the interference microscope. A section (S) mounted on a grid (G) causes a shift Δx in the fringe pattern relative to that in air (A).

Figure 2b shows an example of the fringe pattern as viewed in a Mach–Zender interference microscope. The shift in the pattern is caused by a section mounted on a grid. The relative shift in the fringe pattern needs to be determined as accurately as possible. Using digitized images, the fringe patterns can be integrated separately over the section and over air, and the shift found by correlating the two arrays. Cosslett (1960) and Gillis and Wibo (1971) estimated that the error involved is ~2% when 10 or more measurements are done.

4.2. Methods of Thickness Measurement by Electron Microscopy

4.2.1. Folds

One of the simplest methods of estimating the thickness of a section during electron microscopy is by measurement of the width of any folds

that may occur in the section (Small, 1968). Folds or crimps sometimes occur during electron microscopy on sections that are not well stuck to the grid coating. They frequently occur in sections that are mounted on coated grids. The minimum width of the fold is equal to twice the section thickness.

4.2.2. Thickness Measurement from Re-sectioned Sections

Probably the most direct method of measuring the section thickness is by re-embedding the section, cutting cross-sections and viewing the cross-sections in the electron microscope. This is a destructive method and there is the dilemma of cutting cross-sections from unviewed samples or cutting cross-sections from radiation-damaged samples after electron microscopy. Bedi (1987) has suggested the following solution for resin sections. After trimming a block face, a small score is applied to the face. Upon sectioning, two ribbons are obtained, one of which can be used for experimental viewing in the electron microscope and the other for measuring thickness by the above method.

Cutting cross-sections at any random position in a sample is straightforward. However, if one wants to cut cross-sections across a region precisely identified by electron microscopy, for example across the region where tilt views were obtained, then the meticulous methods developed by Jesior (1982) are to be recommended. Although most of his work was done on negatively stained crystalline material, the method applies just as well to positively stained resin sections. Jesior's method requires application of latex particles along with the test material onto a coated finder grid. The size of the latex particles, 1.3 μm, is chosen to enable viewing at both the electron microscope and light microscope levels. The method is illustrated in Fig. 3. In the electron microscope, the grid is viewed and the path of the cross-section required is noted relative to the grid holes. At the same time, the position of the required object is measured carefully, in terms of its distance and angle relative to the edges of the grid hole, and also the latex particles in the vicinity are mapped. The whole grid is then re-embedded in epoxy resin. The block is trimmed with a glass knife. Once the grid bars surrounding the required object are exposed, the block face and the sloping sides are cut. The required region is then approached by first cutting with the microtome a precise number of thick sections. With the block face devoid of grid bars, a diamond knife can then be used safely for cutting the final sections. Sections from material stained prior to embedding do not require restaining.

4.2.3. Thickness Measurement from Internal Periodicity—Bennett Method

An elegant method devised by Bennett (1974) exploits the internal periodicity of a sample to measure the effect of electron irradiation on the

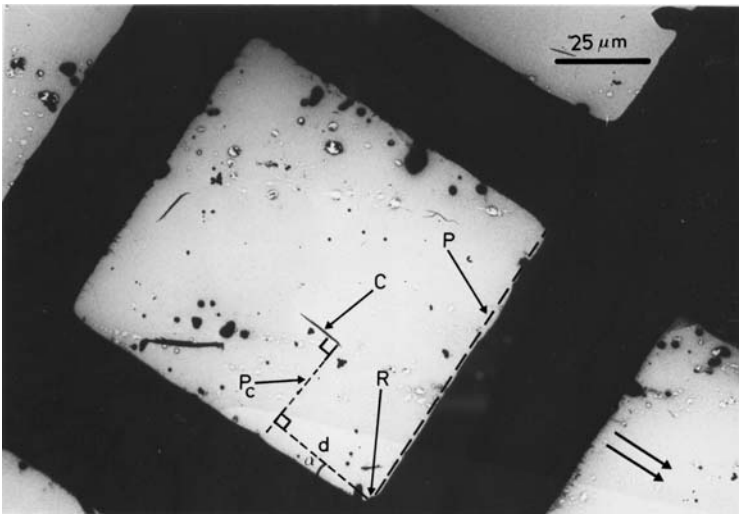


FIGURE 3. Method for cutting cross-sections of a specimen along a path precisely defined by electron microscopy (Jesior, 1982). For the chosen sample, in this case a fribronogen crystal (labeled C), the surrounding latex particles are mapped and the reference plane *P* and distance *d* of the reference plane to the crystal noted. After embedding, the block is trimmed and then oriented in the microtome so that the principal plane is parallel to the knife. The required area is approached by cutting several thick sections. (From Jesior, 1982; reproduced with permission by Elsevier).

relative section thickness. Bennett used a centrifuged pellet of paracrystals of light meromyosin (LMM), which is the α-helical rod part of a myosin molecule, fixed and embedded in Araldite. LMM paracrystals applied directly to a coated grid, negatively stained, and viewed in the electron microscope show a strong banding pattern with a periodicity of 43 nm along their length. As the paracrystals are rod shaped, they tend to lie flat on the coated grids. In the embedded pellet, the paracrystals have a variety of orientations. Consequently, in a section cut from the block, the majority of the paracrystals do not lie in the plane of the section, and hence they do not show any banding pattern. Selecting a paracrystal and tilting the section perpendicular to the long axis of the selected paracrystal brings the banding pattern into view. Only a few paracrystals fortuitously oriented in the plane of the section show a sharp banding pattern in the untilted section. Bennett found that the periodicity measured from the in-plane paracrystals was different from that measured in paracrystals which required tilting of the section. The paracrystals which required tilting had a smaller periodicity than that predicted by the geometry. By assuming a shrinkage to a fraction *f* of the original thickness, she derived the relationship between tilt angle φ required to view the banding pattern sharply and the true periodicity *p* and measured periodicity *q* (Fig. 4):

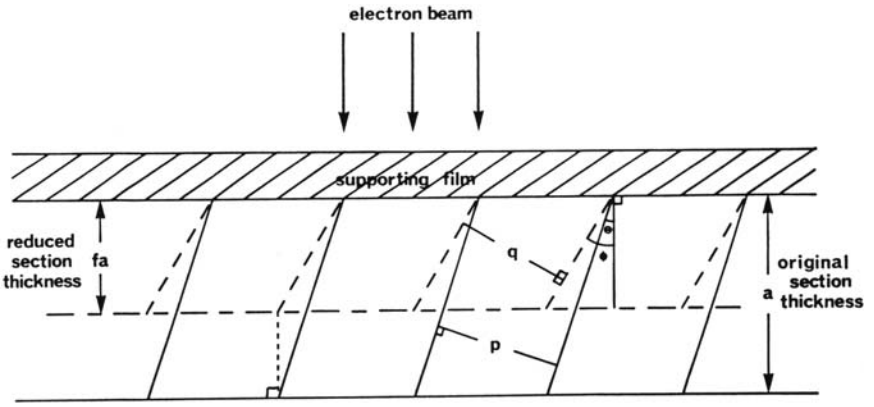


FIGURE 4. Measurement of the relative shrinkage in periodic objects (Bennett, 1974). This method is suitable for samples which have 3D order or 2D order out of the plane of the sample; i.e. the sample has to be tilted in the electron microscope to visualize the repeating pattern. For a sample with original periodicity *p*, the electron beam causes a reduction in thickness by a factor *f* and a new smaller periodicity *q*. Consequently, the sample has to be tilted to a larger angle φ to visualize the banding pattern. The fact that the banding pattern for paramyosin crystals as used by Bennett (1974), and tropomyosin paracrystals as used by Luther *et al.* (1988) can be visualized sharply at the reduced thickness state shows that the shrinkage occurs uniformly through the depth of the sample. The reduction in thickness does not occur due to, for example, etching from each of the surfaces because that would require the periodicity to remain unchanged. (From Bennett, 1974; reproduced with permission by the Company of Biologists Ltd.).

$$q^2 = p^2 - p^2(1 - f^2)\sin^2\varphi$$

Plotting q^2 versus $\sin^2\varphi$ allows the relative thickness and periodicity to be found. Luther *et al.* (1988) modified the above equation by introducing parameters f_x and f_z for the fractional in-plane and depth shrinkage:

$$q^2 = p^2 f_x^2 - p^2(f_x^2 - f_z^2)\sin^2\varphi$$

A serious limitation of the method is that a variable amount of time is required in selecting suitable paracrystals and then in tilting the section until the banding pattern is sharp. The high accumulated dose means that this method is more suited for the measurement of the final thickness in conventional dose electron microscopy. Bennett noted a 50% reduction in the thickness of Araldite sections.

This method gives changes in thickness of a section, not the absolute value. However, the principle involved, that of measuring the relative thickness of a section from knowledge of the periodicity, is widely applicable for samples with crystalline or paracrystalline order, as shown later for cryosections of striated muscle (Sjostrom *et al.*, 1991).

4.2.4. Electron Diffraction

Berriman and Leonard (1986) have used a method first proposed by Dorset and Parsons (1975) for using the electron diffraction pattern to measure the thickness of crystals several unit cells thick. To understand this method, note that the 3D Fourier transform of a crystal one unit cell thick consists of a reciprocal lattice of spikes perpendicular to the a^*, b^* plane. The diffraction pattern of such a crystal tilted to θ° samples the 3D transform along the plane tilted to the same angle (Fig. 5a). By tilting the crystal at various angles, the 3D transform is sampled at different planes and information is built up for 3D reconstruction. In the case of a crystal more than one unit cell thick, the spikes in the 3D transform are broken up into layers or zones (Fig. 5b). The width of these Laue zones is related to the thickness of the crystal, and the separation of the zones is related to the unit cell spacing $\underline{c}$. From the tilt angle, $\underline{c}$ can be calculated geometrically from the separation of the zones. The advantage of this method is that it does not require any specific orientation of the crystal relative to the beam, only a knowledge of the tilt angle. Hence it allows low-dose experiments to be performed.

4.2.5. Use of Gold Particles

The application of colloidal gold particles to a sample for determining its thickness was first investigated by two groups: Berriman *et al.* (1984) and Luther *et al.* (1988). The basic geometry involved in the latter method is described here.

Figure 6 shows a stereo pair of electron micrographs of a section labeled with gold particles on both surfaces. By using gold particles of different diameters (5 and 15 nm, respectively) to mark each surface, areas labeled simultaneously on the top and bottom surfaces can be identified while viewing in the electron microscope. The geometry of the method, illustrated in Figure 7, is derived from the edge-on view of a section tilted to an angle β. Particles 1 and 2 mark one surface, and particle 3 marks the other surface. They are imaged on the micrograph as $1'$, $2'$ and $3'$. The coordinate system for the section has the y-axis parallel to the tilt axis, the x-axis in the plane of the section and the z-axis normal to the plane. Coordinates u' and v' are measured on the micrograph, with the u'-axis parallel to the x-axis. If the shrinkage in the plane is m_x and that normal to the plane is m_z, then:

$$u = m_x\, x \cos\beta + m_z\, z \sin\beta \tag{1}$$

$$v = m_x\, y \tag{2}$$

Shrinkage in the plane of the section m_x can be found from (2), and then the shrinkage m_z normal to the plane can be found.

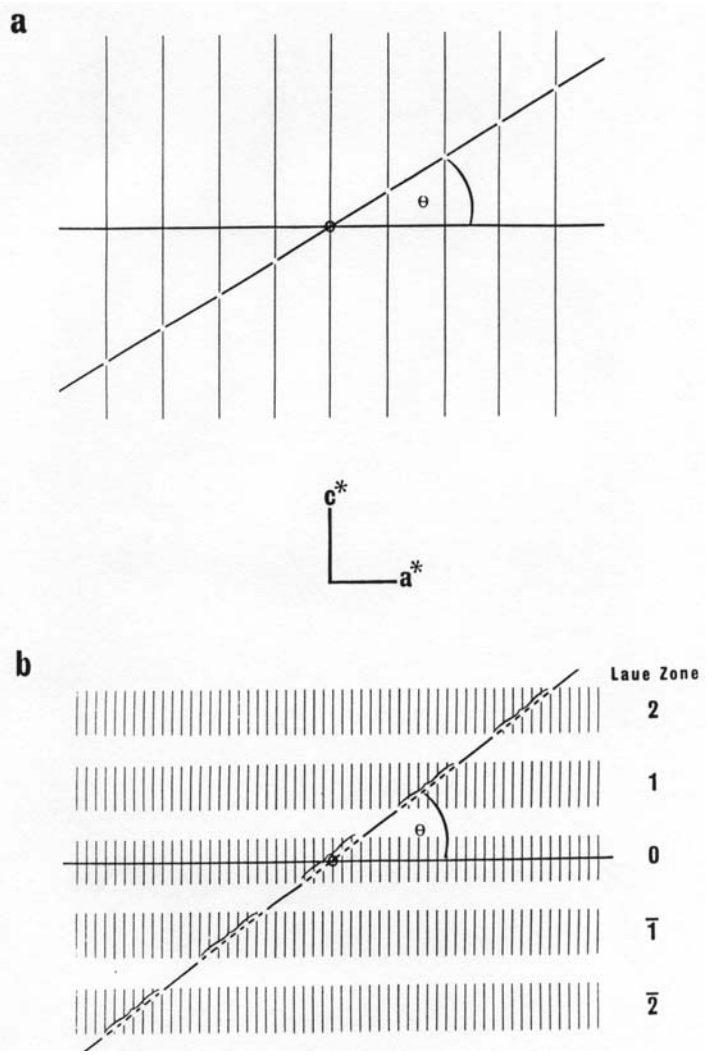


FIGURE 5. Electron diffraction method for measuring lattice spacing and crystal thickness for a thin crystal (Berriman and Leonard, 1986; Dorset and Parsons, 1975). In this figure, the $\underline{c}$ axis is normal to the plane of the crystal. (a) For a crystal one unit cell thick, the diffraction intensities are continuous spikes which can be sampled by tilting at any angle. (b) For a crystal a few unit cells thick, the diffraction intensities lie along bands or zones with a periodicity equal to $\underline{c}^*$. The thickness of the crystal can be calculated from the width of the bands (Dorset and Parsons, 1975). This method allows relative thickness measurements (from $\underline{c}$) to be found using low dose conditions. (From Berriman and Leonard, 1986; reproduced with permission by Elsevier).

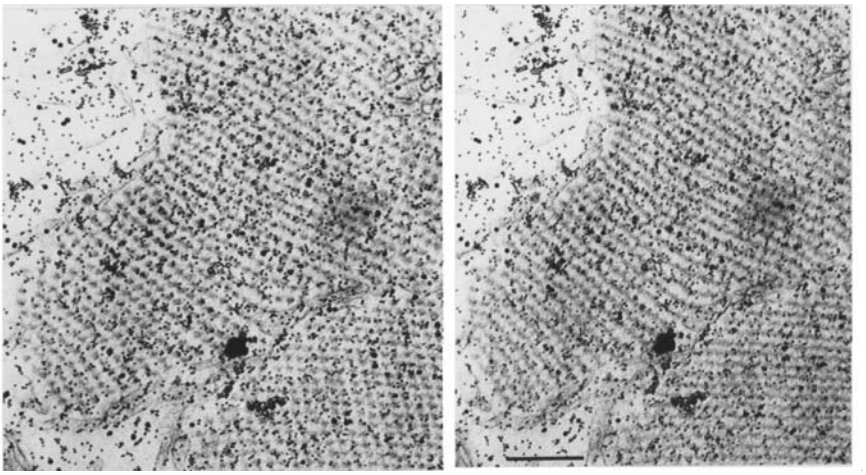


FIGURE 6. Thickness determination by labeling with gold particles. Stereo view of a section marked with gold particles. The two surfaces of the section are labeled with gold particles of different sizes to enable identification in the electron microscope of areas labeled on both sides. Scale bar 200 nm.

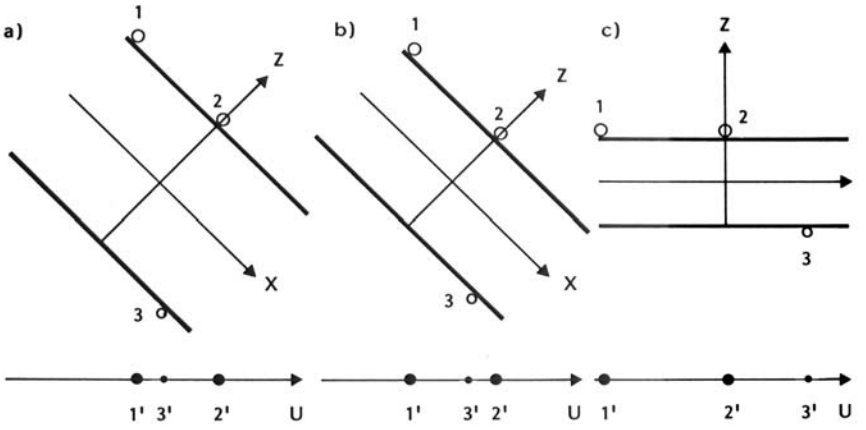


FIGURE 7. Geometry of a section labeled with gold particles viewed edge-on to measure section thickness and collapse. Gold particles 1 and 2 mark one of the surfaces and particle 3 (larger) marks the other. (a) Section tilted to 45°. In the micrograph, the images of the three particles occur at 1′, 2′ and 3′. (b) Subsequent state after section collapse. In the micrograph, 3′ has moved relative to 1′ and 2′. (c) Final image taken at 0° tilt. (From Luther *et al.*, 1988; reproduced with permission by Elsevier).

At the start of the experiment, the sample is tilted to 45° in the electron microscope, a suitable region found at very low magnification, and the microscope focused at an adjacent area. With the selected region exposed continuously to a fixed beam intensity, a series of micrographs are recorded at different times to sample the effects of the irradiation. The effects of very low doses can be investigated since the minimum potential dose is the amount required to record one micrograph. Note that the applied gold particles can obscure valuable structural details. Therefore, the particles must be applied carefully to ensure suitable distribution on each surface. The method used by the author is to float a grid on to a drop of the colloidal gold solution for ~30 s and then blot the grid lightly and allow it to air-dry. An excellent method of preparing colloidal gold particles of different sizes is described by Slot and Geuze (1985). To improve the attachment of the gold particles to the section, 0.02% polylysine may be applied to each surface of the grid (Muller *et al.*, 2000).

5. STUDIES ON SHRINKAGE OF PLASTIC SECTIONS

5.1. Early Studies

Cosslett (1960) carried out pioneering work on the effect of the electron beam on sections, for which she used an interference microscope to measure the section thickness before and after the electron irradiation. She used two different media for the reference beam which allowed her to measure changes in both the thickness and the refractive index of the section. The resins examined were methacrylate, Araldite, Vestopal and Aequon (water-miscible medium). For methacrylate, the effect of the beam was the greatest: the sections reduced in thickness by 50% of the starting value. Sections of the other resins reduced in thickness by 20–30%. In each case, Cosslett noted that the refractive index increased from 1.54 to 1.9. The increase in refractive index was attributed to changes in internal structure of the resins. The work of Bennett (1974), using sections of LMM paracrystals embedded in Araldite, showed a reduction in section thickness by 50% following electron microscopy at conventional doses.

5.2. Characterization of Shrinkage of Plastic Sections

There are now a few detailed studies of the variation in thickness of plastic sections with electron dose (Braunfeld *et al.*, 1994; Kremer *et al.*, 1990; Luther *et al.*, 1988; Trachtenberg *et al.*, 2000). The experiments of Luther *et al.* (1988) using the gold particles method are described here (see Section 4.3.5). Sections of tropomyosin paracrystals embedded in Araldite were used. Their results are shown in Fig. 8. In each of the four panels, the

upper trace shows the shrinkage measured in the plane of the section and the lower trace shows the shrinkage measured in the depth. There are two parts to each curve: at the start of the experiment, low dose rates were used, in the order of 50 e/nm^2/s, then the dose was increased to 400 e/nm^2/s. Despite the low dose rates used at a magnification of 20,000×, 'strict' minimal dose conditions could not be used to record the image of a pristine section, because the beam, although weak, caused violent planar movements in the section and occasionally breakage of the section. Therefore, the sample was first viewed at very low magnification and then the selected area was translated into the beam by hand. There were still gross movements within the irradiated region and, when the first image was recorded after 15 s of irradiation (allowing the section to stabilize), the section had already experienced a dose of 700 e/nm^2. The results in Fig. 8 show that there is sudden collapse to 70% within 3 min (after an accumulated dose of 9000 e/nm^2) and then the shrinkage levels off. The collapse occurred in a similar fashion in areas with resin only (Fig. 8a–c) and in areas with embedded tissue regions (Fig. 8d). Increasing the dose from 50 to 400 e/nm^2/s (after 20 min in (Fig. 8a) and 15 min in (Fig. 8b, c and d)) resulted in a new shrinkage curve which leveled off at a final thickness of 60% of the original. The fact that the curves level off after each set dose level indicates that the changes that occur in a sample allow it to dissipate the energy at the set dose rates. *Hence the changes that occur are not just dependent on the accumulated dose*. This implies that the electron microscopy should be carried out at the lowest dose rate possible with a particular sample. The shrinkage in the plane of the section following irradiation is much lower: there is a small shrinkage by 5% at the lower dose level and a further 5% at the higher dose level.

Luther *et al.* (1988) also discussed the 'clearing' that is observed with a stained section after very low doses (Fig. 9). The effect seen is that of enhanced contrast of the stained material due to the increased transparency of the surrounding resin. To demonstrate the effect, a small disk within the field of view was irradiated with a dose of 1 e/nm^2 at low magnification (×3000). Then the beam was spread to cover a larger region. With the resulting reduced illumination, the exposed disk and the immediate surrounding area were photographed. The process of photography causes the surrounding area to be slightly irradiated as well. In the micrographs, the main exposed region showed a marked improvement in contrast of the stained material compared with the slightly exposed surrounding area. For the total dose received, ~15 e/nm^2, the shrinkage curves in Fig. 8 imply that no significant shrinkage has occurred. Hence Luther *et al.* concluded that clearing could be due to mass loss or some change in chemical structure which causes less electron scattering, and that the efffect must precede the depth shrinkage. Braunfeld *et al.* (1994) investigated this phenomenon and concluded that the change in beam intensity was directly related to mass loss for electron doses up to ~1.5×10^4 e/nm^2.

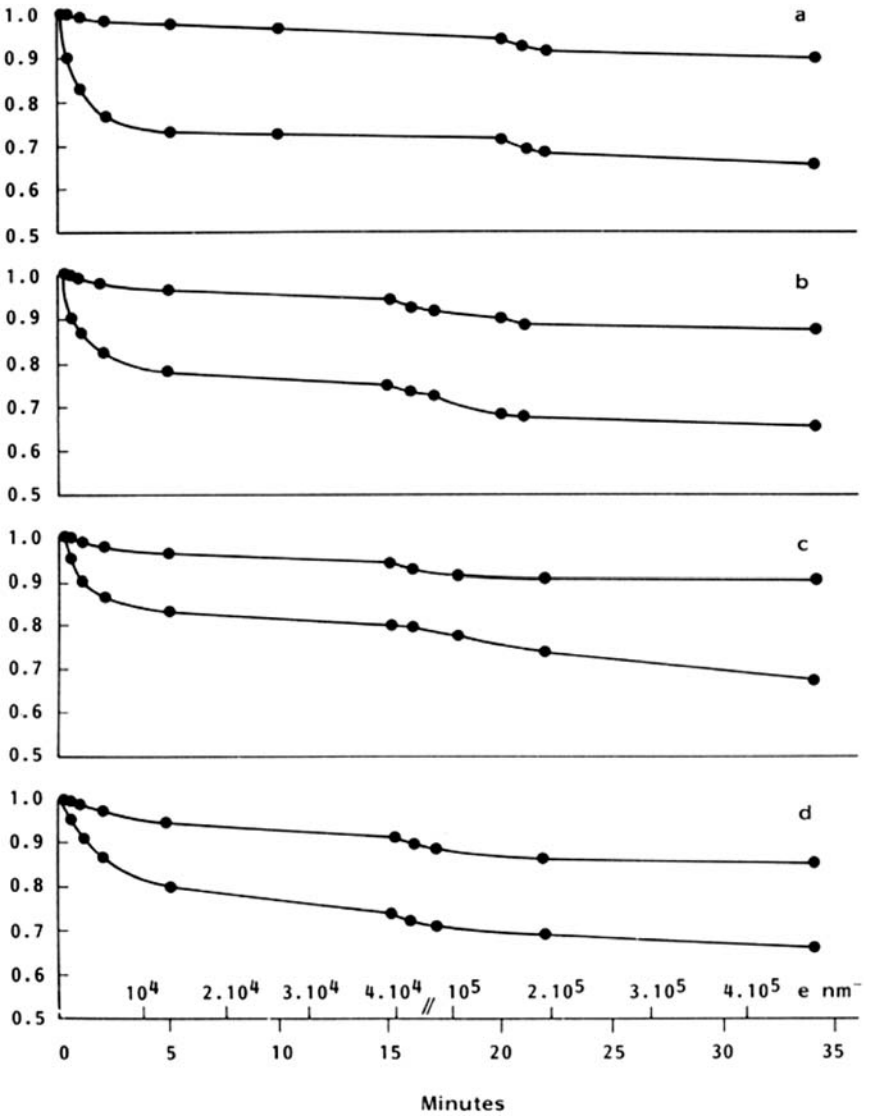


FIGURE 8. Measurement of section collapse (lower trace) and planar shrinkage (upper trace) as a function of electron dose. (a, b and c) Gold particles over resin only (Araldite). (d) Gold particles over a paracrystal. The dose rate, initially 0.5 e/Å^2/s, was increased to 4 e/Å^2/s after 20 min for (a) and after 15 min for (b, c and d). (From Luther *et al.*, 1988; reproduced with permission by Elsevier).

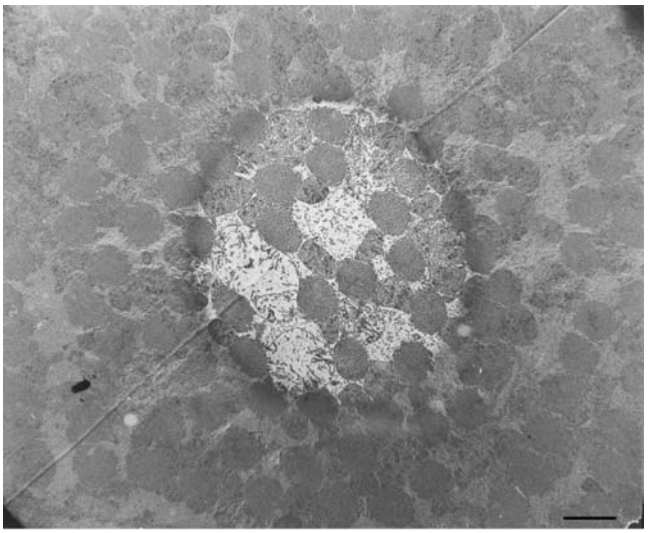


FIGURE 9. The 'clearing' effect as observed in an Araldite section of *Drosophila* flight muscle. The electron dose on the 'cleared' central patch is 0.1 e/Å^2 compared with the much lower dose (0.02 e/Å^2) for photography of the surrounds. Scale bar represents 2 microns. (From Luther *et al.*, 1988; reproduced with permission by Elsevier).

5.3. Investigation of Parameters That May Reduce Section Shrinkage

The Holy Grail for electron tomography of plastic-embedded samples is finding the resin that is most resistant to electron radiation and that experiences the least shrinkage. Only a few resins have been quantitatively tested for their shrinkage behavior in the electron microscope. Some of these studies include: Luther *et al.* (1988), Araldite; Kremer *et al.* (1990), Epox 812/Araldite and LX112; Braunfeld *et al.* (1994), Epon, Epox and Lowicryl; van Marle *et al*. (1995), Epon and Lowicryl; and Trachtenberg *et al*. (2000), Spurr's resin. Hence epoxy resins are the most widely studied, with only two studies on Lowicryl, an acrylic resin. These studies showed that for each resin tested, the response to electron dose has the same characteristic form: an early rapid phase of shrinkage followed by a slow period of thinning. The amount of *z*-shrinkage and planar shrinkage was found to be similar in all of these studies (although slightly less in Lowicryl (Braunfeld, 1994)), hence there is little to choose between the different resins. Most of the recent tomographic studies have been done using epoxy resins. Trachtenberg *et al.* (2000) reported a higher than normal *z*-shrinkage of 45% in their Spurr's-embedded samples. Spurr's resin, an epoxy resin, is preferred by some researchers on account of its very low viscosity.

Kremer *et al*. (1990) and Braunfeld *et al.* (1994) investigated the hypothesis that normally polymerized plastic blocks further polymerized in

a microwave oven are more stable in the electron microscope and suffer less shrinkage. Unfortunately, they found little improvement in the shrinkage properties of sections from the microwave-treated blocks.

Another factor that could have a protective effect on the section shrinkage is higher accelerating voltage electron microscopy (HVEM). This was investigated by Kremer *et al.* (1990), who used a 1 MV electron microscope. They observed shrinkage curves similar to those observed by conventional transmission electron microscopy (TEM) (Luther *et al.*, 1988) and concluded that shrinkage in x–y and z did not seem to be reduced in HVEM.

5.4. Shrinkage Studies of Native Samples (Not Embedded in Plastic)

Jesior and Wade (1987) examined the shrinkage in negatively stained 2D crystalline arrays of the bladder membrane, following conventional dose electron microscopy with a total dose of 10^5 e/nm^2. By embedding the sample and cutting cross-sections, they found that the membrane had 'flattened' to 60% of the native thickness. Berriman and Leonard (1986) used electron diffraction to study the variation in thickness of negatively stained thin crystals of catalase as a function of electron dose. From the electron diffraction pattern, the separation of the Laue zones was measured, and hence the $\underline{c}$ spacing of the unit cell. As shown in trace 5 of Fig. 10, the $\underline{c}$ spacing reduced as the logarithm of the accumulated electron dose. After a dose of 10^5 e/nm^2, the crystals had reduced in thickness to 60% of the starting value. Berriman and Leonard also studied how samples that are cooled to liquid nitrogen temperatures are affected in the electron microscope, and these results will be described in a later section.

The effect of electron irradiation on negatively stained, dried cryosections at room temperature was investigated by Sjostrom *et al.* (1991). The material used was longitudinal cryosections of fish muscle, negatively stained in ammonium molybdate. In the ideal case, tilting such sections about the myofibril axis, to view in turn the [10] and [01] projections of the myosin filament hexagonal array, would require a total tilt of 60°. However Sjostrom *et al.* found that the angle of tilt required to observe the two views was ~120°. From this, they concluded that the sample had collapsed to nearly 33% of the original thickness. Some of this collapse could occur during drying the negatively stained cryosection prior to the electron microscopy.

5.5. Shrinkage Occurs Along Directions Lacking Restraint

Does the shrinkage in a section occur in any preferred direction when irradiated in the electron microscope? Bennett (1974) proposed that shrinkage in z occurs in the electron microscope because this is the only

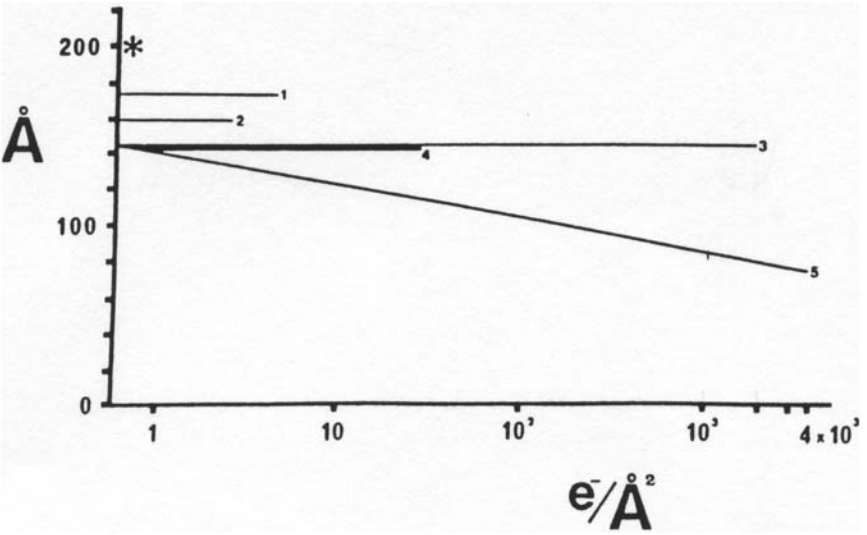


FIGURE 10. Investigation of the *c* unit cell size of thin catalase crystals (hence crystal thickness) with accumulated dose under different experimental conditions. The asterisk indicates the value (206 Å) found by X-ray diffraction. (1) Frozen-hydrated crystals at 120 K. (2) Glucose-embedded at room temperature. (3) Dried, negatively stained crystals in uranyl acetate at 120 K. (4) Negatively stained at 120 K with a contaminating layer of ice or deposited carbon layer. (5) Negatively stained at room temperature. (From Berriman and Leonard, 1986; reproduced with permission by Elsevier).

dimension of the section which is not supported on each side. In the plane of the section, shrinkage is restrained by attachment of the section to the grid bars or support film. Berriman and Leonard (1986) have provided strong evidence for this idea (Fig. 11). They applied catalase crystals to a holey carbon-coated grid and viewed carefully those crystals that partly covered the holes. When viewed at very low illumination levels (100 e/nm^2 accumulated dose), crystal edges were continuous across the hole and the support film. With increased accumulated dose (10,000 e/nm^2), the part of the crystal across the hole contracted in the plane of the grid such that the edge of the crystals was drawn towards the rest of the crystal. The amount of contraction was estimated by viewing the lattice spacings. Unwin (1974) has shown that parallel with the dimensional changes, there are changes in the images and diffraction patterns of stained samples, indicating that radiation causes movement in the stain. Berriman and Leonard (1986) suggest that this redistribution of stain may be due to the stain moving along with the parts of the sample which are free to contract in the electron beam.

Similar shrinkage occurs in the part of a plastic section that does not cover a grid hole completely. In the electron beam, the part that is not attached to the grid bars starts to contract just like the catalase crystal

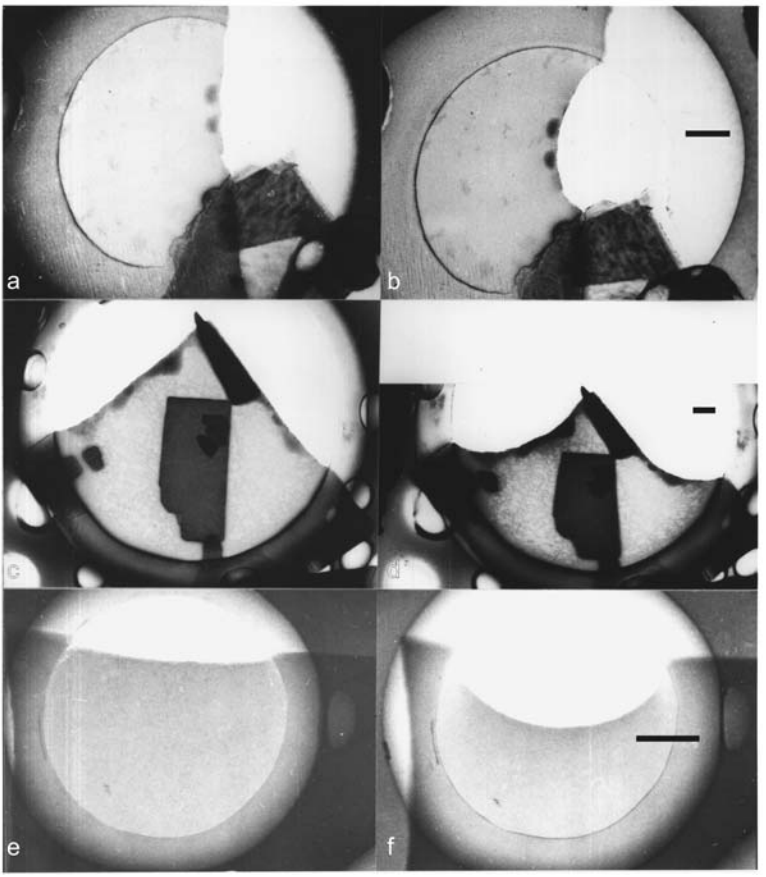


FIGURE 11. Demonstration of sample shrinkage in the electron microscope occurring only along directions lacking physical constraints. The figure shows images of catalase crystals spanning holes in the supporting film, viewed under low dose, 1 e/Å^2 (left panel), and high dose, 100 e/Å^2 (right panel). In each case, the straight edge of the catalse crystals imaged in low-dose conditions becomes rounded with accumulated dose. With the points of attachment remaining unaffected, the roundedness is due to shrinkage of the crystal sheets perpendicular to the edges. Scale bars represent 0.1 μm. (From Berriman and Leonard, 1986; reproduced with permission by Elsevier).

sheets shown in Fig. 11. Hence, for tomography, it is important to select a section that covers a grid hole completely and is well attached to the grid bars.

5.6. Is Shrinkage Uniform in Z?

For a sample to become thinner in the electron microscope, two ways can be envisaged: collapse (shrinkage) or loss of mass from the surface

(etching). In electron tomography studies on sectioned samples, it is now customary to stretch the final 3D reconstruction along z to compensate for this. The assumption is that thinning is caused by shrinkage, and that shrinkage is uniform through the depth of the sample. How can we determine whether the thinning is due to collapse of the section or due to material loss near the surfaces? In the section shrinkage studies of Bennett (1974) and Luther *et al.* (1988) using paracrystals, they found that when the randomly oriented paracrystals were tilted to view the banding pattern, the patterns were sharp across the whole of a selected paracrystal at particular tilt angles. This implicates uniform collapse in z as the origin of the thinning. If the collapse was not uniform in z, then only part of a particular paracrystal would show sharp banding at a particular angle.

The work of van Marle *et al*. (1995) questions the premise of uniform z-shrinkage in sections. They prepared 100 and 200 nm sections of Epon-embedded testis material from rats that was conventionally processed, mounted on formvar-coated grids, and then carbon coated. Resectioning the unexposed sections and viewing the cross-sections in the electron microscope convinced these authors that the sample material was distributed homogeneously throughout the depth of the section. However, after electron microscopy and production of tomograms of the samples, they found that xz projections comprised quite inhomogeneous density distributions through the depth, such that the greatest density was near the surface of the section on the vacuum side, followed by the region close to the formvar surface. van Marle *et al.* also found that there was more shrinkage for a sample cryocooled in the electron microscope than one cooled prior to insertion, implying that the microscope vacuum may also play a part in the shrinkage. These paradoxical results need to be investigated further as they imply great impact on the practice and interpretation of electron tomography.

6. EFFECT OF LOW TEMPERATURES ON RADIATION DAMAGE AND SAMPLE SHRINKAGE

6.1. Characterization of Shrinkage in Native (Unembedded) Crystals at Low Temperature

It is now well established that radiation damage effects are greatly reduced by electron microscopy at low temperatures (Glaeser and Taylor, 1978). For example, studies on purple membrane have shown that at liquid nitrogen temperatures, there is a reduction in radiation sensitivity by a factor of at least four, as judged by the fading of the diffraction spots (Hayward and Glaeser, 1979).

As mentioned in Section 5.4, a systematic study of catalase thickness under various conditions in the electron microscope was carried out by Berriman and Leonard (1986). They monitored the effects of radiation dose by measuring the separation of the Laue zones in the electron diffraction

pattern. They also measured the effect of the preparation on the starting thickness prior to any electron microscopic irradiation (Fig. 10). From X-ray diffraction of crystals in the hydrated state, the cell $\underline{c}$ dimension was measured to be 20.6 nm (asterisk in Fig. 10). With negatively stained crystals, they noted that this dimension was reduced to 70% of the hydrated value at the start of the electron microscope investigation. As mentioned earlier, for the sample kept at room temperature, trace 5 shows a logarithmic shrinkage of a further 50% following a dose of 10^6 e/nm^2. Maintaining the negatively stained crystals at 120 K, they found no change in $\underline{c}$ for high levels of a dose of ~2×10^5/nm^2 (trace 3). Negatively stained crystals with a contaminating layer of ice or a layer of carbon deposited, kept at 120 K, follow the same course (trace 4) as for crystals without contamination (trace 3), but are more radiation sensitive. They also examined crystals 'embedded' in glucose and viewed them at room temperature with low dose. Trace 2 shows that the starting thickness was 80% of the native value. Although the sample did not shrink, it was highly radiation sensitive, and the pattern disappeared after a dose of only 300 e/nm^2. Viewing frozen-hydrated crystals gave the value of $\underline{c}$ closest to the native value (90%, trace 1). There was no shrinkage and the sample, although radiation sensitive, is more resistant than the glucose-embedded sample at room temperature; here the pattern is destroyed after a dose of 800 e/nm^2.

6.2. Characterization of Shrinkage of Plastic Sections at Low Temperature

Characterization of shrinkage of plastic sections cooled to cryotemperatures is a subject of paramount importance for tomography studies. Kremer *et al*. (1990) examined 250 nm thick sections of Epon 812/Araldite coated with TiSi on a high voltage electron microscope operated at 1000 kV with the section cooled to –160°C. Metallic coatings are used at low temperature as carbon becomes an insulator (Rader and Lamvik, 1992). The metallic coatings eliminate charging effects at these temperatures (Braunfeld *et al*., 1994). The shrinkage was monitored by the gold particles method. Using a continuous irradiation of 100 e/nm^2/s, they observed the usual rapid phase of shrinkage, followed by a very slow phase. They observed a final z-shrinkage of 20%, with little lateral shrinkage.

Braunfeld *et al.* (1994) examined sections of Epon resin coated with 3 nm of titanium at –165°C. As shown in Fig. 12, they found that 50% of the total z-shrinkage occurs after 1.4×10^5 e/nm^2. However, their results showed that in the dose range <5000 e/nm^2, there was virtually no shrinkage. Hence they proposed that tilt series could be recorded at low temperatures within the limiting dose of 5000 e/nm^2 with negligible shrinkage. This proposal does not appear to have been tested in practice.

In another study, Trachtenberg *et al*. (2000) examined the surface layer of *Halobacterium salinarum* embedded in Spurr's resin. The sample was fast

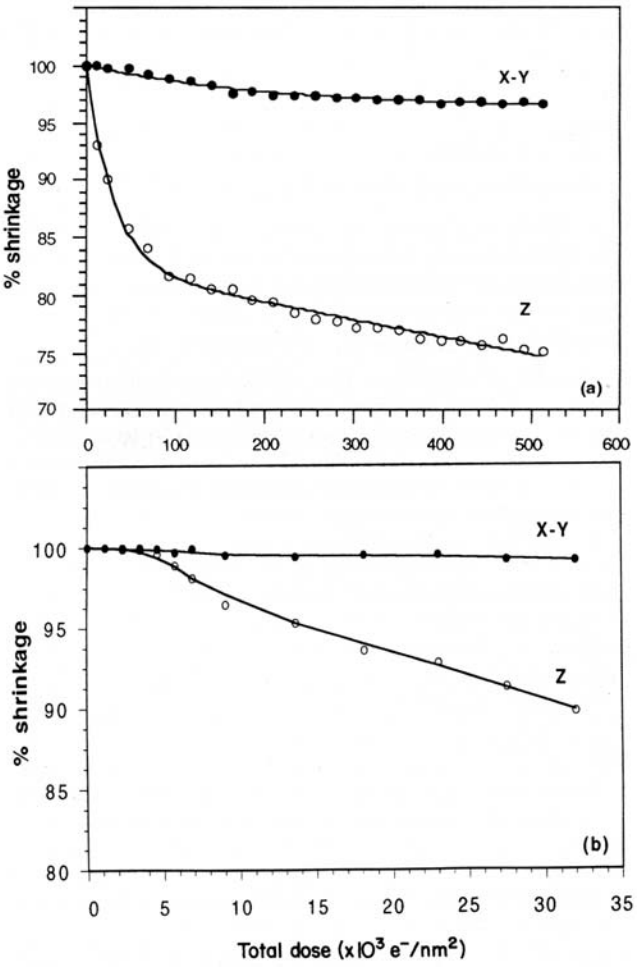


FIGURE 12. Characterization of shrinkage in Epon resin cooled to –165°C. The low-dose data are shown on an expanded scale in (b). In these titanium-coated sections, 50% of the total *z*-shrinkage occurs after 1.4×105 e/nm^2. As shown in (b), in the dose range <5000 e/nm^2, there is virtually no shrinkage. (From Braunfeld *et al.*, 1994; reproduced with permission by Blackwell Publishing).

frozen and freeze substituted and the sections observed at –170°C. They reported that the section was stable after a dose of 10,000 e/nm^2. Unfortunately, they noted a high *z*-shrinkage of ~45% despite their use of low temperatures.

In general, in laboratories where tomography is carried out, cryoholders are routinely available. With current automated methods and availability of cryoholders, it is surprising that tomographic tilt series for plastic sections are rarely recorded at low temperatures, although there is much

evidence that cryoimaging of plastic sections would reduce the shrinkage. Notable tomography studies at low temperature include those of Harlow *et al*. (1998, 2001), but the benefit of cryotemperature was not discussed. The reason for the reluctance to use low temperatures may be the inconsistency in the shrinkage values, as in the study of Trachtenberg *et al*. (2000) above. More tomographic studies need to be done to show the benefits of cryogenic imaging for tomography.

7. REVIEW OF SAMPLE PREPARATION AND IMAGING PROTOCOLS IN PAST TOMOGRAPHY STUDIES

7.1. Comments on Sample Preparation and Tomography

Compared with the 1st Edition of this chapter, there are now several published tomography studies with some spectacular success stories. Excellent reviews have appeared recently highlighting the techniques and achievements (Lucic *et al*., 2005; Marsh, 2005; McEwen and Frank, 2001; McIntosh *et al*., 2005).

It is appropriate to review the methods of sample preparation that have been used in these studies and the protocols used for collecting the tomography data set. A selection of these studies is listed in Table 2. The early studies used conventional chemical fixation for preparing their samples, but recent studies use rapid freezing/freeze substitution methods. For embedding the samples, epoxy resins of the Epon/Araldite type are the most common. A variety of section thicknesses are used, ranging from ultrathin ~30 nm sections of Liu *et al*. (2004b) and He *et al*. (2003) that aim to probe molecular details, to medium thick 500 nm studies that encompass complete organelles and parts of complete cells. The most popular electron microscopes for tomography are operated at 300 kV, fitted with field emission source. HVEM electron microscopes (1 MV) that had been falling out of favor are now back in demand for tomography of thick sections. Bouwer *et al*. (2004) have demonstrated that by using most-probable loss (MPL) energy signal with an intermediate voltage electron microscope (IVEM), it is feasible to record tilt series for tomography of a section up to 5 μm thick.

Two studies are described separately below, one which unravels molecular detail (He *et al*., 2003) and the other cellular detail (Ladinsky *et al*., 1999).

7.2. Tomography of the Desmosome

In this study by He *et al*. (2003), neonatal mouse skin was processed by high-pressure freezing/freeze substitution, embedded in epoxy resin, thin 30–70 nm sections cut, carbon-coated, and viewed in a 300 KV FEG electron microscope. Of the different tomograms calculated, the best detail was obtained using data collected over two orthogonal axes, over the tilt range ±75° in 1.5° steps. Electron tomograms normally have non-isotropic

TABLE 2. Sample preparation and tomography protocols used in a few selected studies

References	Sample	Fixation protocol	Resin	Section thickness (nm)	Acc voltage (kV)	Pre-irradiation	Dose/image (e/nm^2)	Total dose (e/nm^2)	Tilt-series: dual axis, step, range (°)
Horowitz *et al.* (1994)	Chromatin	FF/FS	Lowicryl K11M	90–120	150	5 min	210		1.5, ±70
Landis *et al.* (1996)	Tendon	Chemical	Epon	500	1000				2, ±60
Ladinsky *et al.* (1999)	Golgi	FF/FS	Epon/Araldite	250	1000				DA, 1.5, ±60
Perkins *et al.* (1997)	Mitochondria	Chemical	Durcupan	500	400				1, ±60
Rath *et al.* 1997)	Triad junction in muscle	FF/FS	Polybed 812	125	400		1000		DA, 2, ±60
Harlow *et al.* (1998)	NM junction	Chemical	Epon	70	200			3.5×10^5	2, ±66
Deng *et al.* (1999)	Mitochondria	Chemical	Epon/Araldite	250	400		500		DA, 2, ±60
Lenzi *et al.* (1999)	Synapse	Chemical	Epon or Araldite	250, 500, 1000	400	15 s/image			1 or 2, ±60
Muller *et al.* (2000)	Septal pole cap	HPF/FS	Spurr	250	200	Done			1, –73 to 69
Trachtenberg *et al.* (2000)	Bacterial wall	HPF/FS	Spurr	100	200				1 or 2, ±80
Perkins *et al.* (2001)	Mitochondria	Chemical	Epon 812	500	400	Done			2, ±60
Lenzi *et al.* (2002)	Synaptic body	Chemical	Epon or Araldite	500	400				1 or 2, ±60
He *et al.* (2003)	Desmosome	HPF/FS	LX112–Epoxy	30–70	200	10^5 e/nm^2		10^5	DA, 1.5, ± 5
Marsh *et al.* (2004)	Golgi	HPF/FS	Embed 812/Araldite	300–400	750				DA, 1.5 ±60
Lefman *et al.* (2004)	*E. coli*	Chemical	Cryosection (no resin)	~100	300				0.5, ±70
Liu *et al.* (2004a)	Striated muscle	FF/FS	Araldite	30	300				DA, ±75, cosine rule, 80 images
Uzawa *et al.* (2004)	Spindle pole body	HPF/FS	Epon 812	400	300	15–20 min			1.25, ±70

Where more than one procedure is used, e.g. in the tilt steps and range, only the finest is included here. The dose parameters are included where they are given in the relevant papers. Ladinsky *et al.* (1999) and Marsh *et al.* (2004) carried out double-axis tomography on serial sections and combined them to produce one large volume. Harlow *et al.* (1998) performed the image acquisition at low temperature (–160°C). The Lefman *et al.* (2004) study was done on cryosections prepared by the Tokuyasu method, infiltrated with methyl cellulose/uranyl acetate and dried (no plastic resin was used).
FF/FS = fast freezing/freeze substitution; HPF/FS = rapid freezing by high pressure/freeze substitution.

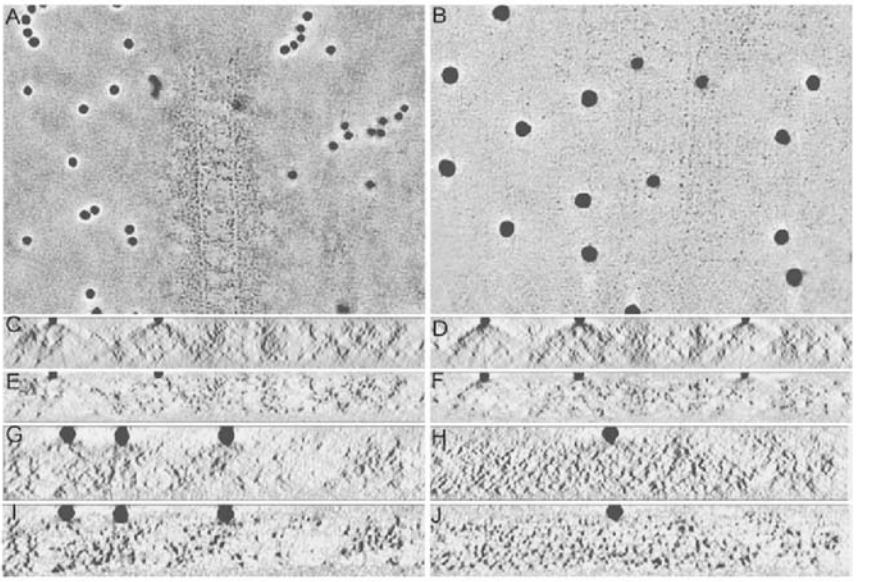


FIGURE 13. Factors affecting isotropic resolution in tomograms. (a and b) *x–y* slices from tomograms with tilt range ±60° (a) and ±75° (b); white rings apparent in (a) are much reduced with higher tilt range. (c and d) *y–z* sections from a ±60° single-tilt axis tomogram show conical streak artifacts. (e and f) *y–z* sections from a ±60° dual-tilt axis tomogram show reduced artifact images. (g and h) *y–z* sections from a ±75°single-tilt axis tomogram show reduced conical streaks. (i and j) *y–z* sections from a ±75°dual-tilt axis tomogram show round gold particles with minimal artifacts. The authors claim that desmosomes are tightly packed and have high radiation resistance, hence shrink little. Excessive *z*-shrinkage in a sample effectively reduces the tilt range, hence the quality of the 3D image. (From He et al., 2003, courtesy of David Stokes).

resolution, since the limited tilt range (e.g. ±60º) results in a missing wedge in Fourier space. He *et al.* claim that their tomogram has an isotropic resolution of 2.3 nm, because of their double-axis high tilt range protocol and because the cadherin molecules are tightly packed within the desmosome and hence may be resistant to shrinkage by radiation. One should therefore try to assess the shrinkage behavior of a sample prior to the tomography. This feature is illustrated in Fig. 13 taken from their Supplement Fig. S2. After segmenting their tomogram with Amira (TGS Inc, San Diego, Californaia) in an iterative process, He *et al.* were able to fit the crystal structure of cadherin to their electron density maps.

7.3. Studies on the Golgi Apparatus

Unlike the ultrathin sections used to elucidate the molecular structure of the desmosome in the previous example, intermediate thickness sections,

250–500 nm, have been used to understand the 3D structure of the Golgi apparatus. In a landmark study, Ladinsky *et al*. (1999) used four 250 nm thick serial sections, collected tilt series over two orthogonal axes over the range ±60 in steps of 2°, calculated each tomogram, and combined them to give a single tomogram of volume $1 \times 1 \times 4\,\mu m^3$. The samples were prepared by rapid freezing/freeze substitution techniques, followed by embedding in Epon/Araldite resin. The 7 nm resolution 3D image allowed them to dissect the Golgi apparatus in different planes, giving new and insightful views of tubules and vesicles associated with the different cisternae. The development of the technique of tomography applied to Golgi research has been reviewed by Marsh (2004), and this is also an excellent general review on tomography.

8. CONCLUSIONS AND RECOMMENDATIONS

For sample preparation, it is clear that rapid freezing (e.g. by high-pressure freezing) followed by freeze-substitution gives the best sample preservation. For embedding the sample, the most common resins used are of the epoxy type, such as Epon or Araldite. The section thickness depends on the size of the object of study, and can range from ~30 nm for study of molecular organization to ~500 nm to study the 3D structure of organelles and cellular entities. The sections may be coated with 0.02% polylysine and then with gold particles ranging from 10 to 30 nm to allow accurate alignment of the tilt views. Some methods of alignment use correlation matching of the different views and do not require gold particles (Winkler and Taylor, 2006). It is essential to carbon-coat the sections to reduce charging effects and improve stability in the microscope. For thin sections, a 100 kV electron microscope is sufficient. However, electron microscopes with 200 and 300 kV fitted with field emission sources are regularly used for thin sections. For the thicker, 250–500 nm, sections, intermediate-voltage, 300 and 400 kV or high-voltage 1 MV electron microscopes are required. The use of automated methods is highly desirable to minimize the radiation dose on the sample, since low-dose methods are used to track, focus and record the image. The saving in labor and electron dose for ~100–200 images required for a dual-tilt axes series is enormous. Recording images directly onto a CCD camera with at least 1000 × 1000 pixels or bigger is another great saving on labor. Pre-irradiation of the section is necessary in order to induce the rapid phase of the shrinkage, to allow collection of data over the slow phase. In practice, with the beam intensity set to the value to be used for recording the images, a few (~3–4) minutes pre-irradiation is sufficient. One can easily ascertain that the rapid phase shrinkage is complete by examining a few successive images of a field of fiducial markers on a grid tilted to a high fixed angle (e.g. 45°). For the data collection, the use of low-dose methods ensures that there is negligible further shrinkage. In the study by He *et al.* (2003), the total pre-irradiation dose was $10^5\,e/nm^2$ and the same amount again for the complete data collection. The beam intensity value

used for the pre-irradiation should not be exceeded for the data collection as it could induce further shrinkage. To obtain the most isotropic 3D images, the tomography data set should include double axes tilt series collected over a large range, about ±75° (He *et al*., 2003). For the sampling steps over the tilt range, 1–2° steps are routinely used with the computer-controlled specimen holders. The alternative is to use the Saxton scheme in which finer steps are used at higher tilt angles to compensate for the increased relative thickness of the sample (Saxton *et al*., 1984).

In the first edition of this book, I wrote 'The overwhelming conclusion . . . to limit the effects of electron irradiation, a sample must be cooled to low temperatures (about –150°C or less)'. Although cryoholders and automated methods for data collection allow relatively straightforward low-temperature work, the take-up has been very poor. It appears that researchers, in a bid to ensure that the tilt series is recorded over the stable part of the shrinkage curve, pre-irradiate the sample with a high dose at room temperature. Considering the success of electron tomography of plastic-embedded sections, it appears that the shrinkage that occurs in room temperature studies is tolerable. Nevertheless, further studies are required to establish the benefit of cryo-cooling of plastic sections for tomography and to recommend suitable protocols. The ultimate goal is to eliminate plastic embedding altogether and cut cryo-sections of vitrified samples and view them by cryo-electron microscopy (see Marko et al., Chapter 2 of this volume).

ACKNOWLEDGEMENTS

I am indebted to J. M. Squire, E. P. Morris, P. Bennett, R. A. Crowther and J. Berriman (1st Edition) and Joachim Frank and Bruce McEwen (2nd Edition) for many helpful discussions and for critically reading the chapter. This work was supported by grants to J. M. Squire from the Leverhulume Trust and the British Medical Research Council (1st Edition) and by grants to P.K.L. from the British Heart Foundation (2nd Edition).

REFERENCES

Amos, L. A., Henderson, R. and Unwin, P. N. (1982). Three-dimensional structure determination by electron microscopy of two-dimensional crystals. *Prog. Biophys. Mol. Biol.* **39**: 183–231.

Bedi, K. S. (1987). A simple method of measuring the thickness of semi-thin and ultra-thin sections. *J. Microsc*. **148**:107–111.

Bennett, P. M. (1974). Decrease in section thickness on exposure to the electron beam; the use of tilted sections in estimating the amount of shrinkage. *J. Cell Sci*. **15**:693–701.

Berriman, J., Bryan, R. K., Freeman, R. and Leonard, K. R. (1984). Methods for specimen thickness determination in electron microscopy. *Ultramicroscopy* **13**:351–364.

Berriman, J. and Leonard, K. R. (1986). Methods for specimen thickness determination in electron microscopy. II. Changes in thickness with dose. *Ultramicroscopy* **19**:349–366.

Bouwer, J. C., Mackey, M. R., Lawrence, A., Deerinck, T. J., Jones, Y. Z., Terada, M., Martone, M. E., Peltier, S. and Ellisman, M. H. (2004). Automated most-probable loss tomography

of thick selectively stained biological specimens with quantitative measurement of resolution improvement. *J. Struct. Biol*. **148**:297–306.

Braunfeld, M. B., Koster, A. J., Sedat, J. W. and Agard, D. A. (1994). Cryo automated electron tomography: towards high-resolution reconstructions of plastic-embedded structures. *J. Microsc*. **174**:75–84.

Cosslett, A. (1960). The effect of the electron beam on thin sections. In *Proceedings of the 1st European Conference on Electron Microscopy*. Delft, Vol. 2, pp. 678–681.

Craig, R., Alamo, L. and Padron, R. (1992). Structure of the myosin filaments of relaxed and rigor vertebrate striated muscle studied by rapid freezing electron microscopy. *J. Mol. Biol*. **228**:474–487.

Deng, Y., Marko, M., Buttle, K. F., Leith, A., Mieczkowski, M. and Mannella, C. A. (1999). Cubic membrane structure in amoeba (*Chaos carolinensis*) mitochondria determined by electron microscopic tomography. *J. Struct. Biol*. **127**:231–239.

Dorset, D. L. and Parsons, D. F. (1975). The thickness determination of wet protein microcrystals: use of Laue zones in cross-grating diffraction patterns. *J. Appl. Phys.* **46**:938–940.

Egerton, R. F., Li, P. and Malac, M. (2004). Radiation damage in the TEM and SEM. *Micron* **35**:399–409.

Giddings, T. H., Jr, O'Toole, E. T., Morphew, M., Mastronarde, D. N., McIntosh, J. R. and Winey, M. (2001). Using rapid freeze and freeze-substitution for the preparation of yeast cells for electron microscopy and three-dimensional analysis. *Methods Cell Biol*. **67**:27–42.

Gillis, J. M. and Wibo, M. (1971). Accurate measurement of the thickness of ultrathin sections by interference microscopy. *J. Cell Biol*. **49**:947–949.

Glaeser, R. M. and Taylor, K. A. (1978). Radiation damage relative to transmission electron microscopy of biological specimens at low temperature: a review. *J. Microsc*. **112**:127–138.

Glauert, A. M. (1998). *Biological Specimen Preparation for Transmission Electron Microscopy*. Portland Press.

Grubb, D. T. (1974). Radiation damage and electron microscopy of organic polymers. *J. Mater. Sci*. **9**:1715–1736.

Gunning, B. E. S. and Hardham, A. R. (1977). Estimation of the average section thickness in ribbons of ultra-thin sections. *J. Microsc*. **109**:337–340.

Harlow, M., Ress, D., Koster, A., Marshall, R. M., Schwarz, M. and McMahan, U. J. (1998). Dissection of active zones at the neuromuscular junction by EM tomography. *J. Physiol. Paris* **92**:75–78.

Harlow, M. L., Ress, D., Stoschek, A., Marshall, R. M. and McMahan, U. J. (2001). The architecture of active zone material at the frog's neuromuscular junction. *Nature* **409**:479–484.

Hayat, M. A. (2000). *Principles and Techniques of Electron Microscopy*. Cambridge University Press.

Hayward, S. B. and Glaeser, R. M. (1979). Radiation damage of purple membrane at low temperature. *Ultramicroscopy* **4**:201–210.

He, W., Cowin, P. and Stokes, D. L. (2003). Untangling desmosomal knots with electron tomography. *Science* **302**:109–113.

Heuser, J. E. (1989). Development of the quick-freeze, deep-etch, rotary-replication technique of sample preparation for 3-D electron microscopy. *Prog. Clin. Biol. Res*. **295**:71–83.

Heuser, J. E., Keen, J. H., Amende, L. M., Lippoldt, R. E. and Prasad, K. (1987). Deep-etch visualization of 27S clathrin: a tetrahedral tetramer. *J. Cell Biol*. **105**:1999–2009.

Hirose, K., Lenart, T. D., Murray, J. M., Franzini-Armstrong, C. and Goldman, Y. E. (1993). Flash and smash: rapid freezing of muscle fibers activated by photolysis of caged ATP. *Biophys. J*. **65**:397–408.

Horowitz, R. A., Agard, D. A., Sedat, J. W. and Woodcock, C. L. (1994). The three-dimensional architecture of chromatin *in situ*: electron tomography reveals fibers composed of a continuously variable zig–zag nucleosomal ribbon. *J. Cell Biol*. **125**:1–10.

Jesior, J. C. (1982). A new approach for the visualization of molecular arrangement in biological micro-crystals. *Ultramicroscopy* **8**:379–384.

Jesior, J. C. and Wade, R. H. (1987). Electron-irradiation-induced flattening of negatively stained 2D protein crystals. *Ultramicroscopy* **21**:313–319.

Koster, A. J., Grimm, R., Typke, D., Hegerl, R., Stoschek, A., Walz, J. and Baumeister, W. (1997). Perspectives of molecular and cellular electron tomography. *J. Struct. Biol*. **120**:276–308.

Kremer, J. R., O'Toole, E. T., Wray, G. P., Mastronarde, D. M., Mitchell, S. J. and McIntosh, J. R. (1990). Characterization of beam-induced thinning and shrinkage of semi-thick section in H.V.E.M. In *Proceedings of the XIIth International Congress for Electron Microscopy* (Peachey, L. D. and Williams, D. B., eds). San Francisco Press Inc., San Francisco, pp. 752–753.

Ladinsky, M. S., Mastronarde, D. N., McIntosh, J. R., Howell, K. E. and Staehelin, L. A. (1999). Golgi structure in three dimensions: functional insights from the normal rat kidney cell. *J. Cell Biol*. **144**:1135–1149.

Lamvik, M. K. (1991). Radiation damage in dry and frozen hydrated organic material. *J. Microsc*. **161**:171–181.

Landis, W. J., Hodgens, K. J., Song, M. J., Arena, J., Kiyonaga, S., Marko, M., Owen, C. and McEwen, B. F. (1996). Mineralization of collagen may occur on fibril surfaces: evidence from conventional and high-voltage electron microscopy and three-dimensional imaging. *J. Struct. Biol*. **117**:24–35.

Lefman, J., Zhang, P., Hirai, T., Weis, R. M., Juliani, J., Bliss, D., Kessel, M., Bos, E., Peters, P. J. and Subramaniam, S. (2004). Three-dimensional electron microscopic imaging of membrane invaginations in *Escherichia coli* overproducing the chemotaxis receptor Tsr. *J. Bacteriol*. **186**:5052–5061.

Lenzi, D., Crum, J., Ellisman, M. H. and Roberts, W. M. (2002). Depolarization redistributes synaptic membrane and creates a gradient of vesicles on the synaptic body at a ribbon synapse. *Neuron* **36**:649–659.

Lenzi, D., Runyeon, J. W., Crum, J., Ellisman, M. H. and Roberts, W. M. (1999). Synaptic vesicle populations in saccular hair cells reconstructed by electron tomography. *J. Neurosci*. **19**:119–132.

Liu, J., Reedy, M. C., Goldman, Y. E., Franzini-Armstrong, C., Sasaki, H., Tregear, R. T., Lucaveche, C., Winkler, H., Baumann, B. A., Squire, J. M., Irving, T. C., Reedy, M. K. and Taylor, K. A. (2004a). Electron tomography of fast frozen, stretched rigor fibers reveals elastic distortions in the myosin crossbridges. *J. Struct. Biol*. **147**:268–282.

Liu, J., Taylor, D. W. and Taylor, K. A. (2004b). A 3-D reconstruction of smooth muscle alpha-actinin by CryoEm reveals two different conformations at the actin-binding region. *J. Mol. Biol*. **338**:115–125.

Lucic, V., Forster, F. and Baumeister, W. (2005). Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem*. **74**:833–865.

Luther, P. K., Lawrence, M. C. and Crowther, R. A. (1988). A method for monitoring the collapse of plastic sections as a function of electron dose. *Ultramicroscopy* **24**:7–18.

Marsh, B. J. (2005). Lessons from tomographic studies of the mammalian Golgi. *Biochim. Biophys. Acta* **1744**:273–292.

Marsh, B. J., Volkmann, N., McIntosh, J. R. and Howell, K. E. (2004). Direct continuities between cisternae at different levels of the Golgi complex in glucose-stimulated mouse islet beta cells. *Proc. Natl Acad. Sci. USA* **101**:5565–5570.

McEwen, B. F. and Frank, J. (2001). Electron tomographic and other approaches for imaging molecular machines. *Curr. Opin. Neurobiol*. **11**:594–600.

McIntosh, R., Nicastro, D. and Mastronarde, D. (2005). New views of cells in 3D: an introduction to electron tomography. *Trends Cell Biol*. **15**:43–51.

Muller, W. H., Koster, A. J., Humbel, B. M., Ziese, U., Verkleij, A. J., van Aelst, A. C., van der Krift, T. P., Montijn, R. C. and Boekhout, T. (2000). Automated electron tomography of the septal pore cap in *Rhizoctonia solani*. *J. Struct. Biol*. **131**:10–18.

Ohno, S. (1980). Morphometry for determination of size distribution of peroxisomes in thick sections by high-voltage electron microscopy: I. Studies on section thickness. *J. Electron Microsc. (Tokyo)* **29**:230–235.

Padron, R., Alamo, L., Craig, R. and Caputo, C. (1988). A method for quick-freezing live muscles at known instants during contraction with simultaneous recording of mechanical tension. *J. Microsc*. **151**:81–102.

Peachey, L. D. (1958). Thin sections. I. A study of section thickness and physical distortion produced during microtomy. *J. Biophys. Biochem. Cytol*. **4**:233–242.

Perkins, G. A., Renken, C. W., Song, J. Y., Frey, T. G., Young, S. J., Lamont, Martone, S. M., Lindsey, E. S. and Ellisman, M. H. (1997). Electron tomography of large, multicomponent biological structures. *J. Struct. Biol*. **120**:219–227.

Perkins, G. A., Renken, C. W., van der Klei, I. J., Ellisman, M. H., Neupert, W. and Frey,T. G. (2001). Electron tomography of mitochondria after the arrest of protein import associated with Tom19 depletion. *Eur. J. Cell Biol*. **80**:139–150.

Rader, R. S. and Lamvik, M. L. (1992). High conductivity amorphous Ti88Si22 substrates for low temperature electron microscopy. *J. Microsc.* **168**:71–77.

Rath, B. K., Marko, M., Radermacher, M. and Frank, J. (1997). Low-dose automated electron tomography: a recent implementation. *J. Struct. Biol*. **120**:210–218.

Reimer, L. (1989). *Transmission Electron Microscopy*. Springer-Verlag, Berlin.

Saxton, W. O., Baumeister, W. and Hahn, M. (1984). Three-dimensional reconstruction of imperfect two-dimensional crystals. *Ultramicroscopy* **13**:57–70.

Shimoni, E. and Muller, M. (1998). On optimizing high-pressure freezing: from heat transfer theory to a new microbiopsy device. *J. Microsc.* **192**:236–247.

Sjostrom, M., Squire, J. M., Luther, P., Morris, E. and Edman, A. C. (1991). Cryoultramicrotomy of muscle: improved preservation and resolution of muscle ultrastructure using negatively stained ultrathin cryosections. *J. Microsc*. **163**:29–42.

Slot, J. W. and Geuze, H. J. (1985). A new method of preparing gold probes for multiple-labeling cytochemistry. *Eur. J. Cell Biol*. **38**:87–93.

Small, J.V. (1968). Measurements of section thickness. In *Proceedings of the 4th European Congress on Electron Microscopy* (S. Bocciareli, ed.). Vol. 1, pp. 609–610.

Sosa, H., Popp, D., Ouyang, G. and Huxley, H. E. (1994). Ultrastructure of skeletal muscle fibers studied by a plunge quick freezing method: myofilament lengths. *Biophys. J*. **67**:283–292.

Spencer, M. (1982). *Fundamentals of Light Microscopy*. IUPAB Biophysics Series.

Stenn, K. and Bahr, G. F. (1970). Specimen damage caused by the beam of the transmission electron microscope, a correlative reconsideration. *J. Ultrastruct. Res*. **31**:526–550.

Studer, D., Graber, W., Al-Amoudi, A. and Eggli, P. (2001). A new approach for cryofixation by high-pressure freezing. *J. Microsc*. **203**:285–294.

Trachtenberg, S., Pinnick, B. and Kessel, M. (2000). The cell surface glycoprotein layer of the extreme halophile *Halobacterium salinarum* and its relation to *Haloferax volcanii*: cryo-electron tomography of freeze-substituted cells and projection studies of negatively stained envelopes. *J. Struct. Biol*. **130**:10–26.

Unwin, P. N. (1974). Electron microscopy of the stacked disk aggregate of tobacco mosaic virus protein. II. The influence of electron irradiation of the stain distribution. *J. Mol. Biol*. **87**:657–670.

Uzawa, S., Li, F., Jin, Y., McDonald, K. L., Braunfeld, M. B., Agard, D. A. and Cande, W. Z. (2004). Spindle pole body duplication in fission yeast occurs at the G_1/S boundary but maturation is blocked until exit from S by an event downstream of cdc10+. *Mol. Biol. Cell*. **15**:5219–5230.

van Marle, J., Dietrich, A., Jonges, K., Jonges, R. de Moor, E., Vink, A., Boon, P. and van Veen, H. (1995). EM-tomography of section collapse, a non-linear phenomenon. *Microsc. Res. Tech*. **31**:311–316.

Williams, M. A. and Meek, G. A. (1966). Studies on thickness variation in ultrathin sections for electron microscopy. *J. R. Microsc. Soc.* **85**:337–352.

Winkler, H. and Taylor, K. A. (2006). Accurate marker-free alignment with simultaneous geometry determination and reconstruction of tilt series in electron tomography. *Ultramicroscopy* **106**:240–254.

Yang, G. C. H. and Shea, S. M. (1975). The precise measurement of the thickness of ultrathin sections by a 're-sectioned' section technique. *J. Microsc*. **103**:385–392.

2

Electron Tomography of Frozen-hydrated Sections of Cells and Tissues

Michael Marko, Chyong-Ere Hsieh and Carmen A. Mannella

Michael Marko, Chyong-Ere Hsieh and Carmen Mannella • Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, PO Box 509, Albany, NY 12201, USA

1. INTRODUCTION

1.1. Cryoelectron Tomography

The technique of cryoelectron tomography of frozen-hydrated biological specimens is opening a new window on cellular structure and organization. This imaging method provides full 3D structural information at much higher resolution (typically 5–10 nm) than is attainable by light microscopy, and can be applied to cells and organelles that are maintained in a state that is as close to native as can be achieved currently in electron microscopy. Not only can cryoelectron tomography be used to visualize directly extended cellular structures, such as membranes and cytoskeleton, but it can also provide 3D maps of the location, orientation and, perhaps, the conformation of large macromolecular complexes, the cell's 'molecular machinery'. This information complements that coming from single-particle cryoelectron microscopy (Frank *et al*., 1996, 2006) and X-ray crystallography, about the subnanometer structure of the same molecular assemblies after isolation. As with studies using single-particle cryoelectron microscopy, specimens smaller than 1 μm in size can be prepared for cryoelectron tomography by plunge-freezing (Dubochet *et al*., 1988). Cells or organelles can be rapidly frozen directly on an electron microscope grid in thin layers of glass-like, amorphous ice, without the formation of ice crystals that would otherwise disrupt fine structure (Kellenberger, 1987). Specimens are imaged directly, without chemical fixation, dehydration or staining with heavy metals. Cryoelectron tomography is made possible by electron microscope automation, which allows the recording of image series from sequentially tilted specimens with a sufficiently low cumulative electron dose, such that damage to high-resolution fine structure is avoided (Dierksen *et al*., 1993, 1995; Koster *et al*., 1997; Mastronarde, 2005; Rath *et al*., 1997; Chapter 4 of this volume).

Examples of specimens studied by cryotomography since the year 2000 include (i) isolated organelles such as mitochondria (Mannella, 2005; Mannella *et al*., 2001; Nicastro *et al*., 2000), axonemes (McEwen *et al*., 2002)

and triad junctions (Wagenknecht *et al*., 2002); (ii) small bacterial cells, such as *Pyrodictium* (Nickell *et al*., 2003) and *Spiroplasma* (Kürner *et al*., 2004, 2005); and (iii) thin parts of intact eukaryotic cells, such as the cytoskeleton in the leading edge of *Dictyostelium* (Bretschneider *et al*., 2002; Medalia *et al*., 2002) and sea urchin sperm flagella (Nicastro *et al*., 2005a). The reader is also referred to reviews by Koster *et al*. (1997), McIntosh (2001), McIntosh *et al*. (2005), Plitzko *et al*. (2002), Steven and Aebi (2003) and Leapman (2005).

Considerable progress has been made in the use of cryotomography to reconstruct and identify macromolecules and macromolecular assemblies, sometimes within larger scale reconstructions of cells or organelles. Examples are the nuclear pore complex (Beck *et al*., 2004; Stöffler *et al*., 2003), viruses (Cryklaff *et al*., 2005; Förster *et al*., 2005; Grünewald *et al*., 2003), ATPase (Rockel *et al*., 2002), ribosomal subunits (Zhao *et al*., 2004), immunoglobin (Bogini *et al*., 2004; Sandin *et al*., 2004) and synaptic macromolecules (Lucic *et al*., 2005a). When the 3D orientation of the macromolecule can be precisely determined, an averaged structure can be computed in order to achieve increased resolution of the *in situ* macromolecule (Böhm *et al*., 2000, 2001; Frangakis and Förster, 2004; Frangakis *et al*., 2002; Förster *et al*., 2005; Grünewald *et al*., 2002; Rath *et al*., 2003; Walz *et al*., 1997; Chapter 14 of this volume).

Reviews containing useful discussions of the role of cryoelectron tomography in structural biology and cellular proteomics include those of Koster *et al*. (1997), Baumeister and Steven (2000), McEwen and Frank (2001), McIntosh (2001, 2005), Frank *et al*. (2002), Baumeister (2002), Plitzko *et al*. (2002), Steven and Aebi (2003), Sali *et al*. (2003), Lucic *et al*. (2005b) and Leapman (2005).

Until recently, cryoelectron tomography could not be applied to excised bulk tissue or to large cells, except at their leading edges, in part because these specimens are too large to be vitreously frozen by plunging into a cryogen. In addition, a specimen thickness exceeding 500 nm poses serious obstacles for high-resolution tomography, even when an accelerating voltage in the range of 300–400 kV is used, with respect to both the requirement for low-dose imaging (Grimm *et al*., 1998) and the number of images needed for the reconstruction (Crowther *et al*., 1970).

1.2. Role of Frozen-hydrated Sections in Tomography

Cryoultramicrotomy of frozen-hydrated specimens currently offers the best possibility to extend electron tomography to large cells and bulk tissue in their native states (Dubochet and Sartori Blanc, 2001). Sections are usually cut from high-pressure-frozen material with diamond knives, and are mounted 'dry' on transmission electron microscopy (TEM) grids for cryoelectron microscopy. Like plunge-frozen specimens, frozen-hydrated sections must be kept below the devitrification temperature (–133 to –138°C; Dubochet *et al*., 1988) at all times, to prevent formation of damaging ice crystals.

Frozen-hydrated sections should be distinguished from 'cryosections'. Cryosections is the term that is commonly applied to sections that are prepared by the technique of Tokuyasu (1986) for immunological studies. In the Tokuyasu method, material is lightly pre-fixed, cryoprotected with sucrose, frozen, and sectioned at –90°C. Sections are then collected on a sucrose solution, and warmed to room temperature for immunolabeling, heavy metal staining and electron microscopy. In contrast, frozen-hydrated sections are not subjected to chemical fixation or staining, are prepared at temperatures below –135°C, and are maintained at those temperatures during electron microscopy.

Successful use of frozen-hydrated sections for electron tomography was first reported a few years ago, with specific application to mammalian tissue (Hsieh *et al*., 2002, 2003; Marko *et al*., 2002). The technique has recently been adopted by other laboratories, for studies of yeast (Schwartz *et al*. 2003), isolated chloroplasts (Nicastro *et al*., 2005b) and algae (Leis *et al*., 2005). The methodology described in this chapter reflects, and expands on, refinements in the technique, as recently reported in Hsieh *et al*. (2006).

The cryoultramicrotome can be used to cut sections with actual thicknesses in the range of ~70–350 nm, as measured by subsequent tomographic reconstruction or by electron energy loss. This means that the thickness of the ice layer is not dictated by the size of the organelle or cell. As pointed out elsewhere in this book, higher tomographic resolution can be obtained for a given number of tilt images (and the same electron dose), if the specimen is thinner. Thus, for frozen-hydrated sections, the investigator can negotiate between resolution and 3D depth. In some cases, it is desirable to make reconstructions at two different section thicknesses.

In previous reports, we compared tomograms from frozen-hydrated tissue sections with tomograms from plastic sections of the same tissues, prepared using high-pressure freezing and freeze substitution. In tomograms of liver (Hsieh *et al*., 2006) and skin (Hsieh *et al*., 2004), the cellular ultrastructure was similar, and the principal differences seen were due to the selective staining and obscuration of features by the heavy metals used for the plastic sections. Such plastic sections may be preferred over frozen-hydrated sections in some situations, for example when the experiment requires immunolabeling or collection of a large number of tomograms. However, in such applications, it is advisable to compare plastic sections with frozen-hydrated sections of the same specimen, to confirm that the structures of interest have not been significantly altered by freeze substitution or staining.

2. FREEZING METHODS

At ambient pressure, formation of vitreous (non-crystalline) ice requires cooling at a rate of ~10^5 °C/s to below the vitrification temperature, about –135°C. During plunge- or slam-freezing, this cooling rate occurs in

specimens only to a depth of one to several micrometers from the cryogen or metal surface, respectively. The rate of cooling deeper into the specimen is slower, resulting in formation of ice crystals. High pressure suppresses ice crystal formation; thus application of a cryogen at high pressure (≥2000 bar) allows biological specimens to be frozen without damage from large ice crystals, at depths where the cooling rate is as much as 10 times slower than at the surface. For specimens that have a naturally very low water content, or those that are infiltrated with sugars or other cryoprotectants, low freezing rates may be adequate for good vitrification of thick specimens, and high-pressure freezing may not be required. For a general reference on freezing techniques for electron microscopy, see Steinbrecht and Zierold (1987).

2.1. Ambient-pressure Freezing

The use of freezing as an alternative to chemical fixation is not a new idea. From the start, it was recognized as potentially offering a means to improve the fidelity of preservation in electron microscopy. In the 1950s, Fernàndez-Moràn experimented with freezing using liquid helium (Fernàndez-Moràn, 1960). More than a decade later, Taylor and Glaeser (1974, 1976) initiated cryoelectron microscopy of macromolecules, using liquid nitrogen as the cryogen.

Sitte (1996) described a technique in which a small droplet of cell suspension is placed on the tip of a pin that fits into the microtome chuck. The pin is plunged in liquid ethane, and then mounted in the microtome, after which the specimen can immediately be sectioned. Only a small number of good sections can be obtained, however, because biological material is vitrified only to a depth of ~1 μm in the absence of cryoprotectant. As noted by Dubochet *et al*. (1988), plunging is best done in liquid propane or liquid ethane, and not in liquid nitrogen. In the latter case, nitrogen gas produced at the specimen surface inhibits heat exchange (the so-called Leidenfrost effect), and reduces the cooling rate, resulting in ice crystal formation.

Slam-freezing achieves the fastest available cooling rate (>10^6 °C/s), and allows vitrification to a depth of ~10 μm. The specimen is rapidly pressed against a polished metal surface (cooled to the temperature of liquid nitrogen or helium) by means of a device that prevents bouncing of the specimen after the initial impact. Unfortunately, the layer that is optimally frozen is close to the impact surface, and cell structures, such as larger vesicles that are not well supported by the cytoskeleton, become distorted. Propane-jet freezing yields good freezing for specimens up to 40 μm in thickness, and is often used for freeze-fracture work, but it has rarely been used for frozen-hydrated sections. For reviews of these techniques, see Gilkey and Staehelin (1986) and Sitte (1996).

2.2. High-pressure Freezing

Currently, the most popular method of freezing tissue, as well as pelleted or flat-cultured cells, is high-pressure freezing (Dahl and Staehelin,

1989; Moor, 1987; Studer *et al*., 1989, 2001). Liquid nitrogen is applied at high pressure (≥2000 bar), which suppresses significant ice crystal formation to sample depths of 100–300 μm, despite the relatively slow cooling rate of 10^4 °C/s. There are numerous reports that high-pressure freezing, unlike plunge-freezing of thin specimens, does not produce purely amorphous ice, but rather a mixture of microcrystalline ice forms. The microcrystals range in size from 2 to 30 nm (Echlin, 1992; Michel *et al*., 1991; Moor, 1987). They are composed of high-pressure forms of ice (ice II, III and IX), as well as cubic and hexagonal ice (Erk *et al*., 1998; Sartori *et al*., 1996; Studer *et al*. 1995). Microcrystals have a greater tendency to form in regions of high water content (e.g. Lepault *et al*., 1997; Richter, 1996). With good technique, ice crystals that might be present are too small to cause noticeable structural damage (Gilkey and Staehelin, 1986; Moor, 1987; Vanhecke *et al*., 2003).

In plastic sections after freeze substitution, the presence of crystalline ice can only be assessed by its effects on specimen ultrastructure, e.g. segregation of cellular components and distortion of membranes (McDonald and Morphew, 1993). In frozen-hydrated sections, the state of the ice can be directly determined either by electron energy-loss spectroscopy (Sun *et al*., 1995) or, more conveniently, by electron diffraction. Vitreous ice is characterized by an electron diffraction pattern having two diffuse rings corresponding to spacings of 0.370 and 0.214 nm when freezing is performed at atmospheric pressure, or 0.357 and 0.214 nm when freezing is performed at high pressure (Sartori *et al*., 1996). The presence of cubic ice is indicated by sharp rings corresponding to spacings of 0.366 and 0.224 nm, while hexagonal ice gives rise to discrete spots, with the most intense occurring at these same spacings (Dubochet *et al*., 1988). The high-pressure crystalline forms, ice III and IX, are characterized by spots or arcs corresponding to spacings of 0.303, 0.279 and 0.179 nm, which are not shared by other forms of ice; a spacing at 0.225 nm, which is in common with both cubic and hexagonal ice; a spacing at 0.215 nm, which coincides with vitreous ice; and spacings at 0.195 and 0.169 nm, which are shared with weak reflections from hexagonal ice (Sartori *et al*., 1996)

We find that electron diffraction from frozen-hydrated sections cut from optimally high-pressure-frozen material is characterized by the pattern of two diffuse rings, and discrete spots are not apparent (Hsieh *et al*., 2006). This indicates either that microcrystals are absent, or else that they are present at concentrations too low to be detected by electron diffraction (discussed in more detail in Section 7.4).

3. OPTIMIZATION OF HIGH-PRESSURE FREEZING

3.1. Types of Specimen Carrier

Generally, for specimens prepared by high-pressure freezing, the depth of optimal freezing does not exceed 100 μm from the cooled surface of the

specimen carrier, unless the cell or tissue has very low water content (Shimoni and Müller, 1998). Several types of specimen carrier have been found to be appropriate for frozen-hydrated section work. The carrier used most often with the HPM 010 high-pressure freezer (Bal-Tec, Balzers, Liechtenstein) comprises pairs of 3 mm diameter aluminum disks that fit together to form an internal cavity 2 mm in diameter. Disks of different types may be assembled to provide cavity depths of 100, 200 or 300 μm. Bal-Tec also supplies specimen carriers that, when opened after freezing, leave a 'dome' of specimen exposed and ready for sectioning (Craig and Staehlin, 1988). The Bal-Tec disk carriers are used for either tissue or cell suspensions. For the EMPACT high-pressure freezer (Leica, Vienna, Austria), a single 3 mm diameter disk, available with a cavity in a variety of sizes and shapes, is used for tissue samples. Recently, a metal tube with an inside diameter of 200 or 300 μm has been used for cell suspensions, with both the Bal-Tec and the Leica high-pressure freezers (Studer *et al*., 2001).

3.2. Specimen Loading

A suspension of cultured cells, bacteria or isolated organelles is concentrated into a paste in the appropriate filler material (see below) before freezing in the disk-type specimen carrier or the metal tube. Alternatively, the cells can be grown as a layer directly in the disk-type carrier (Sawaguci *et al*., 2003), or they can be grown on filters (Morphew and McIntosh, 2003; Walther and Müller, 1997) or sapphire disks (Hess *et al*., 2000) that fit in the disk-type carrier.

Pieces of excised tissue must be small enough to fit into the disk-type carrier. The time between cessation of blood flow and freezing should be minimized, and probably should not exceed 1 min. An ideal loading scheme, proposed by Shimoni and Müller (1998), involves sharpened gold tubes with 100 μm outer diameter and 50 μm inner diameter; these tubes are propelled into the tissue and are used as the specimen carrier for the high-pressure freezer. No filler is needed. However, the yield of tubes that contain tissue is low (E. Shimoni, personal communication). At present, we find that the best technique for tissue is the use of a biopsy needle with 300 μm outer and 200 μm inner diameter (Hohenberg *et al*., 1996; Vanhecke *et al*., 2003), available as a part of a kit from either Leica or Bal-Tec. With the biopsy needle, a core of tissue can be excised, ejected and placed in the high-pressure freezing carrier, pre-loaded with the desired filler, in 30–40 s.

3.3. Fillers

When loading cell pastes or pieces of tissue into the specimen carrier of the high-pressure freezer, it is important to avoid air spaces, since the high compressibility of any gasses that may be present would prevent adequate pressure build-up. The high water content of typical buffers or culture

media is incompatible with good freezing; therefore, aqueous solutions lacking cryoprotectants are not suitable filler materials. A commonly used filler for tissue samples is 1-hexadecene, which does not penetrate the tissue and does not have apparent deleterious effects on tissues with brief contact (Studer *et al*., 1989). Although not compatible with certain freeze substitution protocols, aqueous buffers containing 20% dextran (molecular weight 40,000 Da; Sartori *et al*., 1993) as a filler are ideal for frozen-hydrated sections of both tissue and cell suspensions. Large molecular weight polymers have the advantage that they can be added to the medium of choice without significantly contributing to the osmolarity. Aqueous solutions of 10–20% sucrose have also been used as a filler for both tissue and cell suspensions. Sucrose acts as an internal cryoprotectant at this high concentration, since the increased osmolarity results in osmotic efflux of intracellular water. While this reduction in water content of the biological specimen makes vitreous freezing easier, it also alters the 'native state' of most cells and tissues.

3.4. Quality Variation

As noted above, freezing quality is strongly dependent on the water content, and cells or cellular compartments that have high water content are especially difficult to freeze well. For difficult specimens, such as soft animal tissue, the volume of material to be frozen should be kept as small as possible, with a thickness of 200 μm or less, such that freezing rate can be maximized. Freezing of low water content material may be successful for specimen thickness up to 300 μm. Even when using best practices, there is often considerable variability in freezing quality from one high-pressure freezing run to another, as well as within different areas of the same block of frozen material. Differences in freezing quality are even sometimes observed within a single cell. Thus, considerable time must always be spent at the electron microscope, to find the best-preserved areas. Very poor freezing can already be detected in the cryomicrotome, where irregularities in the block face give it a dull appearance after fine trimming, while with excellent freezing, ribbons of shiny sections are easily produced.

4. CRYOULTRAMICROTOMY

4.1. Development of Cryoultramicrotomy

Taken as a whole, cryoelectron microscopy of frozen-hydrated sections is a demanding technique. However, thanks to gradual technical progress over the years, the microtomy itself, given the appropriate equipment and an ideal specimen, is not much more difficult than is microtomy of plastic sections.

Cryoultramicrotomy started with Fernández-Morán (1952), and was taken up again by Bernhard and Leduc (1967). In both cases, the sections

were freeze-dried before electron microscopy. The earliest report of sectioning unfixed tissue at temperatures as low as –150°C appears to be by Hodson and Marshall (1970). Transfer of frozen-hydrated sections to the transmission electron microscope, for imaging in the fully hydrated state, was first described by Hutchinson *et al.* (1978). Early work in elemental analysis of frozen-hydrated tissue sections in the scanning electron microscope is reviewed by Gupta and Hall (1981). In these studies, however, specimens were not frozen in the vitreous state.

Important contributions to the development of ultramicrotomy of vitreously frozen specimens have come from biologists (Al-Amoudi *et al.*, 2003, 2005; Chang *et al.*, 1983; Dubochet *et al.*, 1987; Frederik *et al.*, 1982, 1984, 1993; McDowall *et al.*, 1983; Michel *et al.*, 1991, 1992; Richter, 1994, 1996; Richter *et al.*, 1991), as well as from microanalysts (Buchanan *et al.*, 1993; Edelmann, 1994; Shi *et al.*, 1996; Somlyo *et al.*, 1977, 1985; Sun *et al.*, 1995; Zierold, 1984, 1987, 1988). For an excellent, comprehensive review, see Sitte (1996).

The number of publications in which frozen-hydrated sections are used for biological research is increasing rapidly, as the technical hurdles are being overcome. These papers include studies of DNA and chromatin (Leforestier, 2001; Sartori Blanc *et al.*, 2001; Woodcock, 1994), bacteria (Al-Amoudi *et al.*, 2004; Matias *et al.*, 2003; Zhang *et al.*, 2004), skin (Norlén *et al.*, 2003, 2004) and liver (Hsieh *et al.*, 2006).

4.2. Cryoultramicrotomes

The cryoultramicrotome was conceived nearly simultaneously with the room temperature ultramicrotome. In the days of the high-speed (20,000 r.p.m.!) rotary microtome, Gessler and Fullam (1946) designed a cryo version. Shortly after the advent of the 'modern' ultramicrotome, Fernández-Morán (1952) described a cryounit. LKB, Sorvall and Reichert all made early commercial models. The development of cryoultramicrotomes is covered in detail by Sitte (1996).

Cryoultramicrotomes, up until the 1990s, were adequate for cutting Tokuyasu sections, but they were not fully adequate for routine work with frozen-hydrated sections. The key improvement was automatic maintenance of a stable temperature of –160°C or lower, without compromising smooth operation of the cutting cycle and the advance. Still, not all cryoultramicrotomes produced today are optimally designed for frozen-hydrated section work. In the selection of a microtome, it is important to ensure that the particular model is specifically designed for work with vitreously frozen specimens at –160°C or lower. The use of diamond knives is recommended (Michel *et al.*, 1992). For work with frozen-hydrated sections, these knives are available in a range of angles (see below), and are supplied without a liquid reservoir. An antistatic ionizer is needed during both trimming and sectioning, to help the sections to glide off the knife edge. Further

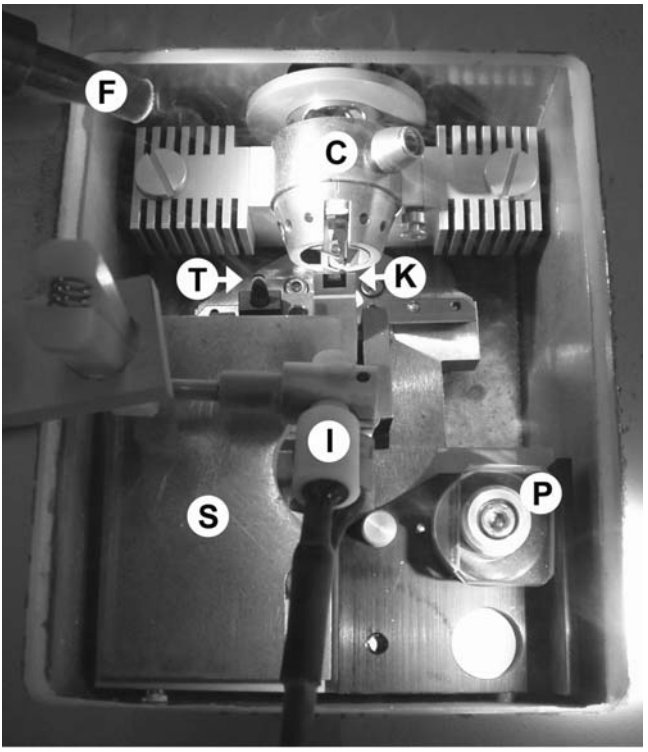


FIGURE 1. Cryochamber of an UCT ultramicrotome with an EM-FCS cryo kit (Leica, Vienna, Austria). The chamber is cooled to –160°C and is filled with cold, dry nitrogen vapor. One of the two auxiliary fiber-optic illuminators is seen, labeled (F). The specimen is held in the chuck (C). The knife stage is shown positioned for trimming, with the diamond trimming tool (T) adjacent to the specimen. For microtomy, the sectioning knife (K) is moved in front of the specimen, and a brass shelf (S) is pushed forward so that it is close to the knife edge. The shelf, fabricated in-house, provides a surface for collecting sections. Both diamond knives (T and K) are from Diatome (Biel, Switzerland). The head of the ionizer unit (I; Static Line II, Haug, Biel, Switzerland) is mounted ~3 cm from the knife edge. The glass screw press tool is mounted at (P). (See color plate)

accessories, such as supplementary cold fiber-optics illumination and custom equipment for mounting sections on grids, are also essential for achieving reliable results. Figure 1 shows the chamber of a suitably equipped cryoultramicrotome.

4.3. Preparation of High-pressure-frozen Specimens for Microtomy

After high-pressure freezing, all preparation steps are carried out in the cryochamber of the ultramicrotome, which is cooled to –160°C or lower,

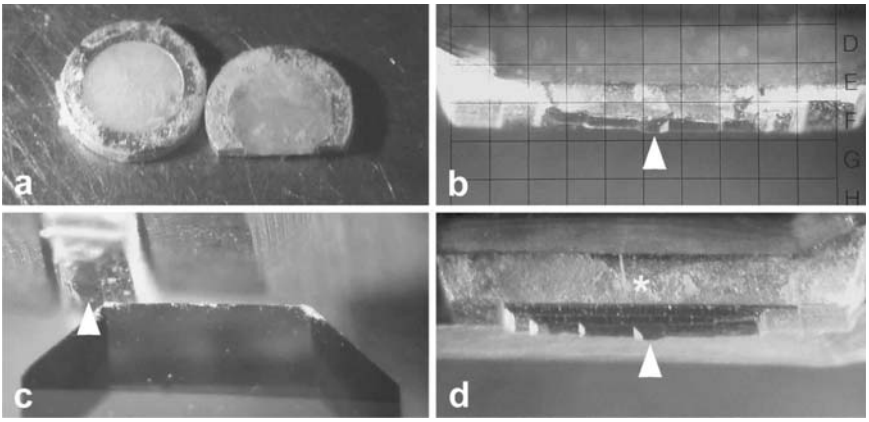


FIGURE 2. Trimming of high-pressure frozen specimens. (a) A 3 mm diameter aluminum specimen carrier for the HPM010 high-pressure freezer (Bal-Tec, Balzers, Liechtenstein) is shown, before and after trimming. The upper portion of the carrier has been removed, revealing the specimen and allowing inspection for air bubbles and suitable areas for microtomy. (b) A top view of the trimmed specimen carrier shows the trimmed block face, which is shaped like a low mesa (arrowhead), 100 μm square, as seen within a 250 μm graticle square of the stereomicroscope. (c) With the chuck rotated 90° from its position in (b), the diamond trimming tool is used to cut a 45° bevel in the metal on the bottom of the specimen carrier, leaving a thin ridge of specimen from which the mesa (arrowhead) has been cut. (d) Bottom view of the trimmed specimen carrier. The specimen is well supported during microtomy by the remaining metal bevel (*). The mesa is indicated by an arrowhead. (See color plate)

and is continuously flushed with cold, dry nitrogen gas so as to prevent frosting. When the metal tube specimen carrier is used (see Section 3.1), it is directly clamped in a suitable microtome chuck, and the surrounding metal is trimmed away with a diamond tool that has sides angled at 45°. The standard disk-type aluminum specimen carriers are clamped in a vice-type microtome chuck, as shown in Fig. 2, and the metal is similarly trimmed to expose the specimen, while leaving it well supported by the remaining portion of the carrier. In the case of the disk-type specimen carriers that reveal a dome of specimen after opening, the metal carrier does not need to be trimmed. For trimming, a microtome feed of 500 nm and a speed of 100 mm/s are used, and the antistatic ionizer is set on full power to keep material from sticking to the knife. With both types of specimen carrier, the specimen is held firmly without the need for 'cryoglue' (Richter, 1994). The final trimming step in all cases is the creation of a low mesa-shaped block face, with a square shape not more than 100 μm on a side. The small block face helps to reduce sectioning artifacts, and the square shape helps in the assessment of compression (see Section 7.2). With the disk-type carrier, the orientation of a piece of tissue can easily be seen, and a great deal of additional material is available for sectioning, once initial electron microscopic

examination has proven that a particular carrier contains optimally well frozen material.

4.4. Sectioning Parameters

Sectioning artifacts (see Section 7) can be reduced, or ease of microtomy can be improved, with the appropriate choice of cutting conditions. The exception is cutting temperature; in our hands, we found no clear evidence of the effect of temperature on section quality, within the range –140 to –160°C. Since we have found that the temperature in the center of the cryochamber is 5–10°C higher than in the bottom of the chamber, where the microtome's temperature sensor is located, we avoid temperature settings warmer than –150°C. We prefer a sectioning temperature setting of –160°C, which gives a generous margin of safety for devitrification. Ease of sectioning, and ease of collecting the sections, is improved, and the chance of frost accumulation is reduced, when the room humidity is not greater than ~40% at 20°C.

Knife marks can be reduced through the use of an undamaged diamond knife and through ensuring that the knife is kept free of frost and debris (Michel *et al*., 1992). After a few hours of use, frost particles become difficult to remove from the knife edge, so the knife should be exchanged. We find that diamond knives for cryo-use must be resharpened more often that those used for plastic sections.

Compression can often be reduced through the use of low-angle knives (Studer and Gnägi, 2000). Although some laboratories prefer the 45° angle that is standard in plastic section work, 35° knives are more commonly used for cryoultramicrotomy, and some laboratories use 25° knives (Zhang *et al*., 2004). For our specimens, 35° knives have proven to be the best choice. We find that wear is excessive with 25° knives, while compression is not significantly reduced. We have also observed that the power setting of the antistatic device can have an effect on compression. When ribbons of sections can be prompted to glide smoothly off the knife through 'tuning' of the power setting, compression is sometimes found to be reduced. Use of a very small block face (<100 μm in width) can also aid in reducing compression.

Chatter can be minimized through the firm mounting of the specimen in the microtome chuck, and through the use of a small block face. The use of very high cutting speeds (e.g. 50 mm/s rather than 0.5 mm/s) has been proposed as a way to minimize chatter (Al-Amoudi *et al*., 2005). In that report, the high cutting speed was also found to reduce crevasse depth, although it increased the numbers of crevasses. In our experience, if specimens are optimally frozen, sections free from chatter and crevasses can be obtained at 'normal' cutting speeds (0.2–1 mm/s). Thus, we find that high cutting speed, while accelerating knife wear, has little effect on sectioning quality. We typically use a cutting speed of 0.4 mm/s. Knife wear is a par-

ticular concern for tomographic work in which sections thicker than 100 nm are often desired, since knife wear increases with section thickness.

5. COLLECTION AND MOUNTING OF SECTIONS FOR TOMOGRAPHY

5.1. Types of Grids

Two types of TEM grids are recommended for tomography of frozen-hydrated sections: folding grids, and standard (single) grids that are coated with a Quantifoil film. Folding grids take the form of two TEM grids attached by a hinge (Fig. 3c). Sections are trapped between the two grids, so loss of sections is rare. Folding grids with a mesh size finer than 200 are not useful for tomography, because their double thickness causes occlusion at high tilt, for specimen areas not close to the centre of the grid opening. A hexagonal grid mesh is preferred, because the alignment of the grid bars relative to the tilt axis has less effect on high-tilt occlusion than when square-mesh grids are used. Folding grids are preferred for sections thicker than 200 nm, because thicker sections attach less well to grids. The Quantifoil film (Quantifoil MicroTools, Jena, Germany) is a 25 nm thick carbon film with a regular array of circular holes. We find that type R3.5/1 (3.5 μm diameter holes spaced 1 μm apart) is most suitable for tomography. The Quantifoil film is available on grids of several different materials and mesh sizes. Molybdenum 200 mesh grids are the most suitable, and are preferred over copper grids because the support film remains flat when cooled (Vonk, 2000; see also Section 6). The regular array of holes in Quantifoil films is also very useful for tracking the area of interest during tilting, and for relocating the area of the specimen after in-plane rotation, during collection of an orthogonal tilt series for double-tilt reconstruction (Mastronarde, 1997; Penczek *et al.*, 1995).

5.2. Support Films

A continuous support film is necessary for both types of grids described above. As mentioned in Section 6 and, as shown in Fig. 4, frozen-hydrated sections have a curved, undulating topology, so the points of contact with a flat support are limited. Thus, a continuous support film maximizes the occurrence of regions of contact. Another reason for the use of a continuous film is that it provides a substrate for colloidal gold particles, which are used as fiducial markers for alignment of the tilt series (see below). Since the Quantifoil grids already have a (perforated) film covering the grid squares, it is only necessary to add a thin (~10 nm) carbon coat, which is ideal for thin-section work. The folding grids require a formvar film, usually 50 nm thick, followed by the carbon film. For the folding grids, the support

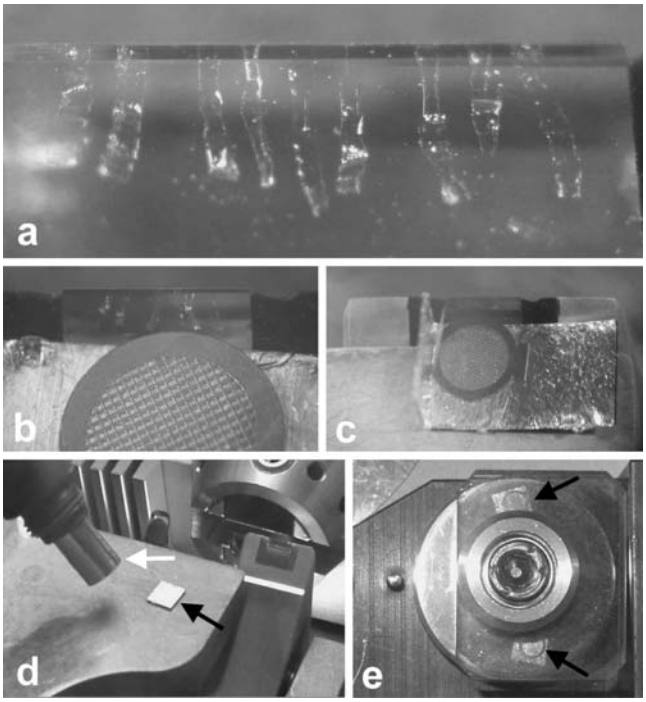


FIGURE 3. Collection and mounting of sections. (a) A ribbon of sections is cut, and then the knife is moved laterally. A series of ribbons is thus collected on the knife edge. Faint gold interference color can be seen in most sections, indicating a thickness of ~140 nm. In places, a purple color is seen, corresponding to a thickness of ~180 nm. Some sections are curled at the edges, or wrinkled. (b) A Quantifoil grid is placed on a fresh piece of indium foil, just adjacent to the knife edge, and sections are transferred to the center of the grid by means of a dog hair. (c) When folding grids are used, the indium foil is pre-folded to a 90° angle and, after placement of the sections, the indium foil is used to push the grid closed. (d) After the grid is latched, the indium foil is folded closed and the 'envelope' (arrow) is pressed with the polished metal rod (white arrow at left). (e) Quantifoil grids (arrows) are placed, still supported on indium foil, in the glass press tool. An impression of the grid squares in the indium foil indicates the evenness of pressing. (See color plate)

films are applied to only one side, so there will be only one set of films after folding.

5.3. Fiducial Markers

Alignment based on fiducial markers is robust and reliable, and is especially useful for the low-contrast images of frozen-hydrated sections. The markers can also aid in tracking and focusing during collection of the tilt series. The colloidal gold solution is applied either immediately after the

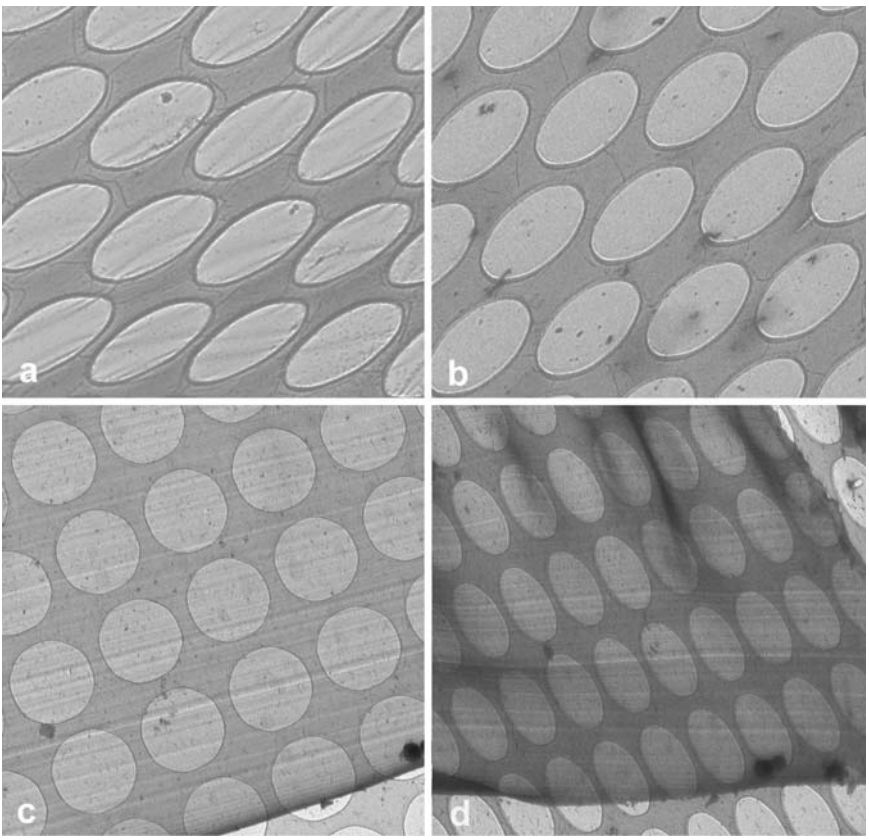


FIGURE 4. (a) Carbon film on a copper Quantifoil grid at –176°C; note wrinkling of the film. (b) Carbon film on a molybdenum Quantifoil grid at –176°C; note the flat film. Both (a) and (b) were recorded at 60° specimen tilt. (c) Frozen-hydrated section at 0° tilt; note the numerous straight knife marks, running from lower left to upper right. The knife marks are parallel to the cutting direction. Chatter, which would be seen as a periodic variation in thickness perpendicular to the knife marks, is absent. (d) The same frozen-hydrated section as in (c), at 60° tilt. The waviness of the section is indicated by curved knife marks and density shading due to increased thickness in projection. Suitable areas for tomography, where there is close contact of the section with the film and where the section is not bent, can be determined by such images, as described in the text. The holes in the Quantifoil film are 3.5 μm in diameter.

thin carbon film is deposited, or after the grids have been cleaned by glow-discharge. Some experimentation is needed to establish the optimal drop size and blotting conditions for each grid type. The particle size is usually 10–20 nm, depending on the magnification that will be used for collection of the tilt series. After storage, grids with fiducial markers should be glow-discharged again just before collecting frozen-hydrated sections, to aid in section attachment (Buchanan *et al*., 1993).

5.4. Collection of Sections

In the microtome cryochamber, the grids are placed near the knife edge, and on the same level as the knife, so that the knife edge and the grids are in focus and in the field of view of the stereomicroscope. This is accomplished by use of a platform (Figs 1 and 3c), which we have fabricated for our Leica FSC cryomicrotomy system. Newer Leica cryomicrotomes provide a platform, and certain models of Diatome cryodiamond knives have integrated platforms. On the platform (Fig. 3b and c), the grids are placed on short strips of fresh indium foil, 0.127 mm thick (Alfa Inorganics, Beverly, MA, USA). The indium aids in the folding and pressing of the folding grids, and it forms a protective envelope around them (Buchanan *et al*., 1993; Fig. 3d). For the Quantifoil grids, the indium foil again helps in the pressing operation (Fig. 3e). A few short ribbons of sections are accumulated on the knife edge, by translation of the knife after each ribbon is cut (Fig. 3a). The sections are then transferred to the adjacent grids by means of a hair or eyelash. The white hair of a Dalmatian dog (Forbes, 1986) is recommended, because it does not break at low temperature.

Suitable sections for tomography are pre-screened at the microtome. Especially wavy sections will not attach well to grids, and high local curvature will interfere with imaging at high tilt (Fig. 4d). As noted earlier, sections with a dull appearance indicate surface roughness due to poor freezing. The amount of compression (see Section 7.2) can be estimated at this stage, especially when a square block face is used. The 16 grid squares in the center of a 200 mesh grid are ideal for high-tilt, double-axis tomography, and we try to place four short ribbons of sections in this area.

5.5. Attachment of Sections to Grids

As noted above, since the frozen-hydrated sections are not flat (see Section 6), only limited areas become attached to the support film. Because the support film carries the gold fiducial markers necessary for alignment of the tilt series, movement of the section relative to the support film during collection of the tilt series will result in incorrect alignment. Even when there is no movement, reconstruction accuracy is highest in the plane of the support film that contains the fiducials. Alignment errors relative to the section are expected to increase as the gap between support film and section widens. Thus, it is important to maximize the area of contact between the lower surface of the section and the support film.

The sections are attached to the grids by pressing with a polished metal rod normally supplied with the microtome. Folding grids are placed inside an indium envelope prior to pressing with the rod (Fig. 3d). For Quantifoil grids, better results are obtained when the glass screw-type press tool supplied with the Leica cryoultramicrotome is used (Figs 1 and 3e). In that case,

the soft indium deforms upon pressing, and acts as a backing for the support film. Since the press tool directly contacts the sections on Quantifoil grids, use of the screw-type press tool provides better shielding against frost, and less likelihood that the sections will be pulled off the grid.

Grids are then placed in small covered grid boxes for transfer to storage under liquid nitrogen and then to the workstation of the cryotransfer TEM specimen holder. Because the sections are exposed on the Quantifoil grids, section loss may occur, and frost accumulation may be experienced, especially after storage for several days under liquid nitrogen. Thus, it is best to examine Quantifoil grids within a day of preparation. After storage for several days under liquid nitrogen, grids of either type should be transferred back to the cryochamber of the microtome, for a check that sections are still present, and in the desired central area on the grid. The grids should then be pressed again, to maximize attachment of the sections.

6. SELECTION OF SECTIONS SUITABLE FOR ELECTRON TOMOGRAPHY

Many of the sections cut in the ultramicrotome will be suitable for general cryoelectron microscopy imaging, but only a small percentage will be suitable for electron tomography. The area of interest must have good structural preservation, and a sufficient number (at least four) of well-distributed gold markers to enable alignment of the tilt series (markerless alignment may be possible in some cases; see Section 10). The area of interest should be near the center of the grid, so as to maximize the available tilt range for either single- or dual-axis data collection. The tilt range should be checked at a magnification low enough to image the whole grid square (the electron dose at this magnification will be minimal). A sensitive TV-rate camera is ideal for this operation.

As described above, frozen-hydrated sections tend to be curved even after pressing, typically with a few undulations in the cutting direction, as shown in Fig. 4d. This topology poses a problem for attachment to the support film, especially if the support film is itself not flat. Carbon support films on copper grids tend to wrinkle when exposed to liquid nitrogen temperatures, but the films remain flat when molybdenum grids are used (Fig. 4b), because molybdenum grids better match the thermal expansion characteristics of carbon films (Vonk, 2000).

For successful tomography, the area of interest in the section should not be more than 100 nm above the support film to which the gold fiducial markers are attached (Fig. 5). Proximity of the section to the carbon surface may be determined by noting the movement of the gold particles relative to features in the specimen, as the specimen is tilted. If the images are displayed so that the tilt axis is vertical, a visual comparison of the positions

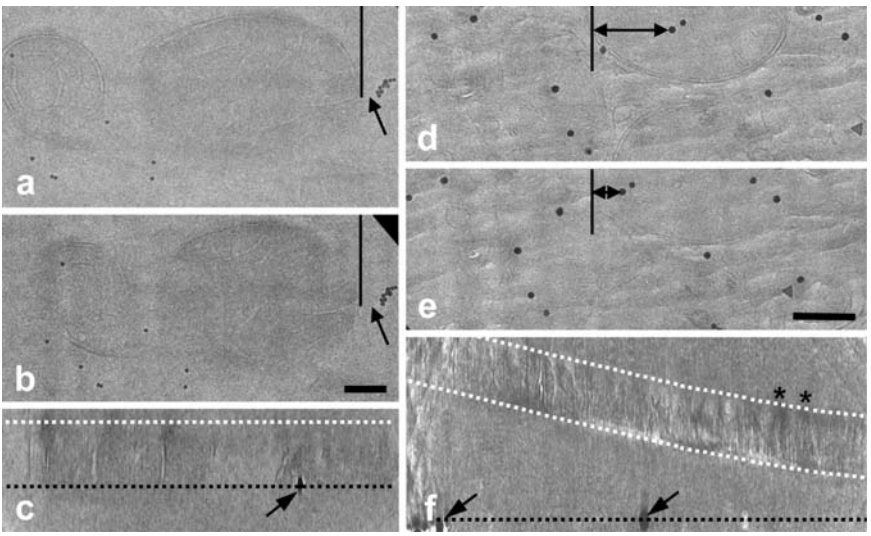


FIGURE 5. Projection images and tomographic cross-sections demonstrating good (a–c) and poor (d–f) attachment of the section to the support film. (a and b) Tilt images recorded at 0 and 30°, respectively, with the tilt axis vertical. Note that the spacing (arrow) between the edge of a mitochondrion (indicated by a black line) and a gold marker changes only slightly, as expected when the section is in close contact with the support film that carries 10 nm gold particles. (c) A gold marker (arrow) is on the lower surface of the section, as seen in a cross-section from the reconstruction (section between the dotted lines), indicating that the support film (black dotted line) is in contact with the section. (d and e) Tilt images at 0 and 30° indicating a wide separation between the section and the support film. The double arrows show that the spacing between a gold marker and the edge of a mitochondrion (indicated by a black line) varies greatly with tilt. The method for calculating the separation between the section and the support film is given in the text. (f) In cross-section, the section (between white dotted lines) is tilted and curved relative to the support film (black dotted line) that contains 20 nm gold markers (arrows). At (*), note the cross-sectional view of crevasses (see text and Fig. 8). Images a, b, d and e were recorded at an electron dose of 0.35 $e^{-}/Å^2$. For these single-axis reconstructions, the cross-sections shown are along the *x*-axis (perpendicular to the tilt axis). The individual gold markers in the images do not correspond to those shown in the projection images. Specimens: (a–c) isolated rat liver mitochondria, 135 nm thick; bar in (b) = 100 nm; (d–f) rat liver tissue, 200 nm thick; bar in (e) = 200 nm.

of gold markers relative to features in the specimen will indicate the degree of separation between the specimen and the support film (Fig. 5d and e).

The actual separation can be determined by the formula $z = (x_1 - x_2)/(2\sin\alpha)$, where z is the separation, $x_1 - x_2$ is the distance by which the marker has moved relative to a specimen feature after tilting, and α is the tilt half angle, for example 5° for a ±5° stereo pair or 15° for images recorded at 0 and 30° (Hama and Arii, 1987).

7. CHARACTERISTICS OF FROZEN-HYDRATED SECTIONS

Sectioning artifacts are well documented (Al-Amoudi *et al*., 2005; Chang *et al*., 1983; Frederik *et al*., 1982, 1984; Hsieh *et al*., 2002, 2006; Michel *et al*. 1991, 1992; Richter *et al*., 1991, 1994; Zierold, 1984). The most common artifacts are knife marks, chatter, compression and crevasses. Rarely observed, and of uncertain cause, is a fine pattern of striations, spaced ~15 nm apart and oriented in the sectioning direction. These can be seen throughout the depth of the tomographic reconstruction (Hsieh *et al*., 2002). We find that with good technique, all sectioning artifacts except for compression and lack of flatness can be minimized.

7.1. Knife Marks and Chatter

Knife marks and chatter are found on both frozen-hydrated and conventional plastic sections, and they can be ameliorated by similar methods for the two types of section. Knife marks are grooves that are often visible on both surfaces of the section (Al-Amoudi *et al*., 2005; Frederik *et al*., 1984; Richter *et al*., 1991). They can be controlled through the use of a fresh knife edge that is kept free of frost. Under ideal conditions, knife marks can be absent from fields as wide as 1 μm. Since knife marks are normally only ~10 nm deep, they do not affect the interior of the tomographic reconstruction. Mild knife marks can be seen on all of the frozen-hydrated sections shown in this chapter, and are especially evident in low-magnification images (Fig. 4c and d).

Chatter appears as a regular variation, by 10–30%, in section thickness with a periodicity of between 0.1 μm and several micrometers (Chang *et al*., 1983). Chatter can be minimized by making sure that the block is firmly mounted. Some investigators believe that friction of the section on the knife surface may contribute to chatter (Al-Amoudi *et al*., 2005). The frequency and severity of chatter may be influenced by cutting speed (Section 4.4; Al-Amoudi *et al*., 2005; Zierold, 1987). For our specimens, even low-magnification images show that the sections are normally free of chatter (Fig. 4c and d).

7.2. Compression

Compression is indicated by a shortening of the section in the direction of cutting, accompanied by an increase in thickness. In plastic sections, compression can be reversed by re-expanding the sections while floating on a water surface, but a liquid on which frozen-hydrated sections will float at temperatures below –135°C has not yet been found. Thus it is unknown if this method would work with frozen-hydrated sections. The amount of compression in frozen-hydrated sections varies with specimen type and section thickness, and is usually ~30–60% (Chang *et al*., 1983; Michel *et al*., 1992; Richter, 1994; Shi *et al*., 1996). Compression is clearly evident in Fig. 6a,

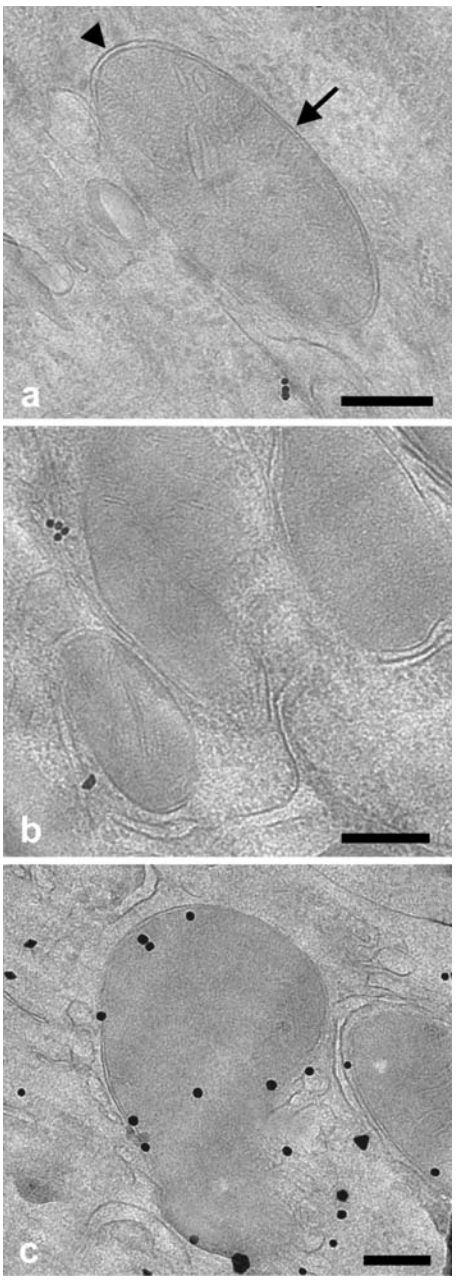


FIGURE 6. Appearance of mitochondria in frozen-hydrated sections of successively greater thickness. As the thickness increases, it becomes more important to use well-frozen specimens in order to avoid formation of crevasses. Shown here are nearly crevasse-free sections of mouse skin at 100 nm thickness (a), and rat liver at 200 nm (b) and 270 nm (c) thickness. Thickness for (a) and (b) was measured by electron energy-loss imaging; thickness for (c) was measured by tomographic reconstruction. These images were recorded at an electron dose of 3–5 $e^{-}/Å^{2}$, ~10 times higher than can be used for recording projection images in tomographic tilt series. In (a), compression is evident from the narrower membrane spacing in the cutting direction (arrow), compared with the spacing in the orthogonal, uncompressed direction (arrowhead). Bars = 200 nm.

where the spacing between outer and inner membranes on the periphery of mitochondria is reduced in the direction of compression.

Compression can sometimes be limited by reducing the angle of the microtome knife from the commonly used 45° to 35, 30 or even 25° (Michel *et al*., 1992; Richter, 1994; Zhang *et al*., 2004). Since low-angle knives are very fragile, the oscillating diamond knife (Studer and Gnägi, 2000) was developed as an alternative, with the oscillation acting to decrease the effective knife angle. Recently, the oscillating knife has been adapted for cutting frozen-hydrated sections; however, results were inconsistent, and only 3% of the sections cut were uncompressed (Al-Amoudi *et al*., 2003). In spite of this disappointing report, further investigation of use of the oscillating knife, with specimens of varying thickness and material may be worthwhile.

In frozen-hydrated sections, compression appears to be non-uniform with respect to different cellular components. A survey of our own images, as well as images published by others, confirms reports that small particles and certain rigid structures such as ribosomes (Dubochet and Sartori Blanc, 2001) and DNA crystals (Leforestier *et al*., 2001) may retain their shape despite distortion in the overall shape of the cell or a membranous organelle. For example, we have observed that desmosomes in mouse skin resist compression in frozen-hydrated sections (Hsieh *et al*., 2004). Therefore, it is not possible to restore the original specimen geometry computationally in tomograms of frozen-hydrated sections, as is sometimes done with plastic sections to compensate for their *uniform* thinning due to mass loss from electron irradiation (Luther *et al*., 1988; van Marle *et al*., 1995; Chapter 1 of this volume).

7.3. Crevasses

Crevasses, named after the cracks that form in glaciers, are an artifact unique to frozen-hydrated sections (Chang *et al*., 1983). We have characterized crevasses using electron tomography (Hsieh *et al*., 2002). They are grooves perpendicular to the cutting direction, usually spaced ~100 nm apart, and with typical dimensions of 100 nm in length, 20 nm in width and 10–50 nm in depth. Most commonly, the depth is little more than 10 nm; thus, in sections at least 100 nm thick, crevasses do not interfere with tomographic study of the section interior. They are found only on the surface of the section that constituted the block face when the section was cut. Various, and contradictory, mechanisms for crevasse formation have been proposed, based on mechanical models (e.g. Al-Amoudi *et al*., 2005; Richter *et al*., 1991). Figure 5 shows typical crevasses, both in projection images and in tomographic cross-section, as do Figs 7d and 8.

We have observed that, when the quality of high-pressure freezing is excellent, based on ultrastructural preservation in the frozen-hydrated sections, the severity of crevasses is reduced. This is especially true for thicker sections. While sections 100 nm or less in thickness can often be cut without

crevasses, only exceptionally well frozen material can be used to produce crevasse-free sections when the thickness exceeds 200 nm (Fig. 6; Hsieh *et al*., 2006).

As noted in Section 2.2, several studies have suggested that high-pressure freezing (unlike plunge-freezing) does not produce uniformly amorphous or vitreous ice, but rather mixtures of ice phases and types, some of which are microcrystalline. Nevertheless, the same studies reported that crystalline ice is not detected in material that has been optimally high-pressure frozen. We have found that, when electron diffraction gives clear evidence of crystalline ice, there is a significant increase in severity of crevasses (Fig. 7), consistent with earlier reports (Chang *et al*., 1983; Erk *et al*.,

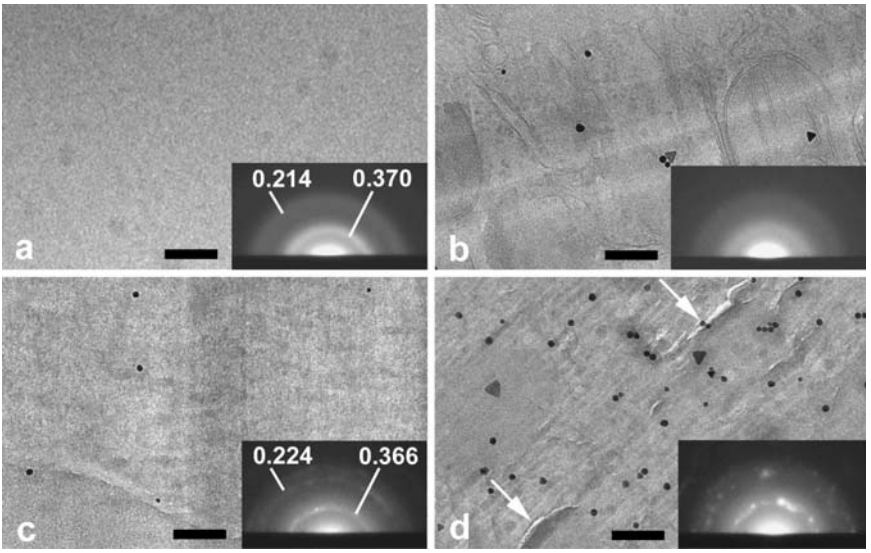


FIGURE 7. Forms of ice in frozen-hydrated sections, as seen in images and corresponding electron diffraction patterns. (a) Vitreous ice prepared by plunge-freezing of a 100 nm thick layer of pure water. In the diffraction pattern, the characteristic diffuse rings at spacings corresponding to 0.214 and 0.370 nm are clearly seen. (b) Vitreous ice from a high-pressure-frozen 200 nm thick rat liver section. The diffraction pattern is less clear for sections of frozen-hydrated tissue in this thickness range, due to a background of diffuse scattering from the biological material. (c) High-pressure-frozen liver section 100 nm thick, with a diffraction pattern showing sharp arcs, corresponding to spacings of 0.366 and 0.224 nm, that are characteristic of cubic ice. Such diffraction patterns are rarely observed, suggesting that the content of cubic ice in this specimen is much greater than usual. (d) A combination of vitreous and hexagonal ice in a 100 nm thick liver section. The presence of hexagonal ice, due to poor freezing, indicated by discrete spots in the diffraction pattern, leads to severe sectioning artifacts, such as crevasses and tears (arrows); the section quality is so poor that artifacts largely obscure the features of the specimen. Bars = 200 nm

1998; Frederik *et al*., 1984; Zierold, 1984). Thus, it may be that increased crevasse severity correlates with an elevated content of microcrystalline ice, due to less successful suppression of crystal formation in the specimen during a particular high-pressure freezing run.

However, we find that crevasses can be observed even when the electron diffraction pattern shows no sign of crystalline ice (see Fig. 3 in Hsieh *et al*., 2006). Thus, the tendency for crevasse formation is either unrelated to minor crystalline ice content, or else it is extremely sensitive to it, such that the content of crystalline ice needed to induce crevasses is below the detection level in electron diffraction patterns. As described in Section 2.2, the sharp diffraction rings and spots associated with crystalline ice tend to occur near or within the diffuse rings due to vitreous ice. These diffuse rings, and the high background of diffuse scattering from the biological material in the section, might mask weak reflections from a minor amount of crystalline ice.

8. IMAGING CONDITIONS FOR TILT SERIES COLLECTION

Tilt series of frozen-hydrated sections are recorded in the electron microscope in the same way as are those of plunge-frozen specimens. Imaging conditions that maximize the signal-to-noise ratio within acceptable cumulative electron dose limits for prevention of damage to macromolecular structure are discussed elsewhere (Grimm *et al*., 1989; Hsieh *et al*., 2006; Marko *et al*., 1999, 2000; McEwen *et al*., 2002; Chapter 4 of this volume). However, frozen-hydrated sections are sensitive to electron irradiation in ways that plunge-frozen specimens are not.

As a result of the common problem of poor attachment of frozen-hydrated sections to the support film, electron irradiation can cause the section to move relative to the support film (Fig. 8a and b). As explained in Section 5.5, tilt series alignment that is based on fiducial markers attached to the support film will fail if such movement occurs, although markerless alignment (see Section 10) might succeed. Thus, if images of the untilted specimen, recorded before and after collection of a tilt series, show movement of the markers relative to the specimen features, the reconstruction should not be attempted using marker-based alignment.

It is well known that surface artifacts in frozen-hydrated sections can be 'ironed out' by electron irradiation (Sartori Blanc *et al*., 1998; Fig. 8c and d). This improves the appearance of the section for general imaging, but is not useful for electron tomography because the electron dose necessary for smoothing (~100 $e^-/Å^2$) exceeds the typical maximum permissible cumulative dose for the tilt series (30–80 $e^-/Å^2$).

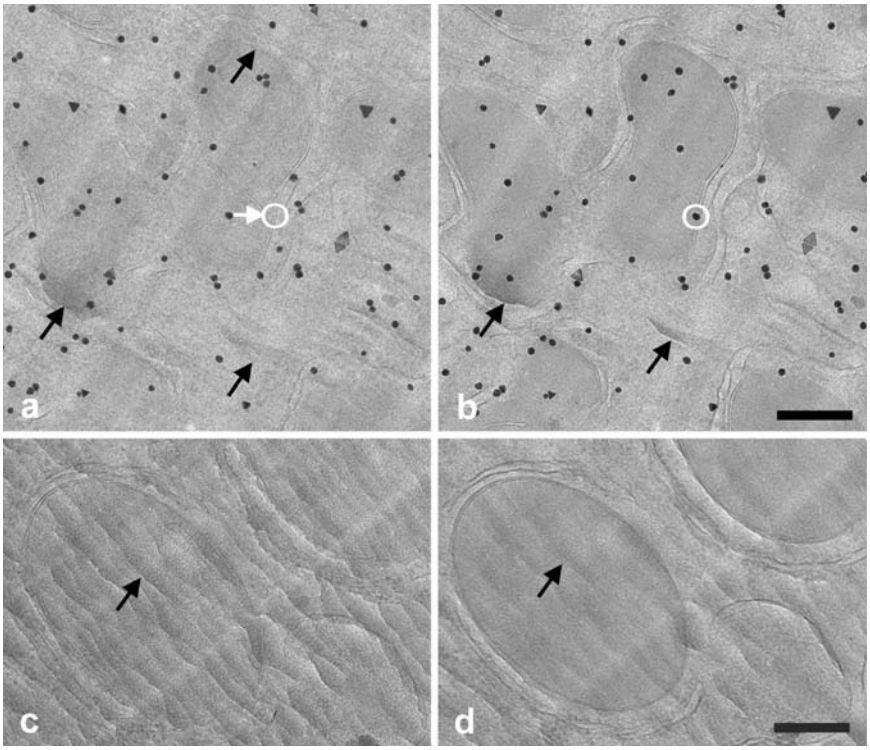


FIGURE 8. Effects of electron irradiation. (a and b) Untilted images recorded before (a) and after (b) the recording of a tomographic tilt series. The accumulated electron dose was 1 $e^-/Å^2$ for (a), and 50 $e^-/Å^2$ for (b). The electron irradiation during the recording of the tilt series was not sufficient to flatten out crevasses (black arrows), but it did cause the section to move relative to the gold markers on the support film (circles and white arrow); a reconstruction was therefore not attempted. The dose used to record (b) was 3 $e^-/Å^2$; thus, it is less noisy than (a), which was recorded at 0.35 $e^-/Å^2$, a typical electron dose for a projection image that is a member of a tilt series. (c and d) Example of flattening of severe crevasses (arrows) in a frozen-hydrated section. The accumulated dose was 3 $e^-/Å^2$ in (c) and 100 $e^-/Å^2$ in (d). Sections are rat liver, 200 nm thick. Bars = 200 nm.

9. SAMPLE TOMOGRAMS

An example of a tomographic reconstruction of a section of high-pressure-frozen rat liver tissue is presented in Fig. 9. The improved clarity of structural detail in the tomographic slices, as compared with the projection image though the whole section, is striking. Membranes of mitochondria and the endoplasmic reticulum are clearly visible, as are ribosomes, particularly those that are attached to the endoplasmic reticulum. These reconstructions show that the invaginations (cristae) of the mitochondrial inner membrane

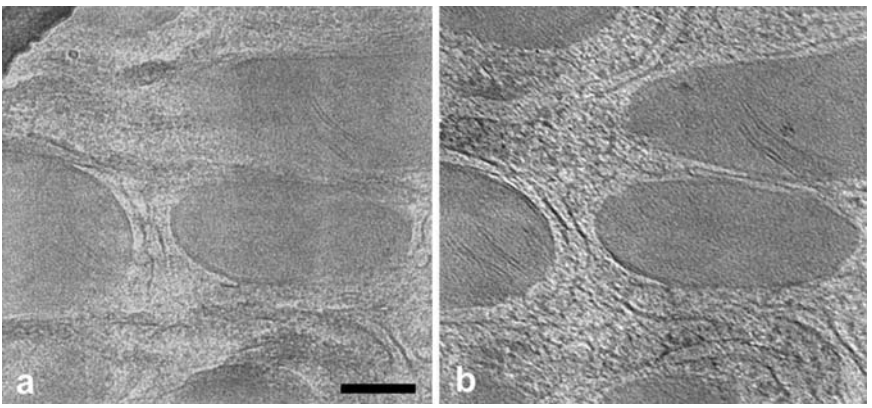


FIGURE 9. Sample tomographic reconstruction of a 200 nm thick liver section. Zero degree tilt series image (a), and 5.4 nm thick *z*-slice from the corresponding reconstruction (b) The angular range of the tilt series was ±60°, with an angular increment of 2°. Imaging was done at 400 kV with zero-loss energy filtering; the total electron dose was 50 $e^{-}/Å^{2}$. Bar = 200 nm. (Tomogram also used in Hsieh *et al.*, 2006.)

connect to the boundary region of this membrane by means of small circular or slit-like openings, as was previously reported for conventionally prepared (Mannella and Frey, 2000; Mannella *et al.*, 1997) and plunge-frozen intact mitochondria (Mannella *et al.*, 2001; Nicastro *et al.*, 2000). Since these 'crista junctions' have been implicated in the regulation of diffusion of metabolites and proteins between mitochondrial compartments (Mannella, 2005; Mannella *et al.*, 1997, 2001), the confirmation of their existence in frozen-hydrated tissue was a bioenergetically significant finding.

We have also carried out tomography of frozen-hydrated sections of yeast (C.-E. Hsieh *et al.*, unpublished), skin (Hsieh *et al.*, 2004) and cyanobacteria (Ting *et al.*, 2005). Such material proved much easier to work with than soft animal tissue, because excellent high-pressure freezing, which is essential for high-quality sections, could be more routinely obtained. We advise investigators to develop skills and acquire experience with such specimens prior to working on more difficult materials such as soft tissue.

10. CURRENT DIFFICULTIES AND FUTURE PROSPECTS

Electron tomographic reconstruction of frozen-hydrated sections offers tremendous promise, as the only currently available technique for 3D imaging of large cells and cells in tissue, in their native state, at resolutions of 5 nm or better.

Currently, the most significant problem for frozen-hydrated sections is the poor yield of sections that can be used for electron tomographic data

collection. A skilled, dedicated individual can achieve a yield of 60% for sections that have sufficient quality for imaging in the electron microscope. However, very few of these sections will satisfy all of the necessary conditions for successful tomography. Interesting regions, free of contaminants and severe artifacts, must be located centrally on the grid and in the grid square, so that a complete tilt series can be recorded without occlusion by a grid bar or the frame of the specimen holder. Gold fiducial markers (Brandt *et al*., 2001a; Chapter 5 of this volume) must be appropriately distributed on the area of interest. Most importantly, the section must be in close proximity to, and firmly attached to, the support film or grid (see Sections 5.5 and 6). Because of the need to keep the cumulative electron dose low, some problems, such as imperfect freezing, may not be discovered until after the tomogram has been computed, especially for thicker specimens. Currently, good tomograms are obtained for fewer than 5% of the frozen-hydrated tissue sections that are cut at the cryoultramicrotome.

In principle, alignment problems caused by lack of proximity between section and support film, and by poor attachment of the section to the support film that carries the fiducial markers could be overcome by markerless alignment methods (Brandt *et al*., 2001b; Winkler and Taylor, 2005; Chapter 6 of this volume). However, for beam-sensitive frozen-hydrated specimens, attainment of high resolution requires fractionation of the total allowable dose to the extent that each individual image has a very low signal-to-noise ratio; this makes alignment by correlation methods problematic. In addition, under very low dose conditions, the dense gold markers become essential for tracking and focusing on the specimen during collection of the tilt series. Thus, better means of enhancing attachment of sections to the support film are being explored, involving 'functionalization' of the support films by structural and chemical surface modification, as well as improvements in mechanical clamping methods.

Progress has been made in controlling sectioning artifacts, but further work is needed before we can fully understand the correlation between freezing quality and crevasse formation. Also needed is a reliable means of reducing compression; such a method might involve further developments in diamond knife technology, including the oscillating diamond knife (Al-Amoudi *et al*., 2003; Studer and Gnägi, 2000). The investigation of the causes and prevention of sectioning artifacts is complicated by the fact that contributing factors combine in different ways, and solutions tend to be specimen dependent.

We are also investigating the use of focused ion beam instrumentation as an alternative approach to cryoultramicrotomy, for the purpose of thinning bulk frozen-hydrated material for cryoelectron tomography (Marko *et al*., 2006). This method has the potential to avoid both the sectioning artifacts and the attachment problems encountered with mechanically cut sections.

ACKNOWLEDGEMENTS

We give special thanks to Shanling Shi of Unilever Co. for introducing us to the technique of cryoultramicrotomy of high-pressure-frozen specimens. We also thank Daniela Nicastro for a careful review of the manuscript. Development of cryoelectron tomography at the Wadsworth Center's Resource for Visualization of Biological Complexity is supported by NIH grant RR01219 (J. Frank, P.I.), through the NCRR Biomedical Research Technology Program, with additional support from Wadsworth Center's Electron Microscopy Core.

REFERENCES

Al-Amoudi, A., Dubochet, J., Gnägi, H., Lüthi, W. and Studer, D. (2003). An oscillating cryoknife reduces cutting-induced deformation of vitreous ultrathin sections. *J. Microsc*. **212**:26–33.

Al-Amoudi A., Norlén, L. P. O. and Dubochet J. (2004). Cryo-electron microscopy of vitreous sections of native biological cells and tissues. *J. Struct. Biol.* **148**:131–135.

Al-Amoudi, A., Studer, D. and Dubochet, J. (2005). Cutting artifacts and cutting process in vitreous sections for cryo-electron microscopy. *J. Struct. Biol.* **150**:109–121.

Baumeister, W. (2002). Electron tomography: towards visualizing the molecular organization of the cytoplasm. *Curr. Opin. Struct. Biol*. **12**:679–684.

Baumeister, W. and Steven, A. C. (2000). Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem. Sci.* **25**:624–631.

Beck, M., Förster, F., Ecke, M., Plitzko, J. M., Melchior, F., Gerisch, G., Baumeister, W. and Medalia, O. (2004). Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. *Science* **306**:1387–1390.

Bernhard, W. and Leduc, E. (1967). Ultrathin frozen sections. *J. Cell Biol*. **34**:757–771.

Böhm, J., Frangakis, A. S., Hegerl, R., Nickell, S., Typke, D. and Baumeister, W. (2000). Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proc. Natl Acad. Sci. USA* **97**:14245–14250.

Böhm, J., Lambert, O., Frangakis, A. S., Letellier, L. Baumeister, W. and Rigaud, J. L. (2001). FhuA-mediated phage genome transfer into liposomes: a cryo-electron tomography study. *Curr. Biol*. **11**:1168–1175.

Bongini, L., Fanelli, D., Piazza, F., De los Rios, P., Sandin, S. and Skoglund, U. (2004). Freezing immunoglobulins to see them move. *Proc. Natl Acad. Sci. USA* **101**:6466–6471.

Brandt, S., Heikkonen, J. and Engelhardt, P. (2001a). Multiphase method for automatic alignment of transmission electron microscope images using markers. *J. Struct. Biol*. **133**:10–22.

Brandt, S., Heikkonen, J. and Engelhardt, P. (2001b). Automatic alignment of transmission electron microscope tilt series without fiducial markers. *J. Struct. Biol*. **136**:201–213.

Bretschneider. T., Jonkman, J., Kohler, J., Medalia, O., Barisic, K. Weber, I., Selzer, E. H. K., Baummeister, W. and Gerisch, G. (2002). Dynamic organization of the actin system in the motile cells of *Dictyostelium*. *J. Muscle Res. Cell Motil*. **23**:639–649.

Buchanan, R. A., Leapman, R. D., O'Connell, M. F., Reese, T. S. and Andrews, S. B. (1993). Quantitative scanning transmission electron microscopy of ultrathin cryosections: subcellular organelles in rapidly frozen liver and cerebellar cortex. *J. Struct. Biol.* **110**:244–255.

Chang, J.-J., McDowall, A. W., Lepault, J., Freeman, R.,Walter, C. A. and Dubochet, J. (1983). Freezing, sectioning and observation artifacts of frozen hydrated sections for electron microscopy. *J. Microsc.* **132**:109–123.

Craig, S. and Staehelin, L. A. (1988). High pressure freezing of intact plant tissues. Evaluation and characterization of novel features of the endoplasmic reticulum and associated membrane systems. *Eur. J. Cell Biol*. **46**:81–93.

Crowther, R. A., DeRosier, D. J. and Klug, A. (1970). The reconstruction of a three-dimensional structure from its projections and its applications to electron microscopy. *Proc. R. Soc. B* **317**:319–340.

Cyrklaff, M., Risco, C., Fernandez, J. J., Jimenez, M. V., Esteban, M., Baumeister, W. and Carrascosa, J. L. (2005). Cryo-electron tomography of *Vaccinia* virus. *Proc. Natl Acad. Sci. USA* **102**:2772–2777.

Dahl, R. and Staehelin, L. A. (1989). High-pressure freezing for the preservation of biological structure: theory and practice. *J. Electron Microsc. Tech*. **13**:165–174.

Dierksen, K., Typke, D., Hegerl, R. and Baumeister, W. (1993). Towards automatic electron tomography. II. Implementation of autofocus and low-dose procedures. *Ultramicroscopy* **49**:109–120.

Dierksen, K., Typke, D., Hegerl, R., Walz, J., Sackmann, E. and Baumeister, W. (1995). Three-dimensional structure of lipid vesicles embedded in vitreous ice and investigated by automated electron tomography. *Biophys. J*. **68**:1416–1422.

Dubochet, J., Adrian, M., Chang, J. J., Homo, J.-C., Lepault, J., McDowall, A. W. and Schultz, P. (1988). Cryo-electron microscopy of vitrified specimens. *Q. Rev. Biophys.* **21**:129–228.

Dubochet, J., Adrian, M., Chang, J.-J., Lepault, J. and McDowall, A. (1987). Cryoelectron microscopy of vitrified specimens. In *Cryotechniques in Biological Electron Microscopy* (R. A. Steinbrecht and K. Zierold, eds). Springer, Berlin, pp. 114–131.

Dubochet, J. and Sartori Blanc, N. (2001). The cell in absence of aggregation artifacts. *Micron* **32**:91–99.

Echlin, P. (1992). *Low-temperature Microscopy and Microanalysis*. Plenum, New York.

Edelman, L. (1994). Optimal freeze-drying of cryosections and bulk specimens for X-ray microanalysis. *Scanning Microsc.* **Suppl. 8**:67–81.

Erk, I., Nicolas, G., Carloff, A. and Lepault, J. (1998). Electron microscopy of frozen biological objetcs: a study using cryosectioning and cryosubstitution. *J. Microsc*. **189**:236–248.

Fernández-Morán, H. (1952). Application of the ultrathin freeze sectioning technique to the study of cell structures with the electron microscope. *Arch. Fysik*. **4**:471–483.

Fernández-Morán, H. (1960). Low-temperature preparation techniques for electron microscopy of biological specimens based on rapid freezing with liquid helium II. *Annu. NY Acad. Sci*. **85**:689–713.

Forbes, M. S. (1986). Dog hairs as section manipulators. *EMSA Bull.* **16**:67.

Förster, F., Medalia, O., Zauberman, N., Baumeister, W. and Fass, D. (2005). Retrovirus envelope protein complex structure *in situ* studied by cryo-electron tomography. *Proc. Natl Acad. Sci. USA* **102**:4729–4734.

Frangakis, A. S., Böhm, J., Förster, F., Nickell, S., Nicastro, D., Typke, D., Hegerl, R. and Baumeister, W. (2002). Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proc. Natl Acad. Sci. USA* **99**:14153–14158.

Frangakis, A. S. and Förster, F. (2004). Computational exploration of structural information from cryo-electron tomograms. *Curr. Opin. Struct. Biol*. **14**:325–331.

Frank, J. (1996). *Three-dimensional Electron Microscopy of Macromolecules*. Academic Press, San Diego.

Frank, J. (2006). *Three-dimensional Electron Microscopy of Macromolecular Assemblies—Visualization of Biological Molecules in Their Native State*. Oxford University Press, New York.

Frank, J., Wagenknecht, T., McEwen, B. F., Marko, M., Hsieh, C.-E. and Mannella, C. A. (2002). Three-dimensional imaging of biological complexity. *J. Struct. Biol*. **138**:85–91.

Frederik, P. M., Bomans, P. H. H. and Stuart, M. C. A. (1993). Matrix effects and the induction of mass loss or bubbling by the electron beam in vitrified hydrated specimens. *Ultramicroscopy* **48**:107–119.

Frederik, P. M., Busing, W .M. and Persson, A. (1982). Concerning the nature of the cryosectioning process. *J. Microsc.* **125**:167–175.

Frederik, P. M., Busing, W. M. and Persson, A. (1984). Surface defects on thin cryosections. *Scanning Electron Microsc.* **1**:433–443.

Gessler, A. E. and Fullam, E. F. (1946). Sectioning for the electron microscope accomplished by the high speed microtome. *Am. J. Anat.* **78**:245–279.

Gilkey, J. C. and Staehelin, L. A. (1986). Advances in ultrarapid freezing for the preservation of cellular ultrastructure. *J. Electron Microsc. Tech*. **3**:177–210.

Grimm, R., Singh, H., Rachel, R., Typke, D., Zilling, W. and Baumeister, W. (1998). Electron tomography of ice-embedded prokaryotic cells. *Biophys. J.* **74**:1031–1042.

Grünewald, K., Desai, P., Winkler, D. C., Heyman, J. B., Belnap, D. M., Baumeister, W. and Steven, A. C. (2003). Three-dimensional structure of herpes simplex virus from cryo-electron tomography. *Science* **302**:1396–1398.

Grünewald, K., Medallia, O., Gross, A., Steven, A. and Baumeister, W. (2002). Prospects of electron cryotomography to visualize macromolecular complexes inside cellular compartments: implications of crowding. *Biophys. Chem*. **100**:577–591.

Gupta, B. L. and Hall, T.A. (1981). The x-ray microanalysis of frozen-hydrated sections in scanning electron microscopy: an evaluation. *Tissue and Cell* **13**:623–643.

Hama, K. and Arii, T. (1987). Three-dimensional analysis of high-voltage electron microscope tilt images: methods and problems. *J. Electron Microsc. Tech*. **6**:185–192.

Hess, M. W., Muller, M., Debbage, P. L., Vetterlein, M. and Pavelka, M. (2000). Cryopreparation provides new insight into the effects of brefeldin A on the structure of the HepG2 Golgi apparatus. *J. Struct. Biol.* **130**:63–72.

Hodson, S. and Marshall, J. (1970). Ultracryotomy: a technique for cutting ultrathin sections of unfixed biological tissues for electron micrscopy. *J. Microsc*. **91**:105–117.

Hohenberg, H., Tobler, M. and Müller, M. (1996). High-pressure freezing of tissue obtained by fine-needle biopsy. *J. Microsc.* **183**:1–7.

Hsieh, C., He, W., Marko, M. and Stokes, D. L. (2004). 3D Tomographic map of desmosome from frozen-hydrated skin sections. *Microsc. Microanal*. **10** (**Suppl. 2**):1188CD.

Hsieh, C.-E., Leith, A., Mannella, C.A., Frank, J. and Marko, M. (2006). Towards high-resolution three-dimensional imaging of native mammalian tissue: electron tomography of frozen-hydrated rat liver sections. *J. Struct. Biol.* **153** (in press).

Hsieh, C.-E., Marko, M., Frank, J. and Mannella, C. A. (2002). Electron tomographic analysis of frozen-hydrated tissue sections. *J. Struct. Biol.* **138**:63–73.

Hsieh, C.-E., Marko, M., Leith, A., Frank, J. and Mannella, C. A. (2003). Electron tomographic comparison of frozen-hydrated and freeze-substituted sections of high-pressure frozen rat-liver tissue. *Microsc. Microanal.* **9 (Suppl. 2)**:1178CD.

Hutchinson, T. E., Johnson, D. E. and MacKenzie, A. P. (1978). Instrumentation for direct observation of frozen hydrated specimens in the electron microscope. *Ultramicroscopy* **3**:315–324.

Kellenberger, E. (1987). The response of biological macromolecules and supremolecular structures to the physics of specimen cryopreparation. In *Cryotechniques in Biological Electron Microscopy* (R. A. Steinbrecht and K. Zierold, eds). Springer, Berlin, pp. 35–63.

Koster, A. J., Grimm, R., Typke, D., Hegerl, R., Stoschek, A., Walz, J. and Baumeister, W. (1997). Perspectives of molecular and cellular electron tomography. *J. Struct. Biol*. **120**:276–308.

Kürner, J., Medalia, O., Linaroudis, A. A. and Baumeister, W. (2004). New insights into the structural organization of eukaryotic and prokaryotic cytoskeletons using cryo-electron tomography. *Exp. Cell Res.* **301**:38–42.

Kürner, J., Frangakis, A. S. and Baumeister, W. (2005). Cryo-electron tomography reveals the cytoskeletal structure of *Spiroplasma melliferum. Science* **307**:436–438.

Leapman, R. D. (2005). Novel techniques in electron microscopy. *Curr. Opin. Neurobiol*. **14**:591–598.

Leforestier, A., Dubochet, J. and Livolant. F. (2001). Bilayers of nucleosome core particles. *Biophys. J.* **81**:2414–2421.

Leis, A., Andrees, L., Gruska, M., Al-Amoudi, A., Sartori, A., Dubochet, J. and Baumeister, W. (2005). Cryo-electron tomography and fluorescence microscopy of unicellular algae in vitreous cryosections. *Microsc. Microanal.* **11**(**Suppl. 2**):330CD.

Lepault, J., Bigot, D., Studer, D. and Erk, I. (1997). Freezing of aqueous specimens: an X-ray diffraction study. *J. Microsc*. **187**:158–166.

Lucic, V., Förster, F. and Baumeister, W. (2005b). Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem.* **74**:833–865.

Lucic, V., Yang, T., Schweikert, G., Förster, F. and Baumeister, W. (2005a). Morphological characterization of molecular complexes present in the synaptic cleft. *Structure* **13**:423–434.

Luther, P. K., Lawrence, M. C. and Crowther, R. A. (1988) A method for monitoring the collapse of plastic sections as a function of electron dose. *Ultramicroscopy* **24**:7–18.

Mannella, C. A. (2006). The relevance of mitochondrial membrane topology to mitochondrial function. *Biochem. Biophys. Acta* **1762**:140–147.

Mannella, C. A., Buttle, K., Marko, M. (1997). Reconsidering mitochondrial structure: new views of an old organelle. *Trends Biochem. Sci*. **22**:37–38.

Mannella, C. A. and Frey, T. (2000). The internal structure of mitochondria. *Trends Biochem. Sci.* **25**:319–324.

Mannella, C. A., Pfeiffer, D. R., Bradshaw, P. C., Moraru, L. I., Slepchenko, L. B., Loew, L. M., Hsieh, C.-E., Buttle, K. and Marko, M. (2001). Topology of the mitochondrial inner membrane: dynamics and bioenergetic implications. *IUBMB Life* **52**:93–100.

Marko, M., Hsieh, C.-E., Mannella, C. A. and Frank, J. (2002) .Electron tomography of frozen-hydrated specimens: application to tissue sections. In *Proceedings of the 15th International Congress on Electron Microscopy* (Cross, R., ed.), Microscopy Society of Southern Africa, Onderspoort, SA, **2**:205–206.

Marko, M., Hsieh, C.-E., Mannella, C. A. and McEwen, B. (1999). Imaging considerations for cryo-tomography of organelles and whole cells at high accelerating voltage. *Microsc. Microanal*. **5**(**Suppl. 2**):414–415.

Marko, M., Hsieh, C., MoberlyChan, W., Mannella, C. A. and Frank, J. (2006). Focused ion beam milling of vitreous water: prospects for an alternative to cryo-ultramicrotomy. *J. Microsc*. **212**:42–47.

Marko, M., Hsieh, C.-E., Rath, B. K., Mannella, C. A. and McEwen, B. F. (2000). Electron tomography of frozen-hydrated samples. *Microsc. Microanal.* **6**(**Suppl. 2**):310–311.

Mastronarde, D. N. (1997). Dual-axis tomography: an approach with alignment methods that preserve resolution. *J. Struct. Biol*. **120**:343–352.

Mastronarde, D. N. (2005). Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**:36–51.

Matias, V. R. F., Al-Amoudi, A., Dubochet, J. and Beveridge, T. J. (2003). Cryo-transmission electron microscopy of frozen-hydrated sections of gram-negative bacteria. *J. Bacteriol*. **185**:6112–6118.

McDonald, K. and Morphew, M. K. (1993). Improved preservation of ultrastructure in difficult-to-fix organisms by high pressure freezing and freeze substitution: 1 *Drosophila melanogaster* and *Strongylocentrotus purpuratus* embryos. *Microsc. Res. Tech*. **24**:254–473.

McDowall, A. W., Chang, J. J., Freeman, R., Lepault, J., Walter, C. A. and Dubochet, J. (1983). Electron microscopy of frozen hydrated sections of vitreous ice and vitrified biological samples. *J. Microsc.* **131**:1–9.

McEwen, B. F. and Frank, J. (2001). Electron tomographic and other approaches for imaging molecular machines. *Curr. Opin. Neurobiol.* **11**:594–600.

McEwen, B. F., Marko, M., Hsieh, C.-E. and Mannella, C. (2002). Use of frozen-hydrated axonemes to assess imaging parameters and resolution limits in cryoelectron tomography. *J. Struct. Biol*. **138**:47–57.

McIntosh, J. R. (2001) Electron microscopy of cells: a new beginning for a new century. *J. Cell Biol*. **153**:F25–F32.

McIntosh, J. R., Nicastro, D. and Mastronarde, D. (2005). New views of cells in 3D: an introduction to electron tomography. *Trends Cell Biol*. **15**:43–51.

Medalia, O., Weber, I., Frangakis, A. S., Nicastro, D., Gerisch, G. and Baumeister, W. (2002). Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* **289**:1209–1213.

Michel, M., Hillman, T. and Müller, M. (1991). Cryosectioning of plant material frozen at high pressure. *J. Microsc*. **163**:3–18.

Michel, M., Gnägi, H. and Müller, M. (1992). Diamonds are a cryosectioner's best friend. *J. Microsc.* **166**:43–56.

Moor, H. (1987). Theory and practice of high pressure freezing. In *Cryotechniques in Biological Electron Microscopy* (R. A. Steinbrecht and K. Zierold, eds). Springer, Berlin, pp. 175–191.

Morphew, M. K. and McIntosh, J. R. (2003). The use of filter membranes for high-pressure freezing of cell monolayers. *J. Microsc.* **212**:21–25.

Nicastro, D., Austin, J., Pierson, J., Gaudette, R., Schwartz, C., Ladinsky, M., Staehelin, L. A. and McIntosh, J. R. (2005b). Visualizing the macromolecular organization of chloroplast membranes using cryo-electron tomography. *Microsc. Microanal*. **11**(**Suppl. 2**):150–151.

Nicastro, D., Frangakis, A. S., Typke, D. and Baumeister, W. (2000). Cryoelectron tomography of *Neurospora* mitochondria. *J. Struct. Biol.* **129**:48–56.

Nicastro, D., McIntosh, J. R. and Baumeister, W. (2005a). 3-D structure of eukaryotic flagella in a quiescent state revealed by cryo-electron tomography. *Proc. Natl Acad. Sci. USA* **102**:15889–15894.

Nickell, S., Hegerl, R., Baumeister, W. and Rachel, R. (2003). *Pyrodictium cannulae* enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *J. Struct. Biol.* **141**(**1**):34–42.

Norlén, L. and Al-Amoudi, A. (2004). Stratum corneum keratin structure, function and formation—the cubic rod-packing and membrane templating model. *J. Invest. Dermatol*. **123**:715–732.

Norlén, L., Al-Amoudi, A. and Dubochet, J. (2003). A cryotransmission electron microscopy study of skin barrier formation. *J. Invest. Dermatol*. **120**:555–560.

Penczek, P., Marko, M., Buttle, K. and Frank, J. (1995). Double-tilt electron tomography. *Ultramicroscopy* **60**:393–410.

Plitzko, J. M., Frangakis, A. S., Foerster, F., Gross, A. and Baumeister, W. (2002). *In vivo veritas*: electron cryotomography of intact cells with molecular resolution. *Trends Biotechnol.* **20**:40–44.

Rath, B. K., Hegerl, R., Leith, A., Shaikh, T. R., Wagenknecht, T. and Frank, J. (2003). Fast 3D motif search of EM density maps using a locally normalized cross-correlation function. *J. Struct. Biol.* **144**:95–103.

Rath, B. K., Marko, M., Radermacher, M. and Frank, J. (1997). Low-dose automated electron tomography: a recent implementation. *J. Struct. Biol*. **120**:210–218.

Richter, K. (1994). Cutting artefacts on ultrathin cryosections of biological bulk specimens. *Micron* **25**:297–308.

Richter, K. (1996). Aspects of cryofixation and cryosectioning for the observation of bulk biological samples in the hydrated state by cryoelectron microscopy. *Scanning Microsc*. (**Suppl. 10**):375–386.

Richter, K., Gnägi, H. and Dubochet, J. (1991). A model for cryosectioning based on the morphology of vitrified ultrathin sections. *J. Microsc.* **163**:19–28.

Rockel B., Jakana J., Chiu W. and Baumeister W. (2002). Electron cryo-microscopy of VAT, the archaeal p97/CDC48 homologue from *Thermoplasma acidophilum*. *J. Mol. Biol*. **317**:673–668.

Sali, A., Glaeser, R., Earnest, T. and Baumeister, W. (2003). From words to literature in structural proteomics. *Nature* **422**:216–225.

Sandin S., Ofverstedt L. G., Wikstrom, A. C., Wrange, O. and Skoglund, U. (2004). Structure and flexibility of individual immunoglobulin G molecules in solution. *Structure* **12**:409–415.

Sartori, N., Bednar, J. and Dubochet, J. (1996). Electron-beam-induced amorphization of ice III or IX obtained by high-pressure freezing. *J. Microsc*. **182**:163–168.

Sartori, N., Richter, K. and Dubochet, J. (1993). Vitrification depth can be increased more than 10-fold by high-pressure freezing. *J. Microsc*. **172**:55–61.

Sartori Blanc, N., Studer, D., Ruhl, K. and Dubochet, J. (1998). Electron beam-induced changes in vitreous sections of biological samples. *J. Microsc*. **192**:194–201.

Sartori-Blanc, N., Senn, A., Leforestier, A., Livolant, F. and Dubochet, J. (2001) DNA in human and stallion spermatozoa forms local hexagonal packing with twist and many defects. *J. Struct. Biol.* **134**:76–81.

Sawaguchi, A., Yao, X., Forte, J. G. and McDonald, K. L. (2003). Direct attachment of cell suspensions to high-pressure freezing specimen planchettes. *J. Microsc*. **212**:13–20.

Schwartz, C., Nicastro, D., Ladinsky, M. S., Mastronarde, D., O'Toole E. and McIntosh, J. R. (2003). Cryo-electron tomography of frozen-hydrated sections of eukaryotic cells. *Microsc. Microanal.* **9**(**Suppl. 2**):1166–1167.

Shi, S., Sun, S. Q., Andrews, B. and Leapman, R. D. (1996). Thickness measurement of hydrated and dehydrated cryosections by EELS. *Microsc. Res. Tech*. **33**:241–250.

Shimoni, E. and Müller, M. (1998). On optimizing high-pressure freezing: from heat transfer theory to a new microbiopsy device. *J. Microsc*. **192**:236–247.

Sitte, H. (1996). Advanced instrumentation and methodology related to cryoultramicrotomy: a review. *Scanning Microsc.* (**Suppl. 10**):387–466.

Somlyo, A. V., Shuman, H. and Somlyo, A. P. (1977). Elemental distribution of striated mucle and the effects of hypertonicity. *J. Cell Biol*. **74**:824–857.

Somlyo, A. P., Bond, M. and Somlyo, A. V. (1985). Calcium content of mitochondria and endoplasmic reticulum in liver frozen rapidly *in vivo*. *Nature* **314**:622–625.

Steinbrecht, R. A. and Zierold, K. (1987). *Cryotechniques in Biological Electron Microscopy.* Springer, Berlin.

Steven A. C. and Aebi, U. (2003). The next ice age: cryo-electron tomography of intact cells. *Trends Cell Biol.* **13**:107–110.

Stoffler, D., Feja, B., Fahrenkrog, J., Walz, J., Typke, D. and Aebi, U. (2003). Cryo-electron tomography provides novel insights into nuclear pore architecture: implications for nucleocytoplasmic transport. *J. Mol. Biol.* **328**:119–130.

Studer, D. and Gnägi, H. (2000). Minimal compression of ultrathin sections with use of an oscillating diamond knife. *J. Microsc*. **197**:94–100.

Studer, D., Graber, W., Al-Amoudi, A. and Eggli, P. (2001). A new approach for cryofixation by high-pressure freezing. *J. Microsc*. **203**:285–294.

Studer, D., Michel, M. and Müller, M. (1989). High pressure freezing comes of age. *Scanning Microsc.* (**Suppl. 3**):253–269.

Studer, D., Michel, M., Wohlwend, M., Hunziker, E. B. and Buschmann, M. D. (1995). Vitrification of articular cartilage by high-pressure freezing. *J. Microsc.* **179**321–332.

Sun, S. Q., Shi, S.-L., Hunt, J. A., Leapman, R. D. (1995). Quantitative water mapping of cryosectioned cells by electron energy-loss spectroscopy. *J. Microsc*. **177**:18–30.

Taylor, K. A. and Glaeser, R. M. (1974). Electron diffraction of frozen, hydrated protein crystals. *Science* **186**:1036–1037.

Taylor, K. A. and Glaeser, R. M. (1976). Electron microscopy of frozen hydrated biological specimens. *J. Ultrastruct. Res.* **55**:448–456.

Ting, C. S., Hsieh, C., Sundararaman, S., Mannella, C. A. and Marko, M. (2005). Comparative three-dimensional imaging of environmentally critical cyanobacteria through cryo-electron tomography. *Microsc. Microanal*. **11**(**Suppl. 2**): 332CD.

Tokuyasu, K. T. (1986). Application of cryoultramicrotomy to immunocytochemistry. *J. Microsc*. **143**:139–149.

Vanhecke, D., Graber, W., Herrmann, G., Al-Amoudi, A., Eggli, P. and Studer. D. (2003). A rapid microbiopsy system to improve the preservation of biological samples prior to high-pressure freezing. *J. Microsc*. **212**:3–12.

Van Marle, J., Dietrich, A., Jonges, K., Jonges, R, de Moor, E., Vink, A., Boon, P. and van Veen, H. (1995). EM-tomography of section collapse, a non-linear phenomenon. *Microsc. Res. Tech*. **31**:311–316.

Vonk, J. (2000). Parameters affecting specimen flatness of two-dimensional crystals for electron crystallography. *Ultramicroscopy* **5**:123–129.

Wagenknecht, T., Hsieh, C.-E., Rath, B., Fleischer, S. and Marko, M. (2002). Electron tomography of frozen-hydrated isolated triad junctions. *Biophys. J.* **83**:2491–2501.

Walther, P. and Müller, M. (1997). Double-layer coating for field-emission cryo-scanning electron microscopy—present state and applications. *Scanning* **19**:343–348.

Walz, J., Typke, D., Nitsch, M., Koster, A. J., Hegerl, R. and Baumeister, W. (1997). Electron tomography of single ice-embedded macromolecules: three-dimensional alignment and classification. *J. Struct. Biol*. **120**:387–395.

Winkler, H. and Taylor, K. A. (2006). Accurate marker-free alignment with simultaneous geometry determination and reconstruction of tilt series in electron tomography. *Ultramicroscopy* **106**:240–254.

Woodcock, C. L. (1994). Chromatin fibers observed *in situ* in frozen hydrated sections. Native fiber diameter is not correlated with nucleosome repeat length. *J. Cell Biol*. **125**:11–19.

Zhang, P., Bos, E., Heyman, J., Gnägi, H., Kessel, M., Petere, P. J. and Subramaniam, S. (2004). Direct visualization of receptor arrays in frozen-hydrated sections and plunge-frozen specimens of *E. coli* engineered to overproduce the chemotaxis receptor Tsr. *J. Microsc*. **216**:76–83.

Zhao, Q., Ofverstedt, L. G., Skoglund, U. and Isaksson, L. A. (2004). Morphological variation of individual *Escherichia coli* 30S ribosomal subunits *in vitro* and *in situ*, as revealed by cryo-electron tomography. *Exp. Cell Res.* **297**:495–507.

Zierold, K. (1984). The morphology of ultrathin cryosections. *Ultramicroscopy* **14**:201–210.

Zierold, K. (1987). Cryoultramicrotomy. In *Cryotechniques in Biological Electron Microscopy* (R. A. Steinbrecht and K. Zierold, eds). Springer, Berlin, pp. 132–148.

Zierold, K. (1988). X-ray microanalysis of freeze-dried and frozen-hydrated cryosections. *J. Electron Microsc. Tech*. **9**:65–82.

3

The Electron Microscope as a Structure Projector

Peter W. Hawkes

1. INTRODUCTION

The intuitive understanding of the process of 3D reconstruction is based on a number of assumptions, which are easily made unconsciously; the most crucial is the belief that what is detected is some kind of projection through the structure. This 'projection' need not necessarily be a (weighted) sum or integral through the structure of some physical property of the latter; in

Peter W. Hawkes • CEMES-CNRS BP 94347, F-31055 Toulouse cedex France

principle, a monotonically varying function would be acceptable, although solving the corresponding inverse problem might not be easy. In practice, however, the usual interpretation of 'projection' is overwhelmingly adopted, and it was for this definition that Radon (1917) first proposed a solution. In the case of light shone through a translucent structure of varying opacity, a 3D transparency as it were, the validity of this projection assumption seems too obvious to need discussion. We know enough about the behavior of X-rays in matter to establish the conditions in which it is valid in radiography. In this chapter, we enquire whether it is valid in electron microscopy, where intuition might well lead us to suspect that it is not. Electron–specimen interactions are very different from those encountered in X-ray tomography; the specimens are themselves very different in nature, creating phase rather than amplitude contrast, and an optical system is needed to transform the information about the specimen that the electrons have acquired into a visible image. If the electrons encounter more than one structural feature in their passage through the specimen, the overall effect is far from easy to guess, whereas in the case of light shone through a transparent structure, it is precisely the variety of such overlaps or superpositions that we use to effect the reconstruction. If intuition were our only guide, we might easily doubt whether 3D reconstruction from electron micrographs is possible: there is no useful projection approximation for the balls on a pin-ball machine! Why then has it been so successful? To understand this, we must examine in detail the nature of the interactions between the electrons and the specimen, and the characteristics of the image-forming process in the electron microscope. Does the information about the specimen imprinted on the electron beam as it emerges from the latter represent a projection through the structure? How faithfully is this information conveyed to the recorded image? These are the questions that we shall be exploring in the following sections.

2. ELECTRON–SPECIMEN INTERACTIONS

2.1. Generalities

The electron microscope specimen is an ordered array of atoms—highly ordered if the object is essentially crystalline or has been organized during the preparation process into some kind of array, and organized, or ordered more locally, in the case of isolated particles. It may be stained, in which case the light atoms of which organic matter is mainly composed will be selectively bound to much heavier atoms, but if very high resolution is desired it will probably be unstained. Its thickness will depend on its nature and on the resolution to be achieved, for reasons that will gradually became clear. The electric potential within the specimen is not uniform, since in the simplest picture, each atomic nucleus creates a Coulomb potential, screened to an extent that depends on the nature of the atom in question by its electrons. In reality, the situation is complicated by the bonding between the atoms,

but the potential distribution will be dominated by the pattern of screened Coulomb potentials associated with the atoms of specimen or stain.

The first step in electron image formation in a conventional (fixed-beam) transmission electron microscope (TEM) consists of irradiating the specimen with a beam of electrons, which may suffer various fates. Many will pass through the specimen, unaffected by its presence in their path. These are said to be unscattered. In the atomic picture, they do not pass close enough to any of the specimen atoms to be deviated significantly from their course; recalling that, from the electrons' point of view, the interior of the specimen is a potential distribution, the unscattered electrons pass through a zone in which the transverse electric field is weak everywhere. Since the mean potential inside the specimen is different from that of the surroundings, the energy of the beam electrons will be altered slightly as they enter the specimen, reverting to its original value as they emerge.

Electrons that pass close to a specimen atom experience a strong transverse electric field and hence will be deflected laterally, or *scattered*; this scattering may or may not be accompanied by a transfer of energy from the beam electron to the specimen. If the energy transferred is negligible, the electron is said to be scattered *elastically*; if, however, there is an appreciable recoil on the part of the specimen atoms, the scattering is *inelastic.* The likelihood of these various events is measured by their scattering cross-sections, which tell us not only the probability of elastic or inelastic scattering as a function of incident beam energy and of the atomic number of the scattering atom but also the angular scattering probability. The mean scattering angle, for example, can thus be determined.

Scattered electrons may then emerge from the specimen or may be scattered afresh, elastically or inelastically; the likelihood of a second scattering event increases with specimen thickness, as we should expect.

This qualitative picture of the electron–specimen interaction goes some way to dispel any doubts that the transmitted beam carries enough information about the specimen for reconstruction to be possible. It also shows us that the simple picture of intensity attenuation must be replaced by the very different idea of a scattering pattern directly related to the electric potential distribution within the specimen, itself determined by the atomic arrangement and dominated by the fields created by heavy atoms, if any are present. In the case of unstained specimens, therefore, we can hope to recover information from all the specimen atoms, but if these are light the signal will be weak or, in other words, only a small fraction of the incident beam electrons will be carrying useful information when they leave the specimen. In the case of stained specimens, whether the stain is negative or positive, the recorded image will be dominated by contrast from the stain.

In order to translate the foregoing discussion into more quantitative language, we now analyze the electron–specimen interaction, especially the elastic case, with the aid of quantum mechanics. We only reproduce here the main steps of the argument, a full account of which may be found in texts on electron microscopy (Reimer, 1997; Williams and Carter, 1996).

It is sufficient for our purposes to represent the incident beam by an extended plane wave

$$\psi = \psi_0 \exp(2\pi i k z) \tag{1}$$

traveling along the microscope axis, which coincides with the z coordinate axis. The effect of imperfect beam coherence (a spread in the incident beam direction and energy) will be considered in Section 3.2. Elastic scattering by a specimen atom generates a spherical wave, so the wave function becomes

$$\psi = \psi_0 \exp(2\pi i k z) + \psi_{sc} \tag{2}$$

where

$$\psi_{sc} = i\psi_0 f(\theta) \frac{\exp(2\pi i k r)}{r} \tag{3}$$

The function $f(\theta)$ is the complex scattering amplitude, $f(\theta) = |f(\theta)| \exp i\eta(\theta)$, and k is the wave number[1]:

$$k = \frac{1}{\lambda} \tag{4}$$

We can use this basic formalism in different ways, and it is important to distinguish between these, since they provide an understanding of the two important contrast mechanisms in the electron microscope. The contrast is always phase contrast, in the sense that virtually all the electrons incident on the specimen emerge beyond it and are, hence, available to participate in image formation. Two very different mechanisms operate to convert the phase information into a visible image and, hence, into intensity variations in some plane downstream from the specimen. We return to these in Section 3, but they inevitably influence the present discussion.

2.2. Amplitude Contrast

We first use equation (3) to calculate the number of electrons scattered through an angle greater than some threshold value α, treating the scattering medium as a film of given mass thickness μ (product of density ρ and thickness t). It is easy to show that the decrease in the number of electrons in the beam scattered through angles smaller than α is given by

$$\frac{dn}{n} = -N\sigma(\alpha)d\mu \tag{5}$$

[1] Note that two conventions are equally common in the literature: $k = 1/\lambda$ and $k = 2\pi/\lambda$.

where N = (Avogadro's number)/(atomic weight) and $\sigma(\alpha)$ is the scattering cross-section:

$$\sigma(\alpha) = \int_{\alpha}^{\pi} |f(\theta)|^2 2\pi \sin\theta \mathrm{d}\theta \tag{6}$$

From (5) we see that

$$\begin{aligned} n &= n_0 \exp\{-N\sigma(\alpha)\mu\} \\ &= n_0 \exp\left(-\frac{\mu}{\mu_t}\right), \quad \mu_t = \frac{1}{N\sigma(\alpha)} \end{aligned} \tag{7}$$

If, therefore, the angle α corresponds to the angle of acceptance of the objective aperture, n/n_0 tells us what fraction of the incident beam will pass through the aperture and reach the image. This is a most important result, for two reasons. First, it tells us that the contrast variations at the final image are directly related to the projection of the density through the specimen along the path of the electron. Secondly, it reassures us that this relationship is single-valued since the exponential function varies monotonically.

The fraction intercepted varies with the nature and position of the scattering atoms and hence generates *scattering contrast*. The phase shifts are closely related to the angular deflections of the electrons in the specimen and depend on the nature of the scattering atoms encountered. They are converted into amplitude contrast by the objective aperture, which intercepts electrons that have been scattered through larger angles and have passed close to the heavier atoms in the specimen. Virtually all the medium-resolution contrast (beyond 2–3 nm, say) seen in electron micrographs is generated by this mechanism, and this therefore includes contrast from negatively stained specimens, for which the resolution is not better than this figure.

Equation (7) is valid only if multiple scattering is rare, though an expression of the same form is obtained if elastic and inelastic single scattering are allowed (see Section 6.1 of Reimer, 1997); the total cross-section is the sum of the elastic and inelastic cross-sections. An estimate of the maximum acceptable specimen thickness can therefore be based on the mean free path between scattering events, which is denoted by μ_t in equation (7). (Published values of this path length frequently represent the product of density and path length, just as the mass thickness is the product of density and thickness.) Figure 1 shows how the total elastic and inelastic cross-sections vary in the range from 20 to 500 kV for elements between carbon (Z = 6, atomic weight = 12, density = 2.2 g/cm^3) and platinum (Z = 78, atomic weight = 195, density = 21.5 g/cm^3). Table 1 shows the true elastic mean free path for a light element (carbon) and a heavy element (platinum) as a function of accelerating voltage. For 100 kV, for example, the elastic mean free path is ~200 nm for pure carbon (graphite) but falls to 40 nm for germanium (Z = 32) and below 10 nm for platinum (Z = 78), which is comparable with the common staining elements osmium (Z = 76) and uranium

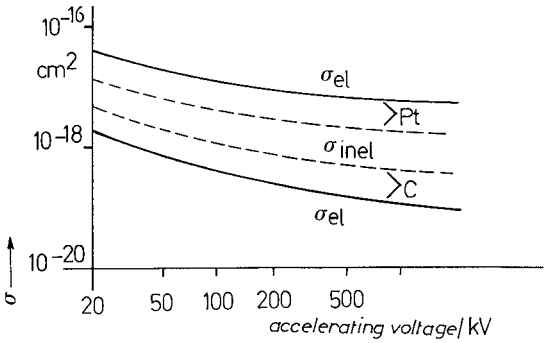


FIGURE 1. The elastic and inelastic cross-sections for carbon and platinum for accelerating voltages in the range 20–500 kV (after Reimer, 1997).

TABLE 1 Elastic mean free paths for carbon and platinum (in nanometers) as a function of accelerating voltage[a]

Accelerating voltage (kV)	Carbon	Platinum
17.3	45.9	0.03
25.2	65.5	3.78
41.5	102	5.41
62.1	145	6.57
81.8	181	7.83
102.2	216	8.95
150	321	10.9
300	518	14.7
750	632	23.6

[a] After Reimer (1997).

(Z = 92). For further data, see Reimer and Sommer (1968), Arnal *et al.* (1977), Reichelt and Engel (1984) and, for guidance through the extensive literature, Chapters 5 and 6 of Reimer (1997).

The simple expression for n can be improved by including inelastic scattering, but is unreliable if the specimen thickness is increased to the point beyond which multiple scattering is common. Nevertheless, in normal circumstances, the current density distribution emerging from the objective aperture does reflect faithfully the pattern of scattering atoms in the specimen, especially the heavy atoms which will probably be those of any stain. Even though the model of the specimen that enabled us to reach this conclusion was crude, we can conclude that, so far as this contrast mechanism is considered, the images used for 3D reconstruction are essentially projections through the specimen. All the early, successful attempts at 3D reconstruction relied on contrast of this kind (DeRosier and Klug, 1968; DeRosier, 1971), and the reasoning employed to justify the reconstruction procedure was considerably less laborious than that given here. The speci-

men is there represented by a mass density distribution, $\rho_t(x, y, z)$, the Fourier transform of which is

$$F(X,Y,Z) = \iiint \rho_t(x,y,z)\exp\{-2\pi i(Xx+Yy+Zz)\}dx\,dy\,dz \tag{8}$$

A central section through this 3D function yields a 2D distribution, and, in particular, for $Z = 0$, we have

$$\begin{aligned} F(X,Y,0) &= \iint \left\{\int \rho_t(x,y,z)dz\right\}\exp\{-2\pi i(Xx+Yy)\}dx\,dy \\ &= \iint \mu(x,y)\exp\{-2\pi i(Xx+Yy)\}dx\,dy \end{aligned} \tag{9}$$

The quantity μ is the projection in the z direction of the mass density distribution, and, as we have seen, the intensity distribution in the image is given by $n/n_0 = \exp(-\mu/\mu_t)$. The foregoing analysis is, however, necessary to understand the assumptions underlying this reasoning.

2.3. Phase Contrast

At higher resolution, a very different contrast mechanism comes into play, and we must re-examine the expression for the wave function emerging from the specimen. The latter is calculated by dividing the specimen into a large number of very thin slices, each of which is represented by a 2D, multiplicative, complex 'transparency'. The incident wave is thus modified by the first slice, allowed to propagate to the following slice, is again modified, propagates further, and so on, until the far side of the specimen is reached. This technique has been studied in very great detail, in connection with crystal structure determination in particular, and is widely used for image simulation. The aspect of this theory of interest here is the *phase-grating approximation*. In certain circumstances, the outgoing wave from the specimen can be represented by a multiplicative specimen transparency function:

$$S(x,y) = |S(x,y)|\exp\{i\varphi(x,y)\} \tag{10}$$

in which the phase shift φ and the amplitude term $|S|$ are projections of the potential and absorption of the specimen. In this approximation, multiple scattering is permitted. A careful examination of the derivation of this result shows that it is essentially a high-energy approximation, strictly valid for vanishing wavelength, when the Ewald sphere[2] becomes indistinguishable from a plane. The phase shift φ is given by

[2] The Ewald sphere is a geometrical construct widely used in the study of crystalline specimens. The diffraction pattern of such specimens can be predicted by identifying the points of intersection of this sphere with the nodes of the reciprocal lattice, which is determined directly by the crystal lattice of the specimen. The radius of the Ewald sphere is $1/\lambda$, so that for short wavelengths its curvature can often but not always be neglected, see Wolf *et al*. (2006).

$$\varphi(x, y) = -\frac{\pi\gamma}{\lambda e\hat{U}}\int V(x, y, z)dz$$
$$\approx -\frac{\pi}{\lambda eU}\int V(x, y, z)dz \qquad (11)$$

in which λ is the electron wavelength:

$$\lambda = \frac{h}{\left(2m_0e\hat{U}\right)^{1/2}} \approx \frac{1.2}{\hat{U}^{1/2}}\,\text{nm} \qquad (12)$$

e is the absolute value of the charge on the electron, m_0 is its rest mass, U is the accelerating voltage, and $\hat{U}$ is its relativistically corrected value:

$$\hat{U} = U\left(1+\frac{eU}{2m_0c^2}\right) \approx U(1+10^{-6}U); \quad \gamma = 1+\frac{eU}{m_0c^2} \approx 1+2\times10^{-6}U$$

(U in volts). The integral is taken through the specimen, in which the potential distribution is $V(x, y, z)$.

Although it is not necessary to adopt the approximations that lead to this phase-grating approximation when calculating the image of a given structure, it is difficult to see how we could proceed in the opposite direction without it. If the specimen cannot be represented by a multiplicative specimen transparency function, or projection, the interpretation of any attempted 3D reconstruction will be obscure. Indeed, attempts to perform such reconstructions would be likely to fail or, at least, prove to be incapable of furnishing high-resolution detail owing to the inconsistency of the data in the different views through the specimen when these views are not strictly projections.

In conclusion, therefore, we may say that the information conveyed by the emergent electron wave is essentially projection information provided that the conditions of the phase-grating approximation are satisfied. The most important of these is that the wave function does not 'spread' laterally as it propagates through the specimen ('column approximation'). This is stated with great clarity by Spence (2003):

> This important qualitative result [that the broadening is typically less than 0.4 nm for thicknesses below 50 nm] indicates that the axial, dynamical image wavefunction on the exit-face of a crystal is locally determined by the crystal potential within a small cylinder whose axis forms the beam direction and whose diameter is always less than a few angströms for typical HREM conditions. This is essentially a consequence of the forward scattering nature of high-energy electron diffraction.

Thus, the thinner the specimen, the higher the accelerating voltage (with little further gain beyond ~600 kV), and the lighter the atoms involved, the better will be the phase-grating and hence the projection approximation.

This qualitative observation is useful, but how can it be rendered quantitative? A valuable rule-of thumb is given by Spence (2003), whose chapter on non-periodic specimens is recommended as background reading. He notes that the effects of Fresnel diffraction in the specimen are neglected in the phase-grating approximation, which is tantamount to assuming that the curvature of the Ewald sphere is negligible. The properties of Fresnel wave propagation can therefore be used to estimate the maximum tolerable thickness (t) for which the phase-grating approximation is acceptable. Spence finds

$$t \leqslant \frac{d^2}{\lambda}$$

where d is effectively the desired resolution. Figure 2 shows t as a function of λ (or accelerating voltage) for $d = 0.5, 1$ and 2 nm.

It is easier to establish the reliability of the phase-grating approximation in the case of crystalline materials. Indeed, any attempt to simulate the image of a structure by the multislice method effectively imposes a periodic character on the latter for, if the structure has no natural periodicity, then for the purposes of the calculation it is regarded as one member of an

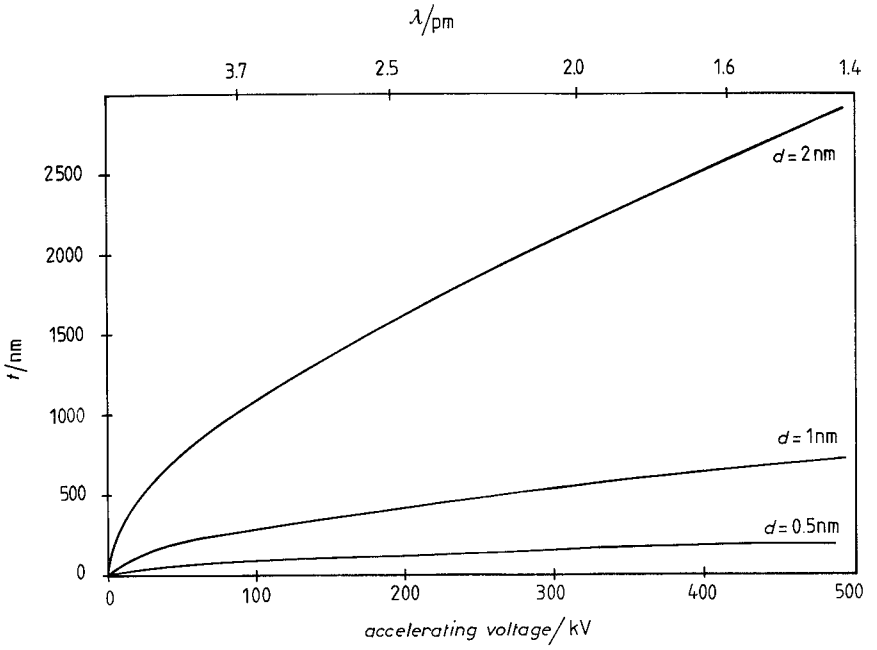


FIGURE 2. The maximum thickness t as a function of accelerating voltage (U) or wavelength (λ) for a resolution of 0.5, 1 and 2 nm.

array of replicas of itself. The amount of calculation becomes proportionately larger, which explains why accurate image simulations for complex, non-periodic specimens remain rare. Among the earlier calculations are those by Grinton and Cowley (1971), for a negatively stained rod-shaped structure (schematically representing a virus), who find that the approximation beaks down at ~10 nm for a resolution of 0.3 nm at ~60 kV ($\lambda = 5$ pm); this is more pessimistic than Spence's estimate (18 nm).

Although it is dangerous to extrapolate their results beyond the rather limited conditions in which the latter were obtained, we next draw attention to the careful study by Jap and Glaeser (1980) of the domains of validity of two single-scattering approximations (the kinematic or first Born approximation and the weak phase object approximation) for two organic crystalline materials: anhydrous cytosine and DISOPS (disodium-4-oxypyrimidine-2-sulfinatehexahydrate). Since these are crystalline, Jap and Glaeser consider both individual diffraction spots and the projected potential. Their findings cannot be encapsulated in a single conclusion, but their curves for the reliability of the projected potential given by the two single-scattering approximations considered do suggest some general guidelines. For resolutions in the 3–5 Å range, the single-scattering approximations are 'quite reliable' for thicknesses of 100–150 Å at 100 kV and of 200–250 Å at 500 kV. (The dissimilarity factor, in terms of which reliability is defined by Jap and Glaeser, is a well-defined quantity for crystalline specimens; the ranges given above correspond to different values of this factor.) We must, however, insist that these results, though indicative, must be treated with caution, even though they agree reasonably with the estimates given earlier. Jap and Glaeser concluded:

> It seems likely that the single-scattering approximations will tend to have a greater validity for large, complex structures than for small, simple structures because it is more likely in the case of small structures that atom centers can overlap in projection. This effect gives rise to large values of the Fourier coefficients of the crystal potential, and a non-linear dependence of the transmitted electron wave function upon the crystal potential. On the other hand, organic structures with much larger unit-cell dimensions will produce a situation in which very many more beams are simultaneously excited than is the case for cytosine or DISOPS. Thus the domain of validity of the single-scattering approximations for larger structures again cannot be accurately predicted from the present results.

Our present concerns being with specimens possessing little or no periodicity, these warnings are particularly apposite. Furthermore, the tilt angles of the outer members of a tilt series are large (±65°), with the result that the thickness traversed by the electrons is much greater than for the untilted specimen. Some subsequent calculations by Glaeser and Downing (1993) are, however, reassuring. On the basis of earlier work (Ho *et al.*, 1988; Jap and Glaeser, 1980), they used dynamical theory calculations to establish 'the domain of specimen thickness, resolution and accelerating voltage within which the usual kinematic theory remained valid to a specified

degree of accuracy'. They conclude 'that the kinematic theory can be used for unstained biological macromolecules up to a specimen thickness of at least 10 to 20 nm for voltages of 100 kV or more'. Glaeser and Downing also describe calculations designed to reveal the impact of dynamical effects on bacteriorhodopsin images. For the smallest thickness considered (4 nm, about the size of the bacteriorhodopsin molecule), such effects are negligible. They increase with thickness but 'none of the indicators of dynamical effects becomes large enough to be considered a significant effect until the specimen thickness is at least 12 nm. These calculations support earlier conclusions . . . that dynamical effects can be ignored, and kinematic diffraction theory can be used to solve crystal structures of biological molecules up to a specimen thickness of 10 to 20 nm, using data obtained with 100 kV electrons'. Certainly these comments are concerned with electron crystallography rather than tomography, but they are indicative of the reliability of the projection approximation. See also Glaeser and Ceska (1989), Chapter 6 of Spence (2003), 'HREM in biology, organic crystals, and radiation damage', and especially Section 6.9 on 'Molecular imaging in three dimensions—electron tomography', and Glaeser *et al.* (2007).

2.4. Amplitude Contrast Re-examined

It is convenient, particularly on a first encounter, to discuss the scattering contrast generated by the objective aperture, which intercepts some of the scattered electrons, and the higher resolution phase contrast as though they were separate and disjoint mechanisms. This is, however, strictly incorrect and, more seriously, leaves a twilight zone somewhere between the high- and medium-resolution regimes unaccounted for. In fact, we could abandon the notion of scattering contrast altogether and include the role of the aperture in the discussion of the mechanisms that convert phase variations in the wave function into amplitude variations in the image. This is indeed done in image simulation. Nevertheless, a great many electron images can be explained satisfactorily by the arguments of Section 2.2, and there is no reason to adopt a complicated train of thought when a simpler one is not misleading. The fact that the two approaches give identical results for single atoms can be demonstrated explicitly (see Section 6.3.3 of Reimer, 1997, for a proof of this).

2.5. Radiation Damage

The problem of the damage that the electron beam may cause as it passes through the specimen goes rather beyond the subject of this chapter and is dealt with more fully by Luther in Chapter 1. Indeed, one of the conditions to be satisfied if the electron microscope is to furnish true projections through the structure is that the passage of the electrons leaves the latter unscathed. Fortunately, many of the requirements for the preserva-

tion of high-resolution information that we have already encountered are the same as those that protect the specimen from radiation damage, since the aim is to keep inelastic scattering as low as possible. The accelerating voltage should not be too low, and the specimen should be thin and consist of light atoms. For stained specimens, the remedy is less obvious, as heavy atoms are numerous and the beam may cause loss of mass and stain migration. Some idea of the magnitude of the effects may be gained by noting that both the elastic and inelastic cross-sections are roughly halved by an increase in accelerating voltage from 100 to 300 kV. The reduction in inelastic scattering in a given specimen should be accompanied by diminution of radiation damage, though other factors are also involved. We refer to Chapter 1 of this volume and to Chapter 10 of Reimer (1997), where the physics of the damage process is thoroughly explored. The earlier papers of Zeitler (1982, 1984) are also of interest in this context. The situation has, however, changed dramatically with the widespread use of cryotechniques in electron tomography (see Chapter 2 by Marko *et al*. and a recent survey by Plitzko and Baumeister, 2006).

3. IMAGE FORMATION IN THE TRANSMISSION ELECTRON MICROSCOPE

3.1. The Coherent Limit

The illumination in an electron microscope is partially coherent. The effective source is small but not pointlike, and the energy spread is narrow but still appreciable. It is, however, convenient to analyze image formation on the assumption that the source is perfectly coherent, i.e. has vanishingly small dimensions and negligible energy spread, and introduce the effects of partial coherence by a perturbation procedure. As before, we represent the wave emerging from the specimen in the form

$$\psi = S(x_0, y_0)\psi_s \tag{13}$$

and we allow this wave to propagate through the objective lens field to the objective aperture and thence to the image plane of the microscope. Each of these steps is represented by a Fourier transform, an immediate consequence of the form of the Schrödinger equation (for details, see Glaser, 1952; Hawkes, 1980a; Hawkes and Kasper, 1994). The effect of the objective lens aberrations may be represented by a multiplicative function in the objective aperture plane (strictly, the plane conjugate to the source, but by considering a plane incident wave, the source is automatically conjugate to the back focal plane of the objective, close to which the aperture is in principle located); see Born and Wolf (1999) for a proof of this. With no new approximations, therefore, and the sole assumption that the illumination is coherent, we may write down the wave function (ψ) and hence the current

density ($\propto \psi\psi^*$) in the image plane of the microscope. In practice, however, the spherical aberration of the objective lens is the only geometrical aberration that need concern us, and this renders the expression for the image intensity substantially more simple.

A straightforward calculation, set out in detail in most modern texts on electron image formation (Hawkes, 1980a; Reimer, 1997; Spence, 2003; Hawkes and Kasper, 1994), tells us that the wave function in the image plane of the microscope, $\psi(x_i, y_i)$ is related to that emerging from the object, $\psi(x_0, y_0)$, by a relationship of the form

$$E_i\psi(x_i, y_i, z_i) = \frac{1}{M}\iint K\left(\frac{x_i}{M} - x_0, \frac{y_i}{M} - y_0\right)\psi(x_0 y_0 z_0)E_0\, dx_0\, dy_0 \qquad (14)$$

in which E_i and E_0 are quadratic phase factors and K is a function characterizing the transfer of information along the microscope:

$$K(x_i, y_i; x_0, y_0) = \frac{1}{\lambda^2 f^2}\iint a(x_a, y_a)\exp[-i\gamma(x_a, y_a)]$$
$$\times \exp\left[-\frac{2\pi i}{\lambda f}\left\{\left(x_0 - \frac{x_i}{M}\right)x_a + \left(y_0 - \frac{y_i}{M}\right)y_a\right\}\right]dx_a\, dy_a \qquad (15)$$

and x_a, y_a are position coordinates in the aperture plane.

The function $a(x_a, y_a)$ is equal to unity in the objective aperture and zero outside. The phase shift $\gamma(x_a, y_a)$ is determined by the spherical aberration coefficient of the objective lens C_s and the defocus Δ measured in the object plane (i.e. the distance from the specimen to the plane optically conjugate to the image plane). An additional term can be included in γ to describe any residual astigmatism, but we shall neglect this since it is not directly relevant to the theme of this chapter. Thus,

$$\gamma(x_a, y_a) = \frac{2\pi}{\lambda}\left\{\frac{1}{4}C_s\frac{\left(x_a^2 + y_a^2\right)^2}{f^4} - \frac{1}{2}\Delta\frac{x_a^2 + y_a^2}{f^2}\right\} \qquad (16)$$

Since (14) has the form of a convolution, its Fourier transform will be the direct product of the transforms of the member functions. Writing

$$\tilde{\psi}_0(p_x, p_y) = \iint E_0\psi(x_0, y_0)\exp\{-2\pi i(p_x x_0 + p_y y_0)\}dx_0\, dy_0$$
$$\tilde{\psi}_i(p_x, p_y) = \iint E_i\psi(Mx_i, My_i)\exp\{-2\pi i(p_x x_i + p_y y_i)\}dx_i\, dy_i$$
$$\tilde{K}(p_x, p_y) = \iint K(x, y)\exp\{-2\pi i(p_x x + p_y y)\}dx\, dy \qquad (17)$$

equation (14) becomes

$$\tilde{\psi}_i = \frac{1}{M}\tilde{K}\tilde{\psi}_0 \qquad (18)$$

This extremely important exact equation tells us that, provided we can neglect geometrical aberrations other than spherical aberration, the electron microscope behaves as a linear scalar filter acting on the complex wave function ψ. If we could record the image wave function, the object wave function could be deduced from it in a straightforward way. Unfortunately, only the amplitude of the wave function, and not its phase can generally be recorded unless we are prepared to abandon the traditional imaging modes and invoke holography or accept the relatively difficult computing needed to solve the 'phase problem'. This difficulty is readily understood in simple physical terms. We have seen in the discussion of scattering that deflection of an electron by the electric field close to an atom is equivalent to introducing a phase variation in the wave function. The phase of the image wave is likewise intimately associated with the direction in which the electrons are traveling at the image plane. We can record the points of arrival of electrons, but we do not know where they were coming from. This is the essence of the phase problem, and shows why it can be solved, in theory at least, by taking two micrographs at different defocus values (or, more generally, in different imaging conditions). Great progress has been made in recent years; see Kirkland and Hutchison (2006) for a survey.

In order to proceed further, we replace the wave function ψ_0 at the exit surface of the specimen by the product of the specimen transparency and the (plane) wave incident on the object:

$$\psi_0 = S(x_0, y_0)\psi_s \tag{19}$$

Since the modulus of S is frequently close to unity, it is convenient to write

$$S = (1 - s)\exp(i\varphi) \tag{20}$$

In the case of a weakly scattering specimen, both s and φ are small; we recall that the presence of s allows for inelastic scattering. The exponential may then be expanded, giving

$$S \approx (1 - s)(1 + i\varphi - \cdots) \approx 1 - s + i\varphi \tag{21}$$

if quadratic and higher order terms are neglected. Substituting the approximation (21) into (18), we find

$$\tilde{\psi}_i = \frac{1}{M}\tilde{K}\{\delta(p_x, p_y) - \tilde{s} + i\tilde{\varphi}\} \tag{22}$$

in which δ is the Dirac delta function and $\tilde{s}$, $\tilde{\varphi}$ are the Fourier transforms of s and φ. From the definitions of K and $\tilde{K}$, equations (15) and (17), we know that

$$\tilde{K}(p_x, p_y) = A(\lambda f p_x, \lambda f p_y) \tag{23}$$

where $A = a\exp(-i\gamma)$. The current density in the image j_i is proportional to $\psi_i\psi_i^*$. After some straightforward calculation, again neglecting quadratic terms in s and φ, we find that

$$\begin{aligned} j_i(Mx_i, My_i) &\equiv M^2|\psi_i(Mx_i, My_i)|^2 \\ &= 1 - 2\iint a\tilde{s}\cos\gamma\exp\{2\pi i(p_x x_i + p_y y_i)\}dp_x\,dp_y \\ &\quad + 2\iint a\tilde{\varphi}\sin\gamma\exp\{2\pi i(p_x x_i + p_y y_i)\}dp_x\,dp_y \end{aligned} \tag{24}$$

Writing $j_i = 1 + C$, we see immediately that

$$\tilde{C} = -2a\tilde{s}\cos\gamma + 2a\tilde{\varphi}\sin\gamma \tag{25}$$

where $\tilde{C}$ is the Fourier transform or spatial frequency spectrum of C, itself a measure of the image contrast.

This result tells us how the phase and amplitude of a weakly scattering specimen are represented in the image. Each is modulated by a sinusoidally varying function, the behavior of which is governed by the spherical aberration coefficient of the objective lens, C_s, and the choice of the defocus value. Since the microscopist has no control over C_s, the value of the defocus is chosen in such a way that as wide a range of values as possible of the spatial frequencies of the specimen phase reach the image undistorted by the function $\sin\gamma$. This in turn implies that a range of sizes of specimen detail will be present in the image and can be interpreted directly.

However, equation (25) is a linear equation. Thus, even if all the information needed for subsequent 3D reconstruction is not present in a single micrograph, it is, in principle, possible to form a linear superposition of several micrographs, taken at different defocus values and suitably weighted, to create an image less disturbed by the transfer function $\sin\gamma$. We shall not give details of this technique (see Saxton, 1986; Schiske, 1968, 1973; Hawkes and Kasper, 1994), but it may be important to ensure that information about the specimen that is carried by the electron beam, but is destroyed or distorted by the passage through the microscope, does appear correctly in the image used for 3D reconstruction.

The form of the phase contrast transfer function, $\propto \sin\gamma$, changes rapidly as the defocus is varied. Before discussing this behavior, we scale, for convenience, the defocus Δ and the coordinates (x_a, y_a) in the aperture plane in such a way as to suppress the dependence of p_x, p_y, and Δ on λ and C_s. We write

$$\begin{gathered} \bar{\Delta} = \frac{\Delta}{(C_s\lambda)^{1/2}}, \quad \bar{f} = \frac{f}{(C_s\lambda)^{1/2}}, \\ \bar{p}_x = (C_s\lambda^3)^{1/4}p_x = \frac{\bar{x}_a}{\bar{f}}, \quad \bar{p}_y = (C_s\lambda^3)^{1/4}p_y = \frac{\bar{y}_a}{\bar{f}} \end{gathered} \tag{26}$$

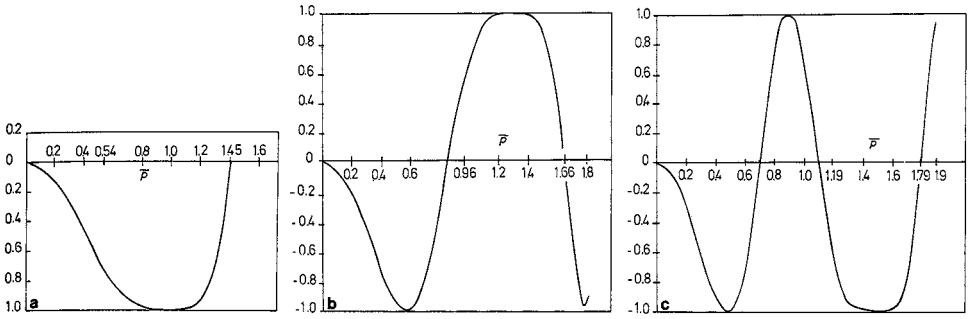


FIGURE 3. The function sin γ as a function of reduced spatial frequency $\bar{p}$ for reduced defocus $\bar{\Delta} = 1$ (a), $\sqrt{3}$ (b) and $\sqrt{5}$ (c).

where $\bar{x}_a = x_a/(C_s\lambda^3)^{1/4}$ and likewise for $\bar{y}_a$. It has been suggested (Hawkes, 1980b) that the scaling factors be given names. A defocus of $(C_s\lambda)^{1/2}$ is thus said to be 1 Sch (one scherzer), and a length of $(C_s\lambda^3)^{1/4}$ is said to be 1 Gl (one glaser). With this new scaling, we have

$$\gamma = \frac{\pi}{2}\bar{p}^4 - \pi\bar{\Delta}\bar{p}^2, \quad \bar{p}^2 = \bar{p}_x^2 + \bar{p}_y^2 \tag{27}$$

Figure 3 shows sin γ as a function of $\bar{p}$ for various values of $\bar{\Delta}$ in the vicinity of $\bar{\Delta} = 1$, $\sqrt{3}$, and $\sqrt{5}$, where the function has wide fairly constant regions and varies least rapidly with $\bar{\Delta}$. These are therefore the values of defocus to be preferred for high-resolution imaging. In practice, the so-called Scherzer focus, corresponding to $\bar{\Delta} = 1$, $\Delta = (C_s\lambda)^{1/2}$, is the most convenient for direct interpretation of the recorded image.

3.2. Real Imaging Conditions: Partial Coherence

In practice, almost all of the assumptions that we have been making are unjustified. The illuminating beam is not perfectly coherent, and the specimen rarely scatters so weakly that φ^2 can be truly neglected. In this section, we briefly consider the relatively innocuous effects of partial coherence. We return to the problems associated with the weak scattering approximation in Section 4.

We have assumed that the illuminating beam has no energy spread (perfect temporal coherence) and is emitted from a vanishingly small source (perfect spatial coherence). In practice, the energy spread may reach a few electron volts, the exact value depending on the nature of the emitter and the gun optics. The finite source size has the effect of spreading the directions of arrival of the electrons at the specimen. Instead of all arriving parallel to the microscope axis (plane wave), they will be incident over a narrow range of angles centred on this same axis.

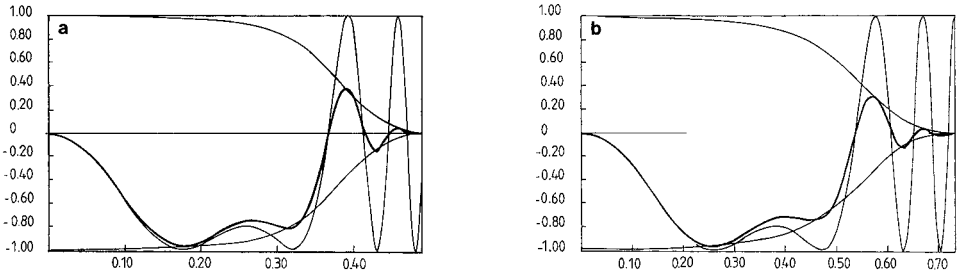


FIGURE 4. The attenuation of the transfer function (sin γ) due to partial coherence for the following conditions: (a) $U = 200$ kV; $C_s = 2$ mm; defocus = –841 Å; beam divergence (partial spatial coherence) semiangle = 0.6 mrad; defocus spread (partial temporal coherence) = 69 Å. (b) $U = 400$ kV; $C_s = 0.9$ nm; defocus = –457 Å; semiangle = 0.64 mrad; spread = 56 Å.

Each of these effects can be represented to a good approximation by an *envelope function*, in the sense that the coherent transfer function $\sin \gamma$ is to be multiplied by a function representing the effect of finite energy spread and another representing that of finite source size. These envelopes attenuate the transfer of information, and it is therefore important to ensure that, if high resolution is to be achieved, the microscope source is sufficiently small and has a narrow enough energy spread for this information to survive in the image. Figure 4 shows a typical situation.

In an attempt to combat the effect of energy spread, monochromators are being incorporated in electron microscopes. See Hawkes (2006) for references to this work and to attempts to correct chromatic aberration.

3.3. Real Imaging Conditions: Specimen Tilt

In practice, the incident beam, however close to perfection the spatial coherence may be, is not incident normal to the specimen surface for the simple reason that a 3D reconstruction is based on a set of closely spaced views through the specimen from different directions. These are commonly obtained by tilting the specimen, and any known symmetry properties of the structure may be used to supplement the information obtained from the tilted images. Alternatively, a single tilted image may be sufficient if the specimen consists of particles that tend to adopt a particular orientation on the specimen support but have random orientations in the plane of the latter. In all cases, however, the problem remains: the microscope transfer function is highly defocus dependent, and we have stated without comment that the microscopist must choose the defocus value with care and forethought. How can we reconcile this with the fact that, for specimen tilts that may reach 45° or more, the defocus difference between different sides of the image may well be of the order of micrometers? This problem has other ramifications of considerable importance. For example, averaging to

enhance the signal-to-noise ratio for low-dose images of fragile specimens is not possible if the defocus varies across the object; the contrast transfer equations themselves cannot be used in the form given above, and the necessary corrections introduce artefacts. Various proposals for overcoming the problem by digital processing of the recorded tilt images before 3D reconstruction have been made over the years (e.g. Amos *et al*., 1982; Henderson and Baldwin, 1986; Henderson *et al*., 1986; Schiske, 1982), but the most satisfactory solution is surely that developed by Zemlin (1989). This requires modification of the microscope to permit the image to be formed as a mosaic of small subimages, each created by illuminating a small zone of the specimen; the illumination spot is then moved to the adjoining zone until the whole specimen area has been covered. In this way, the defocus spread need never exceed that corresponding to the width of an illuminated zone since the microscope defocus can be adjusted as the illuminating spot climbs up the slope of the specimen.

This technique is not a panacea; there are so many conflicting requirements in high-resolution electron microscopy that any perturbation of the delicate balance, however advantageous, is almost certain to have drawbacks as well. In the case of this dynamic focusing idea, care must be taken to ensure that the inevitable loss of spatial coherence due to the decrease in spot size does not cause an unacceptable attenuation of the transfer function. Zemlin estimates that this diminution of the coherence can be kept within acceptable bounds, however. Furthermore, the situation that we have been considering is the most unfavorable; when the structure of isolated particles is to be established, the defocus difference across an individual particle will be much smaller, and the problem consequently less serious. For discussion of the choice of defocus for such specimens, see Radermacher (1988) and Chapter 8 below.

3.4. Real Imaging Conditions: Thick Specimens

For the vast majority of the specimens employed in 3D reconstruction, the image is the result of scattering contrast, as described in Section 2.2. Electrons scattered through large enough angles by the specimen atoms are intercepted by the objective aperture and, hence, are absent from the image. Their absence is indicated by dark contrast against the bright background and, as equation (7) indicates, the number of electrons per unit area at an image pixel is proportional to $\exp(-\mu/\mu_i)$, where μ is the mass thickness through the corresponding object area.

In an ideal recording system, therefore, in which the number of electrons per pixel could be measured, the logarithm of this quantity would represent a projection through the object directly. Cameras that do indeed yield such measurements accurately over a wide range of values of the number of incident electrons are routinely available. If the images are recorded on film, the situation is a little more complicated. At electron microscope energies,

every electron causes sufficient ionization of the grains of the photographic emulsion to render at least one crystal developable. If E electrons are incident per unit area of an emulsion consisting of silver halide crystals of mean cross-sectional area A, there will be on average EA collisions per crystal, so the fraction F of all crystals hit at least once becomes

$$F = 1 - \exp(-EA) \quad \text{(Poisson distribution)} \tag{28}$$

After development, the emulsion is characterized by its optical density D, which is defined in terms of the fraction of light transmitted through the emulsion T by

$$D = \log_{10}\left(\frac{1}{T}\right) \tag{29}$$

This may be written as

$$D = D_{\text{sat}}\{1 - \exp(-EA)\} \tag{30}$$

where D_{sat} is the saturation density, the value reached after an exposure long enough for every grain to be rendered developable. Very often, it is sufficient to expand the exponential and neglect quadratic and higher order terms, whereupon equation (30) becomes

$$D = D_{\text{sat}}AE \propto E \tag{31}$$

(For extensive discussion, see Valentine, 1966, and also Section 4.6.2 of Reimer, 1997.)

To summarize the result of this section, the natural logarithm of the optical density gives a direct measurement of the projected density through the structure, apart from a scaling factor.

The discussion of film has been retained for completeness, but this recording medium has been almost wholly superseded by the CCD camera, which has a wide dynamic range and hence furnishes information about the incident electron current density in a form ready for processing. (Nevertheless, the excellent resolution of film and its long lifetime, important for archival purposes, are very valuable features.) A disadvantage of CCD-based detectors is the indirect nature of the detection process: 'The incident electrons form a visible light image in a thin polycrystalline phosphor, which is imaged onto and recorded by the CCD, often after some demagnification in tapered fibre optics' (Faruqi *et al.*, 2003a). For this reason, a new generation of devices that detect electrons directly is currently being developed. Among the most promising are the 'hybrid pixel detectors', which generate very little readout noise and hence lead to true electron counting. For details of these new detectors, see Faruqi *et al.* (2003a, b, c, 2005), Faruqi and Cattermole (2005) and Evans *et al.* (2005).

3.5. Energy-filtered Imaging and EELS

When an electron beam traverses a specimen, a substantial fraction of the electrons pass through unaffected by their passage. The other electrons interact with the atoms in various ways. If the specimen atoms are mostly light and the specimen is thin, the majority of the interactions will be elastic—the electron loses very little energy during the collision—and, on average, electrons will interact with at most one atom. The collision is described as an elastic scattering event. In other conditions, however, the electrons may be scattered more than once (multiple scattering) and may transfer some of their energy to the specimen; we then speak of inelastic scattering.

When inelastic scattering occurs, the electrons emerging from the specimen will have different energies and the energy spectrum will be to some extent characteristic of the composition of the specimen. The study of such spectra is highly developed (Egerton, 1996) and an electron energy-loss spectrum atlas is available, with the aid of which individual atomic species can be identified (Ahn, 2004). In imaging conditions, however, the inelastically scattered electrons are undesirable. The chromatic aberration of electron lenses has the effect of blurring the image and it is therefore advantageous to remove these electrons. This can be achieved by means of an imaging energy filter.

The transmission electron microscope may be equipped with a post-column magnetic prism or an in-column magnetic filter, typically an Ω-filter (so called from the shape of the optic axis in the device); the most common commercial prism device is the GIF (Gatan Imaging Filter, marketed by Gatan). These accessories allow the microscopist to form an image with electrons in a chosen energy range or to form an energy-loss spectrum from a selected area of the specimen. The collection edited by Reimer (1995) and the book by Egerton (1996) are good background references, largely addressed to physicists, while the papers by Leapman (2003, 2004) and especially Leapman *et al.* (2004) are concerned with biological applications.

Energy filtering is used for various purposes. The energy window may be chosen so that only electrons that have lost very little energy when traversing the specimen are allowed to reach the image, as described above. In this case, only elastically scattered electrons will be selected and any blurring caused by chromatic aberration will be substantially reduced. Examples of the benefit of zero-loss filtering, in low-dose conditions in particular, are to be found in the work of Grimm *et al.* (1996a, 1998). Figure 5 shows the improvement obtained on an image of a *Thermoplasma acidophilum* cell and on the subsequent reconstruction. Alternatively, the window may be chosen in such a way that information about a particular chemical element is transmitted to the image. In the pioneering study of Leapman *et al.* (2004), the filter was set to detect phosphorus in cell sections of the nematode *Caenorhabditis elegans*. The authors explain in detail how the various

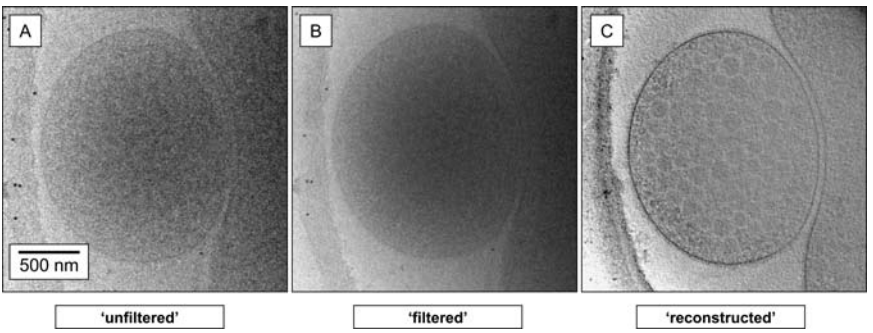


FIGURE 5. Unfiltered (A) and zero-loss (B) filtered images of a *Thermoplasma acidophilum* cell. The specimen thickness is 0.25–0.3 μm, which is greater than the mean free path of 300 kV electrons in ice at liquid nitrogen temperature. The slit width was 20 eV. (C) A section through a tomographic reconstruction based on the filtered image set. Courtesy of J. Plitzko, Max-Planck-Institut für Biochemie, Martinsried.

parameters were chosen. They obtained a tilt series in the angular range between ±55° at energy losses centered on 157 and 120 eV with an energy window of 15 eV. These values are just above and just below a feature in the energy-loss spectrum characteristic of phosphorus (the L_{23} edge). The position of phosphorus in ribosomes is clearly seen in the resulting 3D reconstruction. A third possibility, discussed by Grimm *et al*. (1996b), has been exploited by Olins *et al*. (1989) and more recently by Bouwer *et al*. (2004); examination of an energy-loss spectrum shows that although the spectrum has a fine structure characteristic of the individual elements comprising the specimen, it also exhibits general trends and in particular a 'most-probable' energy loss. By setting the energy selection window around this value, the image will be perceptibly sharper and the signal will not be as low as when the zero-loss region is selected, a very attractive feature. However, the technique is not appropriate for thin specimens, where elastic scattering is dominant. Olins *et al*. showed that sections up to 0.5 μm thick can be imaged using energy filtering, and they used the images to make a tomographic reconstruction of chromatin in the Balbiani ring of *Chironomus tentans*. Bouwer *et al.* reconstructed the spiny dendrite from the mouse hippocampus from a thick copper–lead-stained section. An example of the improvement is shown in Figs 6 and 7. In Fig. 6, unfiltered (left) and most-probable-loss (right) images are compared for a 2 μm thick section; in Fig. 7, a zero-loss image and a most-probable-loss image are compared.

Elemental reconstruction suffers from a number of hazards in addition to those now well known in regular electron tomography, notably radiation damage. Many of the most interesting elements for cell biologists are present in extremely low concentrations, and the background subtraction

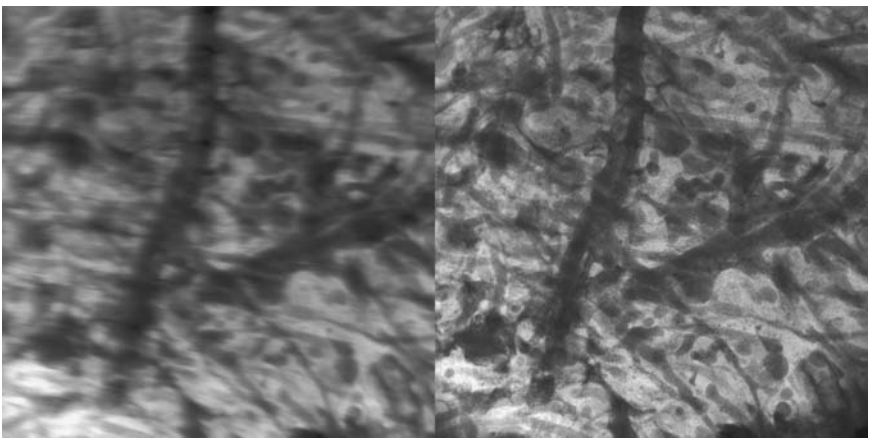


FIGURE 6. Unfiltered and most-probable-loss images of a 2 μm thick section of a spiny dendrite sample. For the filtered image, the slit was 40 eV wide and centered on 150 eV. Magnification ≈ 5000. The reduction of the blurring caused by chromatic aberration is obvious. (From Bouwer *et al.* (2004), reproduced with permission of Elsevier).

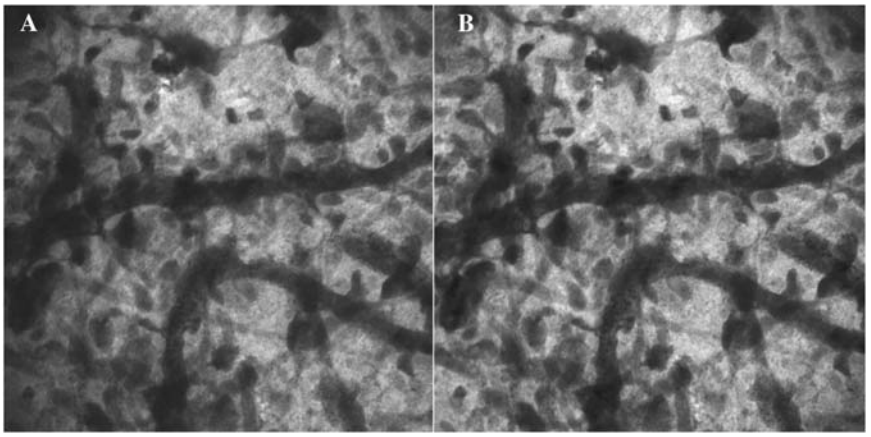


FIGURE 7. Zero-loss and most-probable-loss images of a 2 μm thick section of a spiny dendrite sample. The exposure times were adjusted to produce the same mean image intensity: 1200 s for the zero-loss case (slit width 15 eV) and 10 s for the most-probable-loss image (slit width 35 eV). The most-probable-loss signal is ~120 times higher than the zero-loss signal, while the resolution in the two images is comparable. Magnification ≈ 5000. (From Bouwer *et al.* (2004), reproduced with permission of Elsevier).

that is necessary to reveal the true signal requires great care. This is particularly true at high tilt angles. Also, the markers attached to the specimen for alignment of the individual members of the tilt series must be visible in the energy-filtered image (and not excluded by the choice of energy window).

4. THE MICROSCOPE AS PROJECTOR: ASSESSMENT

The object of this chapter is to discuss the extent to which the electron microscope provides a projection of some aspect of the specimen structure that can be used to form a reliable 3D reconstruction of the latter. There is always a danger that a critical examination of a complicated situation will give the reader the impression that the situation is, if not hopeless, certainly far from rosy. The list of things that can go wrong seems interminable: insufficient source coherence, a strongly scattering specimen, appreciable inelastic scattering, structural dimensions that cross the vague frontier between scattering contrast and phase contrast, loss of information thanks to the form of the contrast transfer function and inadequacy of the column approximation. And yet, there are plenty of successful 3D reconstructions in the literature, and the structures obtained seem in harmony with structural information obtained by other means and enable the molecular biologist, for example, to confirm or invalidate earlier hypotheses about structure and function. In this last section, we attempt to draw together the strands of the interwoven arguments of the preceding sections, and we conclude that, provided that certain cautionary remarks are borne in mind, the electron microscope image does contain the projection information on which a three-dimensional reconstruction can be safely based.

First, we must distinguish between stained and unstained specimens. If a heavy-atom stain is employed, then for several reasons the highest resolutions cannot be expected. The stain will itself not cling to all the fine details of the object and is often incapable of penetrating into all the narrow interstices of the structure. Furthermore, the presence of heavy atoms makes the weak scattering approximation untenable, so all the reasoning based on this becomes invalid. Conversely, the scattering properties of this heavy stain are excellent for creating scattering contrast, and such contrast does indeed represent a projection through the object. We recall that it is important to use not the optical density but its logarithm. A set of views through a stained specimen may be expected to yield a 3D reconstruction that is reliable down to the resolution of the specimen preparation and staining procedures, at best ~2 nm but sometimes worse than this. For a much fuller discussion of all practical aspects of this branch of 3D reconstruction, see the long review by Amos *et al.* (1982) and the book edited by Turner (1981), especially the chapter by Steven (1981), as well as later chapters of this volume.

When we wish to establish the 3D structure of a specimen at higher resolution, better than 1.5 or 2 nm, say, the situation changes radically. Very different techniques of specimen preservation are brought into play, the beam electrons convey information about the positions of light atoms, and the mechanism of image formation is that appropriate to a phase object.

If the information recorded in the image is to be used for reconstruction, and hence represents a projection through the specimen, certain

precautions must imperatively be taken. If not, the different views used for the reconstruction will at best be found to be incompatible, and the reconstruction methods described elsewhere in this book will fail to generate a high-resolution model. At worst, the views will be compatible but wrong, yielding a reconstructed structure that is false.

These precautions are of two kinds. First, we must be sure that the phase shift of the wave function leaving the specimen is essentially a projection through the latter, and, secondly, we must arrange that this phase information is converted faithfully into amplitude variations in the recorded image, at least in the size range of interest. In other words, certain rules must be respected during the preparation of the specimen, and the mode of observation in the microscope must likewise be rigorously controlled. So far as the specimen is concerned, it is most important that it can be represented by a multiplicative transparency (equation 13), with the implications that we have already mentioned: the beam accelerating voltage should not be too low, nor the specimen thickness too great. It is difficult to give figures, since the limiting thickness is a function of beam voltage and of the atomic number of the heaviest atoms present, but for unstained biological material observed between 100 and 200 kV the thickness should probably not exceed 10 or 20 nm.

For transfer along the microscope, we are obliged to use linear transfer theory, since no useful conclusions can be drawn from the more complicated non-linear theory that we obtain if we do not assume that the specimen scatters weakly. The beam convergence must be small, the energy spread small, and the defocus value suitably chosen. Some way of circumventing the problem of the variation in the defocus of a steeply tilted specimen must be found, such as that described in Section 3.3.

For the very highest resolution, all these precautions and many others relating to the actual reconstruction process and described in subsequent chapters are mandatory. For more modest, but still high, resolution, however, we may enquire what is likely to happen if we relax the requirements a little. If the specimen is thicker than ideally we should like, the columns of the 'column approximation' will broaden and the resolution of the phase-grating approximation will deteriorate slowly. Inelastic scattering will increase and add a lower resolution background to the image, which will in consequence display a lower signal-to-noise ratio. The curvature of the Ewald sphere will have a small effect on the fidelity of the reconstruction (Cohen *et al.*, 1984; Wolf *et al.*, 2006).

The theoretical assumption that is the worst respected in practice is, however, that the specimen is a weak scatterer and, hence, that the phase shifts introduced are small. Even if we relax this by regarding the specimen phase variation as a function with only small variations about its mean value, many specimens that preserve very fine detail will not be truly weak scatterers. It is difficult to discuss the consequences of semi-weak scattering, but we can anticipate that the effects will be, if a little misleading, not

catastrophic. If we continue the series expansion of $\exp(i\varphi)$, we obtain an alternating sequence of real and imaginary terms, of which the former will join the amplitude contrast term (s in equation 21) and the latter will alter the absolute value of the phase contrast term. Modest violation of the weak-scattering requirement should therefore not be too serious.

There are hazards in 3D reconstruction of many kinds. Here we have concentrated on a particular kind and have deliberately refrained from comment on those dealt with elsewhere in the book, notably radiation damage, which is extensively discussed by Luther in Chapter 1, and the practical problems of data acquisition. We have shown that at both high and modest resolution projection information can be recorded. How this information is used to generate three-dimensional structures is the subject of the remainder of this book.

Almost all the reviews on the various aspects of 3D reconstruction neglect the question examined in this chapter, taking it for granted that the electron microscope does project the specimen structure onto the image, or at least onto the exit surface of the object. The surveys by Frank and Radermacher (1986) and by Glaeser (1985) are rare exceptions and interest has been revived by Wolf *et al*. (2006) and Lawrence *et al*. (2006). The list of further reading is concerned with 3D reconstruction in general, therefore, and hardly at all with the specific topic of projection.

REFERENCES

Ahn, C. C. (ed.) (2004). *Transmission Electron Energy Loss Spectrometry in Materials Science and the EELS Atlas.* Wiley-VCH, Weinheim.

Amos, L. A., Henderson, R. and Unwin, P. N. T. (1982). Three-dimensional structure determination by electron microscopy of two-dimensional crystals. *Prog. Biophys. Mol. Biol.* **39**:183–231.

Arnal, F., Balladore, J. L., Soum, G. and Verdier, P. (1977). Calculation of the cross-sections of electron interaction with matter. *Ultramicroscopy* **2**:305–310.

Born, M. and Wolf, E. (1999). *Principles of Optics.* 7th ed., Cambridge University Press, Cambridge.

Bouwer, J. C., Mackey, M. R., Lawrence, A., Deerinck, T. J., Jones, Y. Z., Terada, M., Martone, M. E., Peltier, S. and Ellisman, M. H. (2004). Automated most-probable loss tomography of thick selectively stained biological specimens with quantitative measurement of resolution improvement. *J. Struct. Biol.* **148**:297–306.

Cohen, H. A., Schmid, M. F. and Chiu, W. (1984). Estimates of validity of projection approximation for three-dimensional reconstructions at high resolution. *Ultramicroscopy* **14**:219–226.

Deans, S. R. (1983). *The Radon Transform and Some of Its Applications.* Wiley-Interscience, New York.

DeRosier, D. J. (1971). The reconstruction of three-dimensional images from electron micrographs. *Contemp. Phys.* **12**:437–452.

DeRosier, D. J. and Klug, A. (1968). Reconstruction of three-dimensional structures from electron micrographs. *Nature* **217**:130–134.

Egerton, R. F. (1996). *Electron Energy-loss Spectroscopy in the Electron Microscope*, 2nd edn. Plenum, New York.

Evans, D. A., Allport, P. P., Casse, G., Faruqi, A. R., Gallop, B., Henderson, R., Prydderch, M., Turchetta, R., Tyndel, M., Velthuis, J., Villani, G. and Waltham, N. (2005). CMOS active pixel sensors for ionising radiation. *Nucl. Instrum. Meth. Phys. Res.* **A546**:281–285.

Faruqi, A. R. and Cattermole, D. M. (2005). Pixel detectors for high-resolution cryo-electron microscopy. *Nucl. Instrum. Meth. Phys. Res.* **A549**: 192–198.

Faruqi, A. R., Cattermole, D. M. and Raeburn, C. (2003a). Applications of pixel detectors to electron microscopy. *Nucl. Instrum. Meth. Phys. Res.* **A512**:310–317.

Faruqi, A. R., Cattermole, D. M. and Raeburn, C. (2003b). Direct electron detection methods in electron microscopy. *Nucl. Instrum. Meth. Phys. Res.* **A513**:317–331.

Faruqi, A. R., Cattermole, D. M., Henderson, R., Mikulec, B. and Raeburn, C. (2003c). Evaluation of a hybrid pixel detector for electron microscopy. *Ultramicroscopy* **94**:263–276.

Faruqi, A. R., Henderson, R., Prydderch, M., Allport, P. and Evans, D. A. (2005). Direct single-electron detection with a CMOS detector for electron microscopy. *Nucl. Instrum. Meth. Phys. Res.* **A546**:170–175.

Frank, J. and Radermacher, M. (1986). Three-dimensional reconstruction of nonperiodic macromolecular assemblies from electron micrographs. In *Advanced Techniques in Biological Electron Microscopy III* (J. K. Koehler, ed.). Springer, Berlin, pp. 1–72.

Glaeser, R. M. (1982). Electron microscopy. In *Methods of Experimental Physics* (G. Ehrenstein and H. Lecar, eds.), Vol. 20. Academic Press, New York, pp. 391–444.

Glaeser, R. M. (1985). Electron crystallography of biological macromolecules. *Annu. Rev. Phys. Chem.* **36**:243–275.

Glaeser, R. M. and Ceska, T. A. (1989). High-voltage electron diffraction from bacteriorhodopsin (purple membrane) is measurably dynamical. *Acta Crystallogr. A* **45**:620–628.

Glaeser, R. M. and Downing, K. H. (1993). High-resolution electron crystallography of protein molecules. *Ultramicroscopy* **52**:478–486.

Glaeser, R. M., Downing, K. H., DeRosier, D. J., Chiu, W. and Frank, J. (2007). *Electron Crystallography of Biological Macromolecules*. Oxford University Press, New York and Oxford.

Glaser, W. (1952). *Grundlagen der Elektronenoptik.* Springer, Vienna.

Grimm, R., Koster, A. J., Ziese, U., Typke, D. and Baumeister, W. (1996). Zero-loss energy filtering under low-dose conditions using a post-column energy filter. *J. Microsc.* **183**: 60–68.

Grimm, R., Typke, D. and Baumeister, W. (1998). Improving image quality by zero-loss energy filtering: quantitative assessment by means of image cross-correlation. *J. Microsc.* **190**:339–349.

Grinton, G. R. and Cowley, J. M. (1971). Phase and amplitude contrast in electron micrographs of biological material. *Optik* **34**:221–233.

Han, K. F., Gubbens, A. J., Sedat, J. W. and Agard, D. A. (1996). Optimal strategies for imaging thick biological specimens: exit wavefront reconstruction and energy-filtered imaging. *J. Microsc.* **183**:124–132.

Hawkes, P. W. (1980a). Image processing based on the linear theory of image formation. In *Computer Processing of Electron Microscope Images* (P. W. Hawkes, ed.). Springer, Berlin, pp. 1–33.

Hawkes, P. W. (1980b). Units and conventions in electron microscopy, for use in ultramicroscopy. *Ultramicroscopy* **5**:67–70.

Hawkes, P. W. (2006). Aberration correction. In *Science of Microscopy* (Hawkes, P. W. and Spence, J. C. H., eds). Springer, New York and Berlin, pp. 696–747.

Hawkes, P. W. and Kasper, E. (1994). *Principles of Electron Optics*, Vol. 3. Academic Press, London.

Henderson, R. and Baldwin, J. M. (1986). Treatment of the gradient of defocus in images of tilted, thin crystals. In *Proceedings of the 44th Annual Meeting of EMSA* (G. W. Bailey. ed.). San Francisco Press, San Francisco, pp. 6–9.

Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. and Zemlin, F. (1986). Structure of purple membrane from *Halobacterium halobium*: Recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. *Ultramicroscopy* **19**:147–178.

Ho, M.-H., Jap, B. K. and Glaeser, R. M. (1988). Validity domain of the weak-phase-object approximation for electron diffraction of thin protein crystals. *Acta Crystallogr. A* **44**:878–884.

Hsieh, C.-E., Marko, M., Frank, J. and Mannella, C. (2002). Electron tomographic analysis of frozen hydrated tissue sections. *J. Struct. Biol.* **138**:63–73.

Jap, B. K. and Glaeser, R. M. (1980). The scattering of high-energy electrons. II: Quantitative validity domains of the single-scattering approximations for organic crystals. *Acta Crystallogr. A* **36**:57–67.

Kirkland, A. I. and Hutchison, J. L. (2006) Atomic resolution transmission electron microscopy. In *Science of Microscopy* (P. W. Hawkes and J. C. H. Spence, eds). Springer, New York and Berlin, pp. 3–64.

Lawrence, A., Bouwer, J. C., Perkins, G. and Ellisman, M. H. (2006). Transform-based back-projection for volume reconstruction. *J. Struct. Biol*. **154**:144–167.

Leapman, R. D. (2003). Detecting single atoms of calcium and iron in biological structures by electron energy-loss spectrum-imaging. *J. Microsc.* **210**:5–15.

Leapman, R. D. (2004). Novel techniques in electron microscopy. *Curr. Opin. Neurobiol.* **14**:591–598.

Leapman, R. D., Koksis, E., Zhang, G., Talbot, T. L. and Laquerriere, P. (2004). Three-dimensional distributions of elements in biological samples by energy-filtered electron tomography. *Ultramicroscopy* **100**:115–125.

Olins, A. L., Olins, D. E., Levy, H. A., Margle, S. M., Tinnel, E. P. and Durfee, R. C. (1989). Tomographic reconstruction from energy-filtered images of thick biological sections. *J. Microsc.* **154**:257–265.

Plitzko, J. and Baumeister, W. (2006). Cryo-electron tomography. In *Science of Microscopy* (Hawkes, P. W. and Spence, J. C. H., eds.). Springer, New York and Berlin, pp. 535–604.

Radermacher, M. (1988). Three-dimensional reconstruction of single particles from random and nonrandom tilt series. *J. Electron Microsc. Tech.* **9**:359–394.

Radon, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Ber. Verh. K. Sächs. Ges. Wiss. Leipzig, Math.-Phys. Kl.* **69**:262–277. For an English translation, see Deans (1983).

Reichelt, R. and Engel, A. (1984). Monte Carlo calculations of elastic and inelastic electron scattering in biological and plastic materials. *Ultramicroscopy* **13**:279–294.

Reimer, L. (1997). *Transmission Electron Microscopy*, 3rd edn. Springer, Berlin and New York.

Reimer, L. and Sommer, K. H. (1968). Messungen und Berechnungen zum elektronenmikroskopischen Streukontrast für 17 bis 1200 keV-Elektronen. *Z. Naturforsch.* **23a**: 1569–1582.

Reimer, L. (ed.) (1995). *Energy-filtering Transmission Electron Microscopy.* Springer, Berlin.

Robards A. W. and Sleytr, U. B. (1985). *Low Temperature Methods in Biological Electron Microscopy.* Elsevier, New York.

Saxton, W. O. (1986). Focal series restoration in HREM. In *Proceedings of the XIth International Congress on Electron Microscopy* suppl. to *J. Electron Microsc.* **35**, Post-deadline paper 1.

Schiske, P. (1968). Zur Frage der Bildrekonstruktion durch Fokusreihen. In *Electron Microscopy 1968* (D. S. Bocciarelli, ed.), Vol. I. Tipografia Poliglotta Vaticana, Rome, pp. 147–148.

Schiske, P. (1973). Image processing using additional statistical information about the object. In *Image Processing and Computer-aided Design in Electron Optics* (P. W. Hawkes, ed.). Academic Press, New York, pp. 82–90.

Schiske, P. (1982). *A posteriori* correction of object tilt for the CTEM. *Ultramicroscopy* **9**:17–26.

Spence, J. C. H. (2003). *Experimental High-resolution Electron Microscopy*, 3rd edn. Oxford University Press, New York.

Steinbrecht, R. A. and Zierold, K. (eds.) (1987). *Cryotechniques in Biological Electron Microscopy.* Springer, Berlin.

Steven, A. C. (1981). Visualization of virus structure in three dimensions. In *Methods in Cell Biology, Vol. 22, Three-Dimensional Ultrastructure in Biology* (J. N. Turner, ed.). Academic Press, New York, pp. 297–323.
Turner, J. N. (ed.) (1981). *Methods in Cell Biology, Vol. 22, Three-Dimensional Ultrastructure in Biology*. Academic Press, New York.
Valentine, R. C. (1966). The response of photographic emulsions to electrons. *Adv. Opt. Electron Microsc.* **1**:180–203.
Williams, D. B. and Carter, C. B. (1996). *Transmission Electron Microscopy.* Plenum, New York.
Wolf, M., DeRosier, D. J. and Grigorieff, N. (2006). Ewald sphere correction for single-particle electron microscopy. *Ultramicroscopy* **106**:376–382.
Zeitler, E. (ed.) (1982). Cryomicroscopy and radiation damage. *Ultramicroscopy* **10**:1–178.
Zeitler, E. (ed.) (1984). Cryomicroscopy and radiation damage. II. *Ultramicroscopy* **14**:161–316.
Zemlin, F. (1989). Dynamic focusing for recording images from tilted samples in small-spot scanning with a transmission electron microscope. *J. Electron Microsc. Tech.* **11**:251–257.

FURTHER READING

Baumeister, W. (2002). Electron tomography: towards visualizing the molecular organization of the cytoplasm. *Curr. Opin. Struct. Biol.* **12**:679–684.
Baumeister, W. (2004). Mapping molecular landscapes inside cells. *Biol. Chem.* **385**:865–872.
Baumeister, W. (2005). From proteomic inventory to architecture. *FEBS Lett.* **579**:933–937.
Baumeister, W. (2005). A voyage to the inner space of cells. *Protein Sci.* **14**:257–269.
Diociaiuti, M. (2005). Electron energy loss spectroscopy microanalysis and imaging in the transmission electron microscope: example of biological applications. *J. Electron Spectrosc. Rel. Phenom.* **143**:189–203.
Fernández, J.-J., Lawrence, A. F., Roca, J., García, I., Ellisman, M. H. and Carazo, J.-M. (2002). High-performance electron tomography of complex biological specimens. *J. Struct. Biol.* **138**:6–20.
Frank, J. (1973). Computer processing of electron micrographs. In *Advanced Techniques in Biological Electron Microscopy* (J. K. Koehler, ed.). Springer, Berlin, pp. 215–274.
Frank, J. (1980). The role of correlation techniques in computer image processing. In Hawkes (1980), pp. 187–222.
Frank, J. (1981). Introduction *and* Three-dimensional reconstruction of single molecules. In *Methods in Cell Biology, Vol. 22, Three-Dimensional Ultrastructure in Biology* (J. N. Turner, ed.). Academic Press, New York, pp. 119–213, 325–344.
Frank, J. (1989). Image analysis of single macromolecules. *Electron Microsc. Rev.* **2**:53–74.
Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biomol. Struct.* **331**:303–319.
Frank, J., Wagenknecht, T., McEwen, B. F., Marko, M., Hsieh, C.-E. and Mannella, C. (2002). Three-dimensional imaging of biological complexity. *J. Struct. Biol.* **138**:85–91.
Hawkes, P. W. (ed.) (1980). *Computer Processing of Electron Microscope Images.* Springer, Berlin.
Hawkes, P. W. and Spence, J. C. H. (eds) (2006). *Science of Microscopy*. Springer, New York and Berlin.
Henderson, R. and Glaeser, R. M. (1985). Quantitative analysis of image contrast in electron micrographs of beam-sensitive crystals. *Ultramicroscopy* **16**:139–150.
Hoppe, W. and Hegerl, R. (1980). Three-dimensional structure determination by electron microscopy (nonperiodic specimens). In Hawkes (1980), pp. 127–185.
Hoppe, W. and Typke, D. (1979). Three-dimensional reconstruction of aperiodic objects in electron microscopy. In *Advances in Structure Research by Diffraction Methods* (W. Hoppe and R. Mason, eds.), Vol. 7. Vieweg, Braunschweig, pp. 137–190.
Lewitt, R. M. (1983). Reconstruction algorithms: transform methods. *Proc. IEEE* **71**:390–408.

Lucic, V., Förster, F. and Baumeister, W. (2005). Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem.* **74**:833–865.

McEwen, B. F. and Frank, J. (2001). Electron tomographic and other approaches for imaging molecular machines. *Curr. Opin. Neurobiol.* **11**:594–600.

Mellema, J. E. (1980). Computer reconstruction of regular biological object. In Hawkes (1980), pp. 89–126.

Moody, M. F. (1990). Image analysis of electron micrographs. In *Biophysical Electron Microscopy* (P. W. Hawkes and U. Valdrè, eds.), Academic Press, New York, pp. 145–287.

Stewart, M. (1988). Introduction to the computer image processing of electron micrographs of two-dimensionally ordered biological structures. *J. Electron Microsc. Tech.* **9**:301–324.

Stewart, M. (1988). Computer image processing of electron micrographs of biological structures with helical symmetry. *J. Electron Microsc. Tech.* **9**:325–358.

Subramanian, S. and Milne, J. L. S. (2004). Three-dimensional electron microscopy at molecular resolution. *Annu. Rev. Biophys. Biomol. Struct.* **33**:141–155.

4

Cryotomography: Low-dose Automated Tomography of Frozen-hydrated Specimens

Abraham J. Koster and Montserrat Bárcena

Abraham J. Koster • Leiden University Medical Center, Molecular Cell Biology, Einthovenweg 20, Leiden, The Netherlands

Montserrat Bárcena • Universiteit Utrecht, Department of Molecular Cell Biology, Kruyt Building, Padualaan 8, 3584 CH Utrecht, The Netherlands

Electron tomography is an imaging technique that provides 3D images of a specimen with nanometer scale resolution. The range of specimens that can be investigated with this technique is particularly wide, from large (500–1000 nm) unique variable structures such as whole cells to suspensions of thousands of small identical macromolecules (>200 kDa). When applied to cryofixed frozen-hydrated biological material, the technique is often referred to as cryotomography. In combination with automated low-dose data collection and advanced computational methods, such as molecular identification based on pattern recognition, cryotomography can be used to visualize the architecture of small cells and organelles and/or to map macromolecular structures in their cellular environment. The resolution that can be obtained with cryotomography depends on several fundamental and technical issues related to specimen preparation, microscopy and subsequent image processing steps, but will typically be in the range of 5–10 nm.

1. INTRODUCTION

1.1. Basic Concept

Most objects, either biological or inorganic, have a 3D architecture. The higher the complexity of an object, the less revealing is the 2D image that is obtained with transmission electron microscopy (TEM) due to the superposition of multiple 3D structural features into one 2D projection image. 3D imaging of cellular structures has great impact on how we understand the cellular architecture, and provides a powerful tool to expand upon the 2D images obtained of cell biological structures when studied by conventional 2D TEM during the last decades.

Because of the large depth of focus of the instrument, electron micrographs taken with TEM are essentially 2D projections of the imaged specimen. In electron tomography, 2D images of a specimen are acquired as viewed from different angles and then synthesized into a 3D mass density map, often referred to as a tomogram. The specimen holder is tilted incrementally around an axis perpendicular to the electron beam, e.g. from –65° tilt to +65° tilt with 1° increments, and images are taken at each position. Before computation of the tomogram, the projection images must be mutually aligned within a common frame of reference.

The 3D reconstruction (tomogram) is computed either in real space or via interpolation in Fourier space and subsequent back-transformation. The description in Fourier space helps to explain the requirements for data

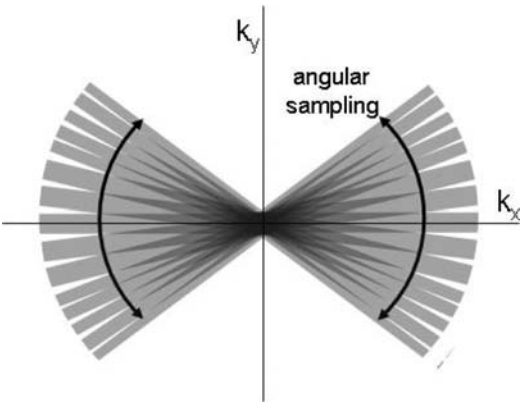


FIGURE 1. Data sampling in Fourier space. The description in Fourier space (k_x, k_y) tells us that a 2D image of the tilt series corresponds to a central section in the 3D Fourier transform of the specimen that is imaged. The collection of a tilt series is equivalent to sampling the specimen information in Fourier space. A projection of an object with a thickness D corresponds to a central slice of thickness $1/D$ in Fourier space. The central section is perpendicular to the projection direction. The sampling is done with angular increments sufficient to obtain a resolution up to $\pi D/N$ where the information of adjacent projections overlaps (black arrows). When a tilt series of a section is acquired, the angular tilt increments become smaller at higher tilt angles.

recording, and its dependence on the size of the specimen and on the desired resolution of the tomogram. The description in Fourier space tells us that a 2D image of the tilt series corresponds to a central section in the 3D Fourier transform of the specimen that is imaged: acquisition of a tilt series with incremental tilts distributed over a large angular range is equivalent to scanning the specimen information in Fourier space by the corresponding set of sections (Fig. 1). The full structural information to a given resolution can be recorded by tilting about a single axis over an angular range of 180°.

The relationship between the attainable resolution d (parallel to the optical axis) is determined by the angular tilt angle increments $\Delta\alpha$. For linear tilt increments, d is given by the relationship $d = \pi D/N$ (Crowther *et al.*, 1970; Grimm *et al.*, 1998; Hoppe, 1969), where D is the diameter of a spherical object, and N the number of projections recorded at equally spaced tilt angles over a range of 180° (also given by $180/\Delta\alpha$). Thus, when 131 images are taken of an organelle 300 nm in size, the attainable resolution d equals 7 nm ($\pi \times 300/131$). This rule of thumb for the attainable resolution gives an indication for the data collection parameters that have to be used. For instance, in order to reconstruct a specimen of 50 nm diameter (D) with a resolution of 2 nm (d), one would need 75 projections (N) and thus a tilt increment of ~2.4°.

The relationship $d = \pi D/N$ is only valid if the geometric thickness of the specimen is independent of the tilt angle (i.e. if the specimen is spherical or cylindrical), and if the sample can be fully tilted, from –90° to 90°. In practice, it is not possible to collect a tilt series over the full angular range of 180°. At high tilts, typically at tilts higher than 65–70°, two kinds of problems arise. First, the grid-bars will eventually obscure the field of view, even using mesh sizes larger than 200. The design of the specimen holders also imposes mechanical constraints to the achievable tilt angle. A second type of problem is that at high tilt angles, the effective specimen thickness rapidly increases. At high tilts, very few electrons will be able to penetrate through the specimen to give a useful contribution to the image. These two effects limit the angular tilt range to a maximum tilt angle (α_{max}). Furthermore, most specimens in electron microscopy, far from being spherical, are quite extended in the *x*- and *y*-directions, but have a limited thickness in the *z*-direction. For a section with a thickness of *T*, the value of *D* equals $T/\cos(\alpha_{max})$ where α_{max} is the maximum tilt angle (McEwen *et al.*, 2002; Radermacher, 1992). For an optimal sampling of specimens of variable thickness, non-equidistant tilt angles may be advantageous (Grimm *et al.*, 1998; Saxton *et al.*, 1984). The basic idea is that the 3D Fourier space corresponding to the specimen under investigation is more efficiently sampled at higher specimen tilt angles. With this data collection scheme, the tilt increments are smaller at high specimen tilts than at low specimen tilts. The advantage is that the number of images that need to be acquired to attain a certain resolution is lower with this scheme than with a linear, equidistant scheme. The non-linear tilting scheme has advantages for low-dose tomographic data collection (Grimm *et al.*, 1998) and will be described in more detail later in the chapter.

In 3D Fourier space, the missing information at angles higher than α_{max} has the shape of a wedge, and therefore this effect is referred to as the missing wedge. The missing wedge leads to an elongation factor in the *z*-direction. The elongation depends on the size of the wedge given by $\mathrm{sqrt}((\alpha_{max} + \sin\alpha_{max}\cos\alpha_{max})/(\alpha_{max} - \sin\alpha_{max}\cos\alpha_{max}))$ where α_{max} is the maximum tilt angle (Radermacher, 1992). For a tilt range of +65 to –65, the elongation factor would be ~1.5. For example, the attainable resolution in the *z*-direction for the 300 nm section mentioned above would be 10.4 instead of 6.9 nm.

These effects due to the missing wedge inspired the development of dual-axis tilt electron tomography (Mastronarde, 1997; Penczek *et al.*, 1995). With the double-axis tilt electron tomography technique, two tilt series are collected. The second tilt series is taken after rotating the object by an angle of ~90°. In this manner, the missing information is much less: in Fourier space, the missing information will not have the shape of a wedge, but of a pyramid. For room temperature holders, double-axis tomography has become the *de facto* standard data collection scheme. For cryotomography, technical issues related to the cryoholders with a suitable mechanical con-

struction for sample rotation and tilting have obstructed routine application of double-axis tilting, Though experiments were done using double-axis tilting (Nickell *et al.*, 2003) with cryotomography, it is expected that more routine application will be possible with improved cryoholders in the near future.

The geometry of sampling in 3D Fourier space (one/two orthogonal single-axis tilt series as well as the missing wedge/pyramid of information) is the reason why the resolution of a tomogram will not be the same in all directions; the tomogram will have an anisotropic resolution (Baumeister and Steven, 2000). This anisotropic resolution has to be taken into account during the analysis of the tomogram. For instance, in some directions within the tomogram of a membrane, connections may be present and in other directions they may not (Nicastro *et al.*, 2000).

To be able to study the samples embedded in a vitreous ice environment, the specimen is investigated at liquid nitrogen temperature. Cooling the specimen further down to liquid helium temperature (4 K) has been shown to be advantageous for thin samples, but it is unclear if this is also true for thicker samples as used in cryotomography (Dubochet *et al.*, 1988; Talmon, 1987). The sensitivity of the specimen to the irradiating electron beam is a fundamental limitation of cryotomography. To avoid unnecessary irradiation of the specimen, electron tomography on frozen-hydrated specimens has to be done in the 'low-dose' mode, and the total dose the specimen receives should be spent on the acquisition of the images of a tilt series. Automated procedures for low-dose data acquisition were developed to satisfy the latter requirement (Dierksen *et al.*, 1993). Because of the inherent low contrast of frozen-hydrated specimens and the sensitivity of the specimen to the electron beam, cryotomograms are characterized by a low signal-to-noise ratio (SNR). Apart from the electronic noise due to the detector itself, a large contribution to the noise is due to statistical fluctuations in the number of electrons counted by the detector. This noise, often described as shot noise, is fundamental and cannot be reduced. The dependence of the resolution on noise and electron dose was discussed in detail previously (Grimm *et al.*, 1997; McEwen *et al.*, 2002).

Though it is possible to give an indication of what resolution can be attributed for a given specimen and for a given data collection scheme, this does not imply that that resolution is indeed obtained. Unfortunately, though progress has been made in defining standards for resolution estimation (Cardone *et al.*, 2005; Penczek, 2002; Unser *et al.*, 2005), straightforward quantitative verification of resolution is not available for electron tomography. In single-particle techniques, the Fourier shell correlation (FSC) approach (Saxton and Baumeister, 1982) is often used to determine the resolution. With FSC, it is possible to measure to what resolution two independent reconstructions contain the same structural details. For electron tomography, this approach is less useful since it is based on

comparing tomograms that are produced from two independent data sets (a data set split in two halves). For cryotomography, the approach is even less useful since the number of projections in a tilt series will often be <100. As a result, the measured resolution as given by the FSC will be largely underestimated. If the specimen allows, a practical approach could be to assess the resolution based on the presence of a known molecular structure (e.g. a known atomic model of the 20S proteasome) within a tomogram. The putative structure as observed within the tomogram with, for instance, 4 nm resolution could then be compared with its known high-resolution model rendered down to the same 4 nm resolution. A number of experimental methods to assess resolution in cryotomography are discussed by McEwen *et al.* (2002).

When repetitive structures are present within a tomogram, such as many copies of a macromolecular structure, it is possible to increase the resolution using 3D averaging techniques (Walz *et al.*, 1997b). For the application of these averaging techniques, it is important to realize that the defocus of the microscope has to be sufficiently small to ensure sufficient contrast transfer. Also important to realize is that structural variations may be present within the tomogram that could be averaged out during the process. Therefore, careful multivariate statistical methods are needed to verify to what extent the averaged result represents the true structures within the tomogram (Walz *et al.*, 1997b). Finally, before aligning, classification and averaging in electron tomograms, it has to be taken into account that the effect of the missing wedge on each of the structures encapsulated in the tomogram is very significant (Bohm *et al.*, 2000; Frangakis *et al.*, 2002). Recently, this averaging approach, often used as a technique extending upon template matching, has matured and been applied in several cases (Forster *et al.*, 2005).

Electron tomography of frozen-hydrated cellular structures at a resolution of 2–5 nm allows the study of the 3D organization of cellular structural components at a level that is sufficient to identify individual large macromolecular assemblies and to visualize connections between cellular components (Frangakis *et al.*, 2002; Grünewald *et al.*, 2003). The cellular architecture, supramolecular organization and interactions can be studied by analyzing their spatial relationships. The attainable resolution range and the size of the objects that can be imaged makes it well suited for studying cell organelles and subcellular molecular assemblies. Cryotomography is a tool to investigate the cellular architecture and networks of large macromolecular assemblies *in situ* and will be most powerful when it is applied in combination with other techniques, both with higher resolution (X-ray and nuclear magnetic resonance (NMR)) and with lower resolution (fluorescence light microscopy), as well as genomics and proteomics approaches (Baumeister, 2002; Baumeister and Steven, 2000; Baumeister *et al.*, 1999; Frank *et al.*, 2002; McEwen and Frank, 2001; Russell *et al.*, 2004; Steven and Aebi, 2003; Subramaniam and Milne, 2004).

1.2. Historical Perspective

The basic concept of electron tomography is that 3D reconstructions can be obtained from a set of 2D projections (Radon, 1917). The approach was put forward in 1968 in three independent papers (DeRosier and Klug, 1968; Hart, 1968; Hoppe, 1969). The possibility of identifying macromolecular structures in frozen cells or tissue was already stated in Hart's paper of 1968 and re-addressed again in a review in 1997 (Koster *et al.*, 1997) by Baumeister's group (Max-Planck-Institute for Biochemistry, Martinsried, Germany). In 1968, the technical challenges to carry out (cryo)tomography were substantial, and it took 30 years before the technique could be applied in practice. It was not until the 1980s that cryoelectron microscopy was developed (Dubochet *et al.*, 1988) and the technical requirements to collect data at very low dose under cryogenic conditions could be met. Moreover, the computational power required to process data sets and to visualize the noisy tomograms was just not available.

During the 1980s, significant pioneering research and development on (cryo)tomography was carried out by Baumeister's group. Fundamental high-risk technological developments were pursued by several members of that group. In particular, Dieter Typke and Reiner Hegerl were pivotal in several research initiatives related to instrumentation development (lenses and high-tilt specimen holders), to aspects of automation, to novel approaches in reconstruction and application of statistical image analysis techniques during the 1980s, 1990s and early 2000s. In the early 1990s, the first automated data collection system for cryotomography was realized, and was optimized during the next few years (Dierksen *et al.*, 1992, 1993, 1995; Grimm *et al.*, 1998; Horowitz *et al.*, 1997; Ratho *et al.*, 1997; Walz *et al.*, 1997a). In the early 1990s, computer-controllable TEMs and sensitive electron image detectors became available. Pioneered by the group of David Agard and John Sedat (UCSF, Californa), the mechanical housing and computer control of cooled slow-scan CCD (charge-coupled device) cameras were adopted from applications in wide-field light microscopy to meet the requirements of electron microscopy. Once the CCD interfacing was sorted out, automated tomography became possible (Koster *et al.*, 1992). Also, the computational power of desktop computing increased rapidly and the required computational methods for electron tomography were optimized (Böhm *et al.*, 2000; Frangakis *et al.*, 2001; Walz *et al.*, 1997b). At the beginning of 2000, many of the technological improvements were realized and commercial instrumentation for cryotomography became available. Cryotomography became a tool for both cell biology and structural biology applications.

With cryotomography, a major practical obstacle is the sensitivity of unstained ice-embedded samples to the electron beam irradiation. This sensitivity requires that electron tomographic data collection has to be done under extreme low-dose conditions, typically <5000–10,000 e/nm^2. Though

the dose can be distributed over a large number of projections during data collection (McEwen *et al.*, 1995), low-dose data collection of low-contrast frozen-hydrated samples will result in 3D reconstructions with a very low SNR (De Carlo *et al.*, 2002). A major step forward to overcome this limitation was taken in the late 1990s when it was recognized that it is not indispensable to identify the 3D molecular structures by direct visual inspection. Instead, the identification of a given 3D molecular structure, often referred to as a template (Böhm *et al.*, 2000), in noisy electron tomograms could also be based on pattern recognition methods. The recognition of the template may be statistically significant even if the particle is hardly recognizable by eye.

Since 2000, interest in cryotomography has increased enormously, and novel methodological approaches were proposed and realized by several groups worldwide to collect data sets more automatically (Bouwer *et al.*, 2004; Nickell *et al.*, 2005; Rath *et al.*, 1997; Zhang *et al.*, 2001, 2003; Zheng *et al.*, 2004; Ziese *et al.*, 2002a). Besides freezing suspensions of molecules and organelles, approaches to section vitrified cells and tissue, possibly using vibrating knives, were investigated by several groups, including those of Joachim Frank, Mike Marko and Bruce McEwen (Albany), Jacques Dubochet (Lausanne) and Wolfgang Baumeister (Martinsried) (Al-Amoudi *et al.*, 2004, 2005; Hsieh *et al.*, 2002; Sitte, 1996), and improved methods to transfer cryosamples from the laboratory into the electron microscope with a minimum amount of perturbation were proposed. Furthermore, dedicated image processing techniques to improve the reconstruction step (Liu *et al.*, 1995; Van Aert *et al.*, 2002) and visualization and interpretation were developed (Bajaj *et al.*, 2003; Baumeister, 2005; Bohm *et al.*, 2000; Frangakis and Hegerl, 2002; Frangakis *et al.*, 2001; McEwen *et al.*, 2002; Perkins *et al.*, 1997; Ress *et al.*, 2004; Sali *et al.*, 2003; Volkmann, 2002).

In this chapter, we will focus on the principles and basic procedures of electron tomography as they are applied to the 3D imaging of biological objects prepared for TEM with the best possible structural preservation currently available: fully frozen-hydrated specimens. In cryotomography, four phases can be distinguished: specimen preparation, data acquisition, reconstruction and visualization. Successful application of the technique depends on a large number of factors, ranging from critical specimen preparation conditions to the correct application of complex image analysis tools. Though several technical aspects of electron tomography on frozen-hydrated samples are addressed to identify the limitations and the potential of the technique, in this rapidly developing field we cannot hope to be complete and apologize for having not referred to important work and developments.

1.3. Applications of Cryotomography

Thanks to technical improvements of instrumentation, cryotomography became more applicable to biological systems in the late 1990s. Though

it is difficult to make a clear division, one can try to distinguish between four different types of applications.

First, cryotomography can be applied on cryofixed suspensions of multiple copies of macromolecular complexes embedded in a thin layer of vitrified buffer. High-resolution information, better than 4 nm resolution, can be attained by averaging multiple subtomograms containing individual copies of the macromolecular structure. It is important for this kind of application that the structures under investigation are in principle identical, although possible conformational changes can be explored by an extensive use of statistical analysis techniques. Characteristic of this approach is the acquisition of a large number of tomograms to increase the resolution. Furthermore, to attain even higher resolutions, a combination of single-particle electron microscopy with electron tomography can be carried out (Walz *et al.*, 1997b, 1999). In order to align subtomograms containing individual particles, sufficient information within those subtomograms needs to be present (Henderson, 1995). The implications of Henderson's work are discussed in a review by Frank (2002), where it is noted that close to atomic resolution is feasible, provided the particle exceeds a certain size. To approach that resolution, the number of particles that have to be averaged has to be significantly high. The number is roughly the cube of $1/d$ (d being the resolution). In the case of a ribosome, a resolution in the range of 1 nm can be achieved by analyzing ~30,000 particles. Therefore, to attain 0.3 nm, it would require $\sim 27 \times 30{,}000$ particles: close to a million particles. In practice, with electron tomography, the averaging of substructures was carried out with significantly lower numbers and at much lower resolution. Examples of the approach where multiple molecules are averaged are the acrosomal bundles (Sherman *et al.*, 1997), the thermosome from the archeon *Thermoplasma acidophilum* (Nitsch *et al.*, 1998), the tricorn protease (Walz *et al.*, 1997a) (Fig. 2), VAT, a CDC48/p97 ATPase homolog in *T. acidophilum* (Rockel *et al.*, 1999), nuclear pore architecture (Beck *et al.*, 2004; Stoffler *et al.*, 2003) (Fig. 3), individual *Escherichia coli* 30S and 50S ribosomal subunits (Zhao *et al.*, 2004), and immunoglobulins (Bongini *et al.*, 2004; Sandin *et al.*, 2004).

Secondly, cryotomography can be applied on isolated similar, but not identical, macromolecular structures or small organelles that are sufficiently thin for direct observation with TEM. With these samples, inherent variability is expected, and attention is paid to those features that are similar. Cryotomography can show the overall 3D morphology of these structures, and higher resolution information can possibly be attained by averaging over multiple repetitive features on or within these organelles or structures. Technical issues that can play an important role in these applications are careful isolation methods and the approach to locate the structure of interest on the electron microscope grid. Examples are parts of the yeast spindle pole body (Bullitt *et al.*, 1997) (Fig. 4), chromatin fibers composed of a continuously variable zig–zag nucleosomal ribbon (Horowitz *et al.*, 1997), axonemes (McEwen *et al.*, 2002), isolated triad junctions

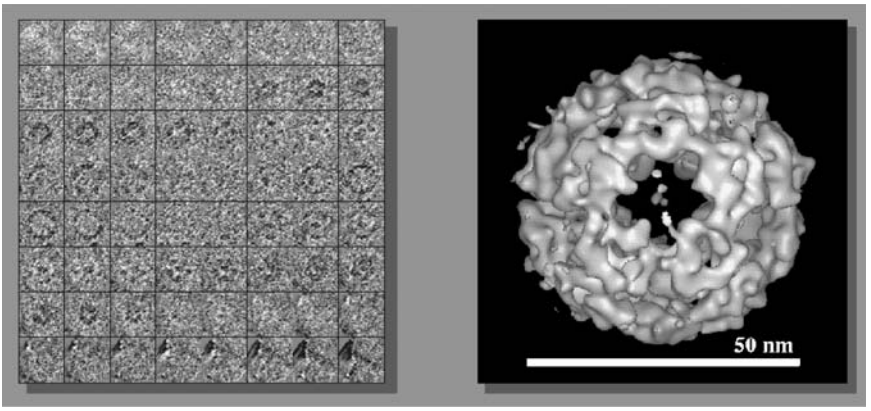


FIGURE 2. The effect of averaging subtomograms. Electron tomography on isolated ice-embedded tricorn capsids of 14.6 MDa (from the archaebacterium *Thermoplasma acidophilum*). Electron tomography was carried out on seven particles, and each series is composed of 25 images (–60° with 5° intervals. The magnification corresponded to 0.5 nm per pixel, and the total dose for data collection to 30 e/nm^2. In the tomogram, 4.2 nm resolution (FSC) was obtained. Slices through the subtomogram of one individual particle (left) and an isosurface display after averaging seven particles. The 5-fold axis of the particle becomes visible. The 4.2 nm tomogram was used as the starting reference for a refinement procedure incorporating 2D images of the particle to attain 1.3 nm resolution (Walz *et al.* 1999). (From Walz *et al.* (1997a), reproduced with permission of Cell Press).

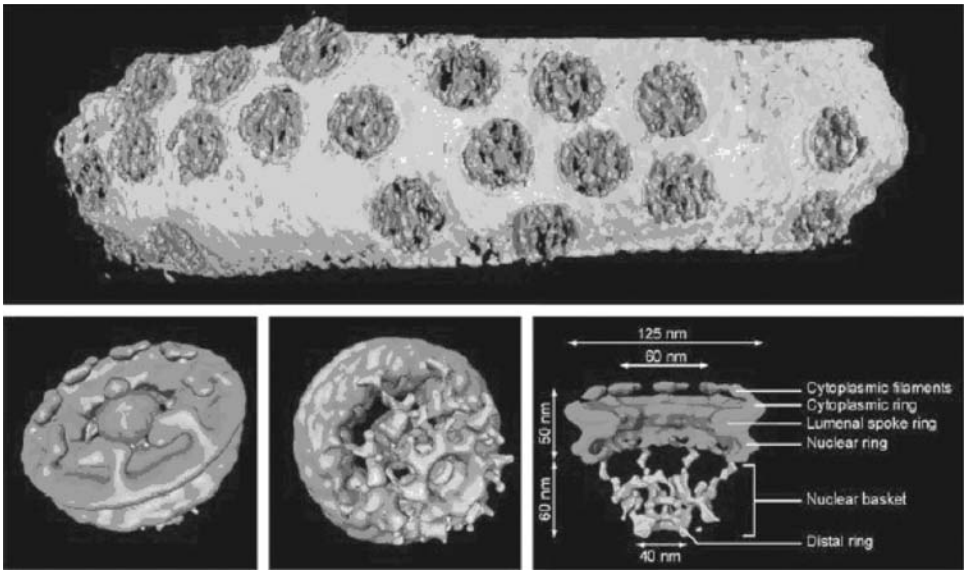


FIGURE 3. Nuclear pore complex. Isosurface representation of a tomogram showing the nuclear envelope. Individual nuclear pore complex (purple) and membrane (yellow) are segmented manually. Bottom left and center: different views of the averaged nuclear pore complex structure Bottom right: overview of the structural elements of the nuclear pore complex. (From Beck *et al.* (2004), reproduced with permission of AAAS).

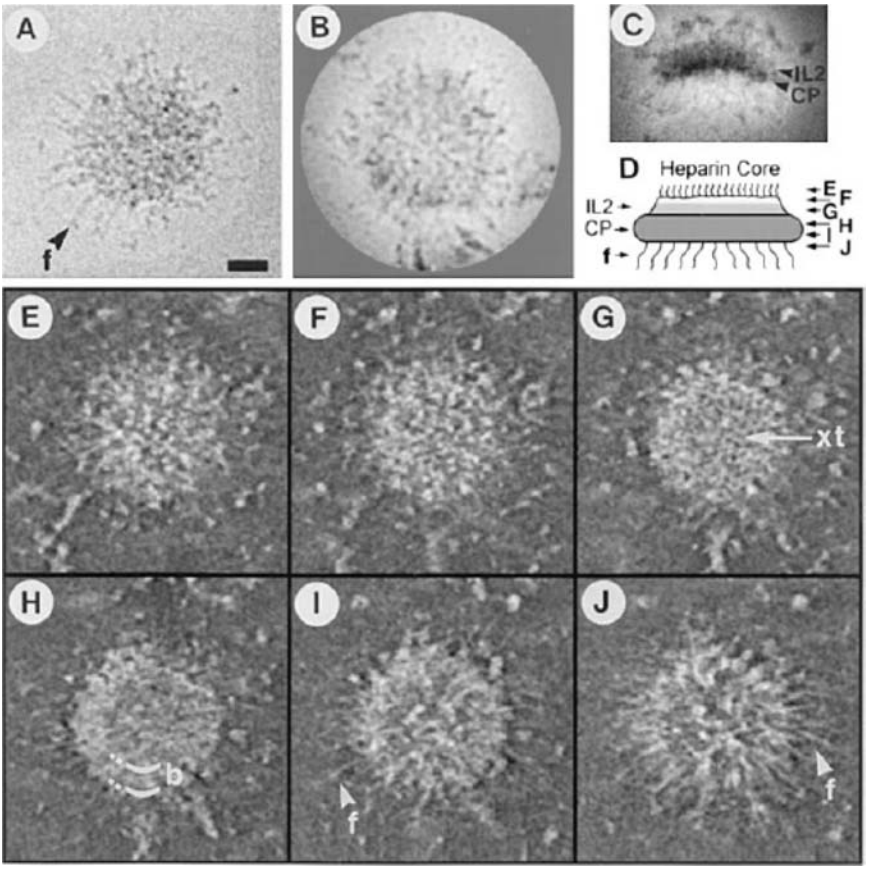


FIGURE 4. Images and tomographic reconstruction of the spindle pole body (SPB) core produced by heparin treatment. (A and B) Face-on views of the frozen-hydrated and negatively stained SPB cores (protein is black). f = filaments. Scale bar = 50 nm. (C) Edge-on view of a negatively stained SPB core (protein is black). (D) Schematic diagram of the metastable SPB core, with approximate vertical positions of the slices shown in (E–J) indicated. (E–J) Successive slices through the tomographic 3D map of frozen-hydrated SPB core (protein is white). The SPB core is viewed face-on, beginning at the cytoplasmic side and progressing toward the nucleoplasmic side, with 14 nm between slices. (E and F) A disordered region above second intermediate layer (IL2). (G) An ordered punctate appearance is visible in the central region of this slice (xt). (H) A ring of protein is visible forming a border (b) located at the base of the hexagonal IL2 crystal and at the top of the cential plague (CP). (I) Region in the CP in which filamentous proteins are visible (f). (J) Filamentous proteins (f) are seen emanating from the bottom of the SPB core. (From Bullitt *et al.* (1997), reproduced with permission of Cell Press.)

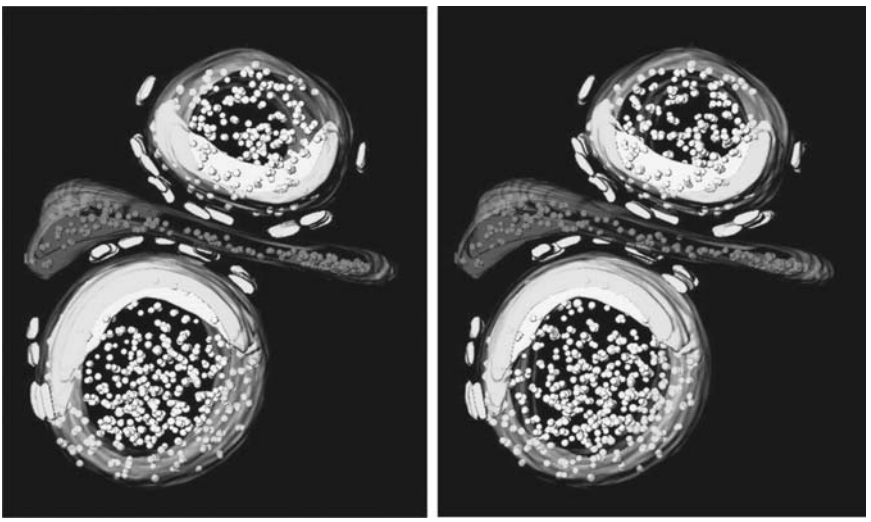


FIGURE 5. Isolated triad junctions. Stereo-pairs of a surface-rendered representation of a tomographic reconstruction of a triad junction. Sarcoplasmic reticulum- (SR) and transverse tubule-derived vesicles are shown in green and red, respectively. Yellow spheres in the lumen of the SR represent calsequestrin, and continuous yellow slabs near junctional surfaces of the SR represent the condensed calsequestrin. Blue structures correspond to feet/ryanodine receptors. (From Wagenknecht *et al.* (2002), reproduced with permission of the Biophysical Society). (See color plate)

(Wagenknecht *et al.*, 2002) (Fig. 5), the nuclear pre-mRNA processing machine (Medalia *et al.*, 2002a), herpes simplex virus (Grunewald *et al.*, 2003), human immunodeficiency virus (HIV; Benjamin *et al.*, 2005) or a retrovirus envelope glycoprotein (Beck *et al.*, 2004; Forster *et al.*, 2005).

Thirdly, cryotomography can be applied to study parts of the cellular architecture, or macromolecular arrangements in their cellular context, in bacteria, large viruses, small cells or regions of cells, that are sufficiently thin for direct observation with TEM. These applications are of particular interest considering that cells are not merely collections of individual molecules and that functions are performed by ensembles of molecules which constitute functional modules. Whereas stable complexes (e.g. the ribosome) present in the cell at all times can be isolated for detailed studies, other functional modules are transient and will escape biochemical detection and, as a consequence, characterization. Ideally, such molecular ensembles should be studied where they act, i.e. in their cellular context. With the application of cryotomography, the cell arrangement and macromolecular organization of vitrified cells can be studied. Though the potential of whole-cell imaging is very large, since the TEM can only image structures with sufficient resolution up to 1000 nm thickness, only relatively small cells or bacteria (<500–1000 nm in diameter) can be investigated with tomography.

Nevertheless, these specimens are relatively thick for TEM, so the application of a high voltage (e.g. 300 or 400 kV) together with an imaging energy filter often plays an important role. In addition, to compensate partly for the limited resolution, extensive use of visualization and computational modeling is applied. For instance, visualization of the actin cytoskeleton in an unperturbed, close-to-life state was achieved with the application of cryotomography to vitrified eukaryotic cells of *Dictyostelium discoideum* (Medalia *et al.*, 2002b). For this study, cells grown on electron microscopy grids were sufficiently thin and flat to be examined at least locally by electron tomography. 3D reconstructions of these cells provided snapshots of the dynamic actin filament network. Other examples are actin and vesicles (Grimm *et al.*, 1997), prokaryotic cells (Grimm *et al.*, 1998), the protein import pore of the outer membrane of mitochondria (Ahting *et al.*, 1999), *Neurospora* mitochondria (Nicastro *et al.*, 2000), rat liver mitochondria (Mannella, 2006; Mannella *et al.*, 2001), phage genome transfer into liposomes (Bohm *et al.*, 2001), T4 bacteriophage (Messaoudi *et al.*, 2003), archeon *Pyrodictium cannulae* entering the periplasmic space (Nickell *et al.*, 2003), vaccinia virus (Cyrklaff *et al.*, 2005), the cytoskeletal structure of *Spiroplasma melliferum* (Kurner *et al.*, 2005), the organization of the core proteins of the yeast spindle pole body (Muller *et al.*, 2005) and molecular complexes in the synaptic cleft (Lucic *et al.*, 2005).

Fourthly, for even larger specimens such as the majority of cells and tissue, a type of application that has not matured yet but has great potential is cryotomography on sections. For these applications, the cryofixation and subsequent sectioning of frozen-hydrated material is limiting due to the technical hurdles in specimen preparation. Pioneering work on this area has been carried out on tissue sections of rat liver where internal features within mitochondria could be seen (Hsieh *et al.*, 2002), and on sections of *E. coli* and *Pseudomonas aeruginosa* (Matias *et al.*, 2003). For an extensive discussion on sectioning frozen-hydrated material, see Chapter 2 of this volume.

2. SPECIMEN PREPARATION

2.1. Cryofixation

The specimen preparation step is of paramount importance as it defines how useful the outcome of the electron tomography efforts will be. Cryofixation has the virtue that the native structure of the sample is preserved by vitrification of the water content. The vitrification process enables the structure of macromolecules and the cellular architecture to be studied in a frozen-hydrated near native state, without considering possible specimen preparation artifacts that may be related to chemical fixation and/or staining.

Small objects, such as isolated macromolecules in a suspension, organelles, or small cells that have dimensions less than a few micrometers, can be cryofixed using a plunger (Adrian *et al.*, 1990; Cyrklaff *et al.*, 1990;

Dubochet *et al.*, 1988; Lepault *et al.*, 1983). A plunger is a device that allows the rapid freezing of the sample by quickly thrusting the grid into liquid ethane (or propane) at liquid nitrogen temperature. Several types of plungers, for instance automated, temperature-controlled and humidity-controlled, are commercially available that can make this step in the procedure quite reproducible and controllable. The speed of vitrification is quite high, in the millisecond range, which can be important for interpretation of results in performing time-resolved experiments (Chestnut *et al.*, 1992; Heymann *et al.*, 2003; Lepault *et al.*, 1991; Steven *et al.*, 2005; Talmon *et al.*, 1990; White *et al.*, 2003).

For thicker samples, the plunging technique will not be very successful because of physical reasons that prevent successful vitrification (Dubochet and Sartori Blanc, 2001). Ice crystals would form that preclude high-resolution interpretation. For samples thicker than a few micrometers, such as large organelles, small cells or parts of tissue, a high-pressure freezing device is the instrument of choice. With high-pressure freezing, the pressure in the vicinity of the sample is rapidly raised and the temperature simultaneously lowered. As a result, the sample will vitrify and the content of the sample, the cellular architecture or network of macromolecular connections, will be fixed, fully frozen-hydrated.

One can ask the question of why high-pressure freezing is not the specimen preparation step used by all of those interested in high-resolution morphological information using electron microscopy. There are several reasons for this. The sample might simply not be suitable for cryofixation as the applicability of the technique depends heavily on the size of the sample and on its components. The speed and quality of vitrification depend on the constituents of the sample (e.g. high sugar content). Although better vitrification can be obtained by adding cryoprotectants (e.g. dextran), these substances may influence the native ultrastructure of the sample and compromise the interpretation of the final outcome of the experiment. In practice, high-pressure freezing is still an approach that requires considerable technical skills and understanding of the procedure. High-pressure freezing works quite reliably for objects less than ~300 μm in size. For larger specimens, no techniques are available to vitrify them successfully (Studer *et al.*, 1989).

2.2. Sectioning Frozen-hydrated Material

Most specimens prepared for TEM such as cells or tissue will be several micrometers thick and not suitable for direct imaging with TEM. An approach to overcome this limitation is to slice these samples in sufficiently thin sections using cryoultramicrotomy (Dubochet *et al.*, 1988; Sitte, 1996). Though sectioning frozen-hydrated material is a technique not strongly defined as yet, results show that frozen-hydrated sections can be produced reliably to a thickness of ~100 nm. For an extensive discussion on section-

ing frozen-hydrated material, see Chapter 2 of this volume. Tools and procedures to produce very thin sections (20–50 nm) of frozen-hydrated material have been shown to be highly reliable and can contribute greatly to provide answers to many structural questions (Zhang *et al.*, 2004). In sectioning frozen-hydrated material to produce thicker sections (thicker than 100 nm), several cutting artifacts may occur, notably compression and crevasses (Al-Amoudi *et al.*, 2004, 2005; Frederik *et al.*, 1982; Hsieh *et al.*, 2002; Zhang *et al.*, 2004). An investigation involving a tomographic reconstruction showed that knife marks were 10–40 nm deep and located on the 'knife face' of the section, while crevasses were 20–50 nm deep and found on the 'block face' (Hsieh *et al.*, 2002). In the experiments reported, a 15-nm banding pattern perpendicular to the cutting direction was observed in the interior of the section, probably associated with section compression. The banding was most evident in the uniformly dense, protein-rich material of the mitochondrial matrix. These artifacts could be minimized, if not totally avoided, by future developments. In this direction, the most remarkable advance has been the design of a new oscillating diamond knife that greatly reduces the cutting artifacts (Al-Amoudi *et al.*, 2003; Studer and Gnaegi, 2000).

2.3. Freeze-Substitution

Considering the fact that sectioning frozen-hydrated material is by no means a matured technique yet, large-scale morphological studies that would involve sectioning frozen-hydrated material cannot be routinely performed. For those applications, an alternative approach for specimen preparation could be cryofixation followed by freeze-substitution (Erk *et al.*, 1998; Humbel *et al.*, 1991; von Schack *et al.*, 1993). With freeze-substitution, the water in the cryofixed block is replaced by organic solvents in successive steps and at low temperature. The outcome of the treatment is a plastic block of material with reasonably well preserved ultrastructure. Of these plastic blocks, sufficiently thin sections can be produced using ultramicrotomy and imaged with TEM. Moreover, as a result of staining, membranous structures can clearly be visualized with electron tomography. Note, however, that ultrastructural interpretation has to be carried out with caution because of possible variations within the sample due to fixation, staining, embedding and shrinkage during data collection (Braunfeld *et al.*, 1994). Although in this chapter we will restrict ourselves to electron tomography on frozen-hydrated specimens, structural work using freeze-substituted material can be very powerful (e.g. to study desmosomes (He *et al.*, 2003), cisternae at different levels of the Golgi complex (Marsh *et al.*, 2004), activated contracting insect flight muscle (Taylor *et al.*, 1999), centrosomes from early *Drosophila* embryos (Moritz *et al.*, 1995) and multivesicular bodies (McIntosh *et al.*, 2005; Murk *et al.*, 2003).

2.4. Whole Cells

A first attempt to prepare whole cells for cryotomography was carried out by plunge-freezing the archeon *Sulfolobus* (Grimm *et al.*, 1998). Tilt series were recorded (50–140 images) with an energy-filtering electron microscope at 120 kV. The embedded cells were between 300 and 750 nm thick. The resolution allowed the identification of the cytoplasmic membrane (4 nm in width), the S-layer (10 nm width) and the space in between, the (quasi-) periplasmic space (20 nm wide). Later, in a similar approach, cryotomography was carried out on cells grown on electron microscope grids where the actin filamentous arrangement in a eukaryotic cell, *D. discoideum*, was studied (Kurner *et al.*, 2004; Medalia *et al.*, 2002b). The peripheral of these cells was sufficiently thin to be imaged with cryotomography. The cells were incubated on the carbon-coated copper grids for 30–60 min. Coating on both sides of the grid proved to be essential. For data collection, a 300 kV TEM was used (with energy filter) and data were collected from –60 to + 60° with 1° increments. The nominal defocus was 15 μm, the pixel size 0.85 nm and the resolution ~5.5. nm. Tomograms of these cells provided snapshots of the dynamic actin filament network (Fig. 6), which occupies 6% of the volume of the cortical region. It was shown to consist of two types of

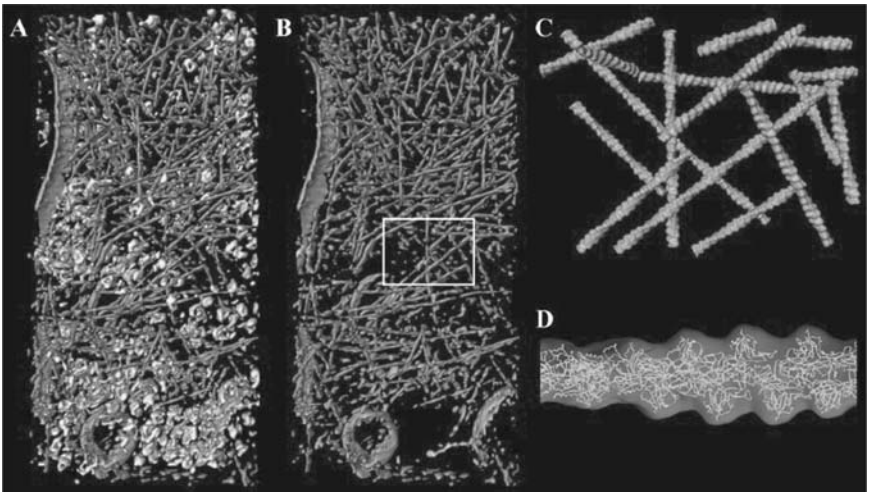


FIGURE 6. Cytoskeletal network of a *Dictyostelium discoideum* cell. (A) Three-dimensional volume rendering representation of a cortical region of the cytoplasm (400 × 870 × 97 nm) showing the actin filament network (reddish), membranes (blue) and cytoplasmic macromolecular complexes (green), many of them resembling 80S ribosomes in size and shape. (B) Visualization of the network of actin filaments. (C) Idealized representation of the actin network region marked in (B) at a resolution of 2 nm using the 3D density map obtained from the atomic model of an actin filament. (D) Docking of the atomic model of actin into the density map of an actin filament. (Panel A adapted from Medalia *et al.* (2002b); panels B–D reproduced from Kurner *et al.* (2002) with permission by Elsevier).

arrays: almost isotropic networks and parallel arrangements of actin filaments that are linked together by bundling proteins. The tomograms showed the interaction between individual microfilaments, their branching angles and membrane attachment sites.

2.5. The Fiducial Markers

Once the specimen is inserted in the TEM, a tilt series of the specimen can be collected. As the holder is tilted, the specimen will move and focus will change due to mechanical imperfections of the goniometer and the specimen holder. Although automated data collection is able to compensate for the major movements, accurate alignment of the set of 2D images to a common origin has to be performed after data collection.

In essence, there are two approaches to align a tilt series. One is based on using high-density markers, such as gold beads 5–10 nm in diameter, which are included in or on the specimen (Frank *et al.*, 1996; Fung *et al.*, 1996; Kremer *et al.*, 1996; Nickell *et al.*, 2005). The other approach is based on using the structural elements within the sample to align one image with the other (Brandt *et al.*, 2001; Liu *et al.*, 1995; Owen and Landis, 1996). This latter strategy has the advantage that no additional preparative steps have to be taken. Often the alignment of the tilt series based on structural elements within the sample is tightly connected to reconstruction scheme in an iterative manner.

However, in some situations, where the sample is relatively thick and the image contrast is not sufficient for alignment, the alternative method based on fiducial markers can be more effective. The presence of a sufficient amount of high-contrast markers within the field of view together with the structure of interest may be crucial for the required image processing (alignment and reconstruction). When very low dose images are collected, the SNR of each individual image in the tilt series can be so low that the images are not suitable for accurate alignment based only on the presence of structural elements within the sample. High-contrast markers such as gold beads within the sample can be used for alignment of the data set (Fig. 7). In most cases, 5–10 fiducial markers within the field of view should be sufficient to allow for alignment of the data set and the subsequent reconstruction. In addition, careful analysis of the fiducial markers throughout the data set can also be used to measure (and correct) resolution-limiting effects due to magnification change and image rotation that might be present in the tilt series.

3. OPTIMIZATION OF ELECTRON MICROSCOPY

3.1. Specimen Thickness

For electron tomography, the specimen has to be sufficiently thin to provide images that are projection images of the sample. If the sample is

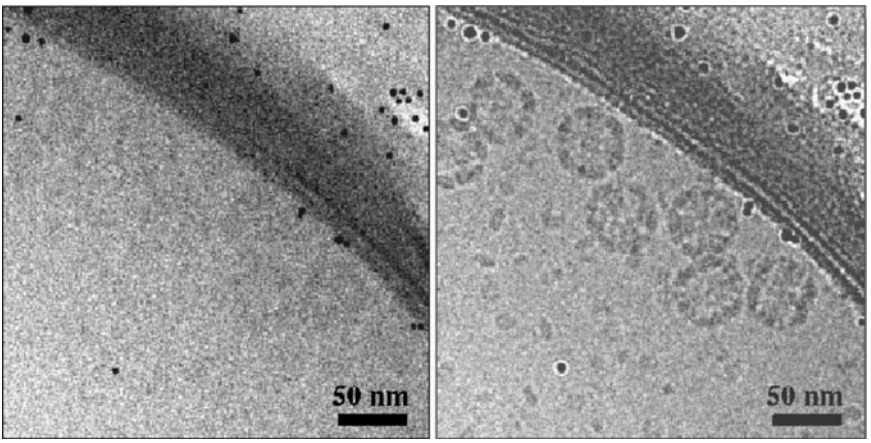


FIGURE 7. The effect of the defocus on the appearance of images for cryotomography. Images of ice-embedded particles (tricorn capsid of *Thermoplasma acidophilum*) at 0.5 μm (left) and 5 μm (right) underfocus. The spherical shaped particles are clearly visible at large defocus. (From Walz *et al.* (1997a), reproduced with permission of Cell Press).

too thick, too many scattering events of electrons within the sample will result in images that are not true projections of the sample. Since electron tomography is a technique aimed at exploiting the third dimension (the thickness) of a sample, most users will investigate a sample as thick as possible. However, it has to be taken into account that at high tilt, the specimen will appear to be thicker compared with the untilted situation. The increase in thickness depends on the specimen tilt angle, α, i.e. the thickness increases with $1/\cos(\alpha)$. This means that, for example, at 60° tilt, a 100 nm thick specimen will appear to be 200 nm thick. Therefore, the increase in thickness when tilting the specimen is a factor that cannot be neglected in practice.

Several factors determine the allowable specimen thickness. For an extensive description, we refer the reader to an overview paper (Koster *et al.*, 1997). When the electron beam penetrates the specimen, three types of events can take place (Fig. 7). (i) When the specimen is thin enough, the majority of the electrons will transfer through the sample as if the sample was not present (*unscattered* electrons). (ii) A part of the fraction of electrons that scatter within the specimen experiences an energy loss in the process (*inelastic scattering*). These inelastically scattered electrons generate a blurry background. (iii) A part of the fraction of electrons that scatter will do so with negligible energy loss (*elastically scattered* electrons). Contrast formation with these elastically scattered electrons is described by the phase contrast imaging theory (Burge *et al.*, 1977; Patwardhan, 2003; Toyoshima and Unwin, 1988; Zhu *et al.*, 1997), since contrast is determined

by phase shifts between the unscattered and elastically scattered electrons. This fraction of elastically scattered electrons provides high-resolution information. Past a certain—allowable—thickness, this contribution of elastically scattered electrons to the image contrast becomes small, and high-resolution contrast will be hidden by the (background) contribution of inelastically scattered electrons.

The experimenter has some means to maximize the allowable thickness. For thick specimens, in the size range of 200–500 nm, the use of medium high voltage TEM (e.g. 300 kV) may be a prerequisite. Medium to high voltage increases the penetration power of the electron beam and reduces the deteriorating influence of multiple scattering. The higher the voltage, the smaller the fraction of the electron beam that will lose energy. The fraction of electrons that loses energy is described by the mean free path, and is determined by both the high voltage of the microscope and the composition of the specimen (Feja and Aebi, 1999; Grimm *et al.*, 1996). An indication of the relative contribution of inelastically scattered electrons to the image can be obtained by electron energy loss spectroscopy (EELS; Leapman and Sun, 1995; Somlyo and Shuman, 1982). Although outside the scope of this chapter, it is worth noting that those electrons that lose energy can be very effective in characterizing the elemental composition of the specimen and consequently give 3D information on the distribution of specific elements such as iron in ferritin (Zhang *et al.*, 2005). The electrons that do not lose energy contribute to the zero-loss peak. For thin specimens, the zero-loss peak is large compared with the overall spectrum. However, for very thick specimens, the zero-loss peak can become insignificant (Bouwer *et al.*, 2004). Though one might reason that very high voltage electron microscopy would enable the investigation of even thicker specimens, there is a limit to the effectiveness of a high acceleration voltage for frozen-hydrated specimen. It was shown that the high voltage enhances the radiation damage due to knock-on events, i.e. inelastic scattering events, in which the energy transferred to an atom is greater than its binding energy, and that the possible gain in penetration power is limited (Grimm *et al.*, 1998). Increasing the accelerating voltage from 100 to 300 kV gives an increase in the penetration power with a factor of 2, while the additional gain by a further increase from 300 kV to 1.2 MV is only a factor of 1.5. An accelerating voltage of 300–400 kV thus seems to be a practical compromise.

A piece of instrumentation that will enhance the image contrast in electron tomography is an imaging energy filter (Grimm *et al.*, 1997; Marco *et al.*, 2004). Using zero-loss energy filtering, only those electrons that did not lose energy will contribute to the image contrast. As a result, the low-resolution contribution (blur) of those electrons that lost energy within the sample is removed and the image will be clearer. An alternative way to remove the disturbing image contrast of inelastically scattered electrons is to apply image restoration techniques. With image restoration, a series of images taken at different defocus are combined into one image (Coene *et al.*, 1992), the

contribution of inelastically scattered electron is removed and the image contrast will improve (Han *et al.*, 1995, 1997).

Nevertheless, there is a limit to the extent energy filtering or image restoration can help in the improvement. The techniques cannot overcome the fundamental limitation that the fraction of elastically scattered electrons will become less and less when the specimen gets thicker. At some thickness, the allowable thickness, the remaining fraction of (single) elastically scattered electrons will be extremely small and the tiny amount of high-resolution image contrast will be obscured by the (counting) noise in the image recorded at very low intensity. Current developments aimed at developing spherical aberration and chromatic aberration correctors for TEM imaging are attachments that do not intend to remove electrons, but to make them contribute to the high-resolution information in the image (Freitag *et al.*, 2005; Rose, 2005; Sawada *et al.*, 2005). Potentially, these attachments would make it possible to image sections thicker than several micrometers. An unsolved fundamental limitation is that the image contrast is reduced due to multiple elastic scattering events, which cannot be remedied by energy filters or aberration correction. To minimize the influence of this effect, the use of a small objective aperture that blocks the electrons at high scattering angles will have a positive effect, but will also limit the resolution in the image.

3.2. Allowable Electron Dose

In practice, high-resolution imaging of frozen-hydrated specimen is limited by radiation damage of the specimen since unstained specimens are very sensitive to the electron beam. When the specimen is exposed to an amount exceeding the allowable dose, the specimen will change due to the electron irradiation, and eventually even effects on the ice will become visible (bubbling). Therefore, the electron dose will need to be minimized and it is important to determine how much electron dose the specimen can take before specimen damage occurs at the resolution aimed for. This dose, the allowable total electron dose, has to be distributed over the number of images that will be taken with the microscope (Hegerl and Hoppe, 1976; McEwen *et al.*, 1995). In most cases, the allowable electron dose will be the limiting factor for cryotomography.

An upper limit to the number of projections over which the dose can be fractionated is given by the requirement that the 2D images in the tilt series have to be aligned to a common origin before a 3D reconstruction can be computed. The dose fractionation will reach its limit when the individual 2D images within the data set cannot be aligned accurately because the SNR becomes too low. In this regard, the presence of high-density markers (e.g. 5 or 10 nm gold beads) in the sample can be of great help.

There is still another practical limitation to dose fractionation: sufficient electrons must be recorded for each image to give a statistically significant image on the recording device (CCD camera) (Grimm *et al.*, 1998). These authors stated that, although it will be possible to obtain resolutions approaching 2 nm for objects that are <100 nm in thickness because small objects do not require such a fine angular sampling, specimens thicker than 500 nm will be limited to 6–10 nm resolution. Later it was pointed out (McEwen *et al.*, 2002) that it was not taken into account that as the desired resolution increases, so does the total electron dose required to ensure statistical significance of each volume element at the resolution limit. This requirement is independent of the size of the object or the recording media used. As stated, a poor recording medium can impose further limits to resolution, but a superior medium cannot overcome the statistical nature of the electron beam. McEwen and colleagues deduce that it will be difficult to push resolution past 5 nm without severe irradiation damage to the specimen.

The allowable dose differs depending on the combination of specimen and the desired structural details which are hoped to be revealed. For example, cryotomography of isolated triad junctions (Fig. 5) at 200 kV was carried out with a total dose of 5000–8000 e/nm^2 to attain a resolution of 5–6 nm (Wagenknecht *et al.*, 2002). Cryotomography of vaccinia virus particles with a 300 kV TEM was performed with a total dose of 2500–3500 e/nm^2 and the resolution achieved was ~5 nm (Cyrklaff *et al.*, 2005). These examples indicate the resolution and range of electron dose that can be allowed to image vitrified samples. Depending on the type of question, procedures that include averaging of substructures within the tomogram could be used to obtain higher resolution information (Forster *et al.*, 2005).

3.3. Image Detector

In most cases, a digital image recording device, such as a cooled CCD camera, is attached to the TEM to perform automated data collection. Digital image recording is required to enable automated compensation for image shifts and defocus changes that occur during the tilt series (Dierksen *et al.*, 1992). Concerning the camera, two parameters need to be taken into account to optimize data collection: the sensitivity and the resolution of the device. Since a CCD camera has considerably less resolution than photographic film, choosing a sufficiently high magnification is a critical factor for cryotomography. When the magnification selected is too low, the image will not show sufficient detail. On the other hand, one major difference between CCD cameras and exposure on film is that the field of view is by far greater for photographic film, even in comparison with a 4000 × 4000 pixel camera (Typke *et al.*, 2005). Therefore, when the magnification selected is too high, sufficient image detail can be retrieved, but the overall

area that is imaged will be very small. Therefore, it is important to select a compromise magnification that will show sufficient image detail for the specific question at hand.

In CCD cameras for TEM, the electrons are converted into photons by means of a scintillation screen at the entrance of the camera. The light is guided via fiber optic elements, or lenses, to the CCD chip. Thus, the distribution of the electron intensity impinging on the entrance of the camera is transferred to a distribution of photons going through the optical elements to the CCD chip. This conversion from an electron distribution to an image intensity distribution can be described by a point-spread function. To date, the point-spread function will be sharper (better) for a microscope operating at low voltage (e.g. 100 kV) than at high voltage (e.g. 300 kV). The sensitivity will be higher when the phosphor layer (or YAG screen) on top of the CCD chip is chosen to be somewhat thicker. However, a thicker layer will result in less resolution. Most CCD cameras for electron microscopy are equipped with large-area chips, containing 1024^2–4096^2 or more pixels with a pixel size in the range of 15–30 μm. The CCD chip in the camera housing is cooled, for instance to –36°C, to reduce electronic noise in the images taken.

To have an idea of the compromises to be made, suppose we would like to obtain a tomogram of a small organelle 200 nm in diameter with a resolution of 2 nm. To allow interpretation of structures to a given resolution, in most cases at least three resolution elements are needed, in this example ~0.7 nm. A resolution element in an image is in most cases not the same as the CCD pixel, but instead is determined by the point-spread function of the CCD camera (which is, in turn, mostly determined by the phosphor layer on top of the CCD chip). Suppose we have a 200 kV TEM at our disposal with a CCD camera outfitted with a phosphor layer of 20 μm thickness. For such a setup, the resolvable image detail that can be obtained will be in the order of 30 μm (Downing and Hendrickson, 1999; Koster *et al.*, 1992; Sherman *et al.*, 1996; Zhang *et al.*, 2003; Zuo, 2000). For a 300 kV TEM, the scintillator will be thicker and the smallest independent resolution element will be ~50 μm. Optimally, the pixel size of our available microscope and camera (30 μm, or 30,000 nm) would exactly match the resolution element we aim for in imaging the specimen (0.7 nm). Therefore, the magnification has to be 30,000 nm divided by 0.7 nm, which equals ~42,000. Note that this is the magnification from the specimen to the image detector (here the CCD camera). Depending on the specific instrumental setup, the magnification factor will differ between the different detectors (e.g. a camera mounted at the end of an energy filter, or a camera at a 35 mm port much closer to the specimen).

So far we assumed that the 30 μm spot generated by the electrons in the scintillator matched the CCD pixel size. However, this is typically not the case. Some cameras are equipped with 2048^2 pixels of 14 μm. In spite of the fact that the camera has more, and smaller, pixels, the images will contain

only ~1024^2 resolvable details. To adapt the camera settings to the highest resolution image content, most designs of the readout electronics of the CCD chip are capable of transferring the data from the chip in different ways. For instance, the electronics can transfer the content of the 2048^2 pixels one at a time. Another way would be to transfer packages of pixels at a time, for instance groups of 2×2 pixels (binning of 2). With these settings, the field of view is ~700^2 nm^2 (1024^2 pixels) with a pixel size of 0.7 nm (28 μm).

In choosing the camera settings and illumination intensity, it is important to consider the sensitivity of the camera. Suppose the electron dose per exposure of the specimen is 100 e/nm^2 (1 e/$Å^2$). Let us estimate if we can see this image clearly using a CCD camera. As discussed above, each pixel of the image gives a number (counts) that is linearly related to the impinging electron intensity. The linear factor, generally indicated by ADU (analog to digital units), is one of the specifications/characteristics of the CCD and indicates its sensitivity.

Let us suppose now that we use a camera optimized for low-dose electron microscopy that has large pixels and is cooled to reduce the inherent noise of the chip. With a CCD camera having 28 μm pixels at a magnification of 42,000, one pixel corresponds to 0.7^2 nm^2. Therefore, the number of electrons per pixel equals 100 e/nm^2 times 0.7^2 nm^2 which corresponds to 49 electrons per pixel. The sensitivity can be 9 ADU per electron (for a particular high voltage and a particular type of scintillator). At very low close levels, the noise produced within the CCD chip may become stronger than the signal produced by the electron's image intensity. In most cases, the specifications of the CCD camera contain an indication of the inherent noise, for instance the standard deviation of the noise can be 2 counts. These 49 electrons will produce an image with a value of ~450 counts, much higher than the noise level.

Let us suppose now that we use a camera not optimized for low-dose applications that has small pixels of 6.5×6.5 μm and is not cooled. The sensitivity of such a camera can be 0.5 ADU per electron (for a particular high voltage and a particular type of scintillator). The inherent noise (standard deviation) can be 4 counts. At 42,000 magnification, one pixel corresponds to ~0.16^2 nm^2. With the same illumination intensity as mentioned above, the number of electrons per pixel equals 100 e/nm^2 times 0.16^2 nm^2 which corresponds to 2.5 electrons per pixel. With a sensitivity of 0.5 ADU per electron, the image will have an average count of ~1. The noise per pixel will be ~4 counts. In this example, for a non-optimized camera for low-dose imaging, the image will be drowned in the noise and will not be usable.

3.4. Optimizing Image Contrast

For a more complete description on image formation and how this impacts electron tomography, see Chapter 3 of this volume.

An important point to realize in discussing image formation is that in order to minimize specimen damage, the total electron dose that is needed to collect the tilt series has to be kept as low as possible (e.g. ~5000 e/nm^2). The interaction between the specimen and the imaging electron beam has to generate suitable image contrast that allows the image intensity distribution to be related to a distribution of mass density within the sample. In practice, under these conditions, weak phase objects can only be visualized by choosing a large defocus, enhancing the relatively low frequencies as described by the phase contrast transfer theory. When imaging thin layers of suspensions containing macromolecules, it may be an advantage to mix the solution with ammonium molybdate. The mass den-sity of the macromolecule is less than that of the ammonium molybdate and will produce high contrast images, often referred to as cryonegative staining (Adriant *et al.*, 1998; De Carlo *et al.*, 2002).

Several parameters can be used to optimize the image contrast. One could select a small objective aperture (e.g. 10 μm). The small aperture will block electrons that are scattered over a large angle and will be effective in blocking multiple elastically scattered electrons. Although a small objective aperture precludes high-resolution imaging, this may be a valid option for relatively low resolution imaging. On the other hand, the aperture chosen cannot be too small. As is shown in Fig. 8, depending on the thickness of the specimen, the relative contribution of elastically scattered electrons comprises the greatest fraction of the electrons that would be filtered out by an aperture, the opposite of what one would want to do. Therefore, only choosing a higher acceleration voltage can help to optimize the contrast at high resolution.

A limitation in selecting too small an objective aperture can be the performance of the specimen holder. For example, if during collection of a tilt series the holder moves laterally over a large distance (e.g. >5 μm), the optics have to be adjusted over that distance to keep the feature of interest centered. It may well be that at some point the optical correction will become less optimal and the edge of the objective aperture will (partly) block the illuminating beam. Nevertheless, in most instruments, the movement of (cryo)holders during a tilt series will be within a range of 2 μm and should not be the cause of serious problems.

To enhance the image phase contrast, the tilt series in cryotomography are often taken at large defocus (>5 μm). To choose the suitable defocus, we need to consider the image formation process as described by the phase contrast transfer theory. In the case of a (weak) phase contrast specimen (i.e. a thin frozen-hydrated sample), the image formation can be mathematically described by the phase contrast transfer function (CTF). Several programs are available to display a phase CTF for a specific type of microscope (defined by a number of TEM specifications, such as the spherical aberration constant, the chromatic aberration constant, and the temporal and spatial coherence of the electron source) and for a particular defocus.

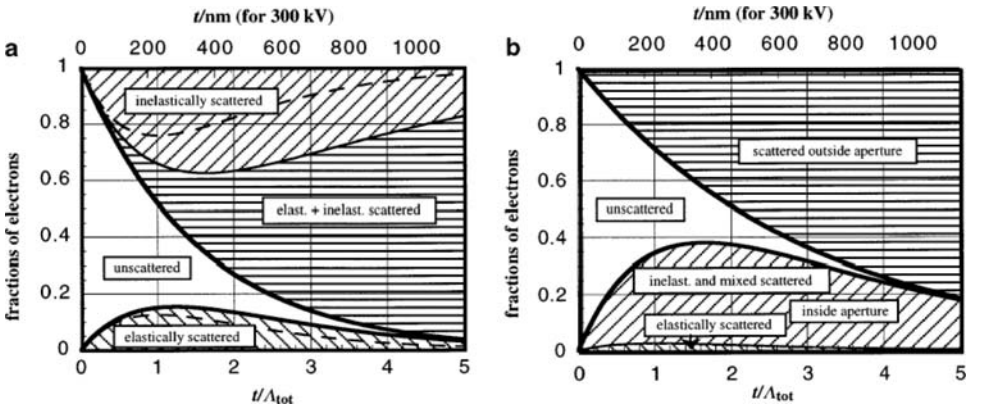


FIGURE 8. Distribution of scattered electrons for vitreous ice. When the electron beam penetrates the specimen, three types of scattering events can take place. When the specimen is thin enough, the majority of the electrons will transfer through the sample as if the sample were not present (*unscattered* electrons). A fraction of electrons that scatter within the specimen experience energy loss (*inelastic scattering*). Another fraction of electrons that scatter will do so with negligible energy loss (*elastically scattered* electrons). (a) Distribution (without an aperture) over the elastic (lower diagonally hatched area), inelastic (upper diagonally hatched area) and mixed (horizontally hatched area) scattering channels for vitreous ice as a function of thickness (in multiples of the total mean free path Λ_{tot} and, for 300 kV, in nm). The dashed lines mark fractions of single (elastic or inelastic) scattering. (b) Distribution of electrons hitting or passing the objective aperture (diameter corresponding to 0.4 nm resolution). (From Koster *et al.* (1997), reproduced with permission of Elsevier).

Figure 9 shows that a transfer function for a Tecnai 20 FEG at 200 kV at a defocus of –3.4 μm has its first zero at 2.8 nm. Because CTF corrections in the noisy images of a cryotilt series are very difficult to carry out, the defocus is often chosen such that the first contrast reversal as described by the CTF corresponds to the highest resolution aimed for. For instance, in the cryotomography experiments on visualizing the actin cytoskeleton, a defocus of 15 μm was chosen to optimize contrast at 5.5 nm resolution with a 300 kV microscope. The interpretation of images at higher resolution is not straightforward because of the contrast reversals which occur. However, novel instrumental developments aimed at compensating for the spherical and chromatic aberration coefficients within the TEM (Freitag *et al.*, 2005) could provide means to facilitate image interpretation at higher spatial frequencies.

Another practical approach can be to optimize the contrast by choosing the defocus to match the resolution of the digital camera. For instance, suppose the pixel size of a 2048×2048 CCD camera is 0.7 nm. As explained in the previous section, because of the imperfect transfer of the scintillator on top of the CCD camera, in many circumstances it will be effective to collect images with 2×2 binned resolution, corresponding to a pixel size of

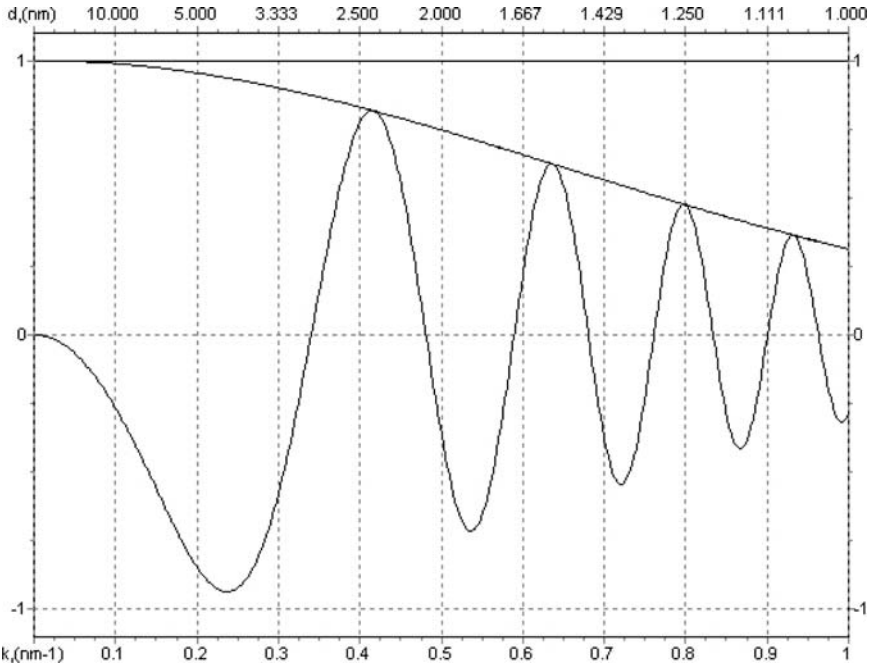


FIGURE 9. Contrast transfer function. Theoretical contrast transfer functions for an FEG instrument at an acceleration voltage of 200 kV. The spatial coherence (α = 0.1 mrad) and temporal coherence (energy width of the incident beam ΔE = 0.7 eV) lead to small attenuation of the contrast transfer function. The defocus value (–3.4 μm) was chosen such that the first zero of the CTF occurs at (2.8 nm)–1.

1.4 nm. Then, following image sampling theory, the best resolution one can ever hope to retrieve will be by sampling with twice the pixel size (i.e. the Nyquist frequency), corresponding to 2.8 nm. Therefore, the limitations of sampling make it sensible to use, at least, a defocus value that positions the first zero of the transfer function at 2.8 nm. Using a lower defocus would only produce an unnecessary reduction of the contrast in our images.

When aiming at achieving better contrast, zero-loss filtering can be of great help for most specimens imaged with electron tomography (Grimm *et al.*, 1996, 1997; Koster *al.*, 1997). As mentioned before, the main purpose of such a filter will be to remove inelastically scattered electrons from the image. Because of the chromatic aberration of the objective lens, these inelastically scattered electrons contribute significantly to a blurry background in the image. Often an energy window of ~10 eV is applied. To determine experimentally whether or not zero-loss filtering may be of help to produce better images for the type of samples investigated, one can image energy loss spectra at different specimen tilt angles (Angert *et al.*, 2000; Grimm *et al.*, 1997; Marco *et al.*, 2004; Midgley and Weyland, 2003;

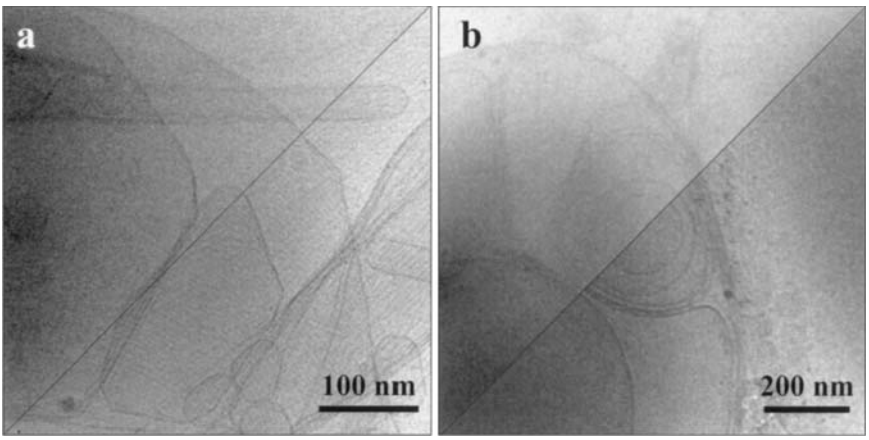


FIGURE 10. Filtered and unfiltered images of ice-embedded specimens. (a) Actin filaments and vesicles (250 nm thick ice). The half of the image below the diagonal is zero-loss filtered, while the upper half is unfiltered. Whereas contrast is comparable in both images, the resolution is improved in the case of the filtered image, such that individual actin filaments can be identified. (b) Multilamellar vesicles (600 nm thick ice). The division of the image is as in (a). Only a few lamella can be distinguished in the unfiltered case. (From Koster *et al.* (1997), reproduced with permission of Elsevier).

Somlyo and Shuman, 1982; Zhu *et al.*, 1997). These experiments showed that the amount of electrons that do not lose energy becomes less at higher specimen tilts (because the path through which the electrons pass in the specimen will be longer at high tilts). While the effect is not readily visible to the eye for thin specimens (less than ~100 nm thickness), the effect of zero-loss filtering becomes clearer for samples in the intermediate thickness range of 100–500 nm. This is illustrated in Fig. 10 with two different specimens. In those image regions where a sufficiently high number of electrons have penetrated the specimen without energy loss, details are observed with higher resolution than in the unfiltered image. Within thicker objects, small details are sometimes more readily seen in the unfiltered image due to the low-resolution information of the inelastically scattered electrons. However, the contrast is better with zero-loss filtering. When the specimen is beyond a certain thickness, very few electrons will pass the specimen without energy loss or multiple scattering, and the images will be very dim and not possible to align. Though high-resolution information will be lacking, it may be an effective, alternative, approach to use the energy-filtering slit to select those electrons in the spectrum at the maximum of the plasmon peak (Bouwer *et al.*, 2004). Thickness variations in thick specimens may produce strong intensity differences in these images, but a CCD camera, due to its large dynamic range, will be able to record such images properly.

As an example of how imaging conditions can be optimized to obtain a cryotomogram of a particular structure, let us look at the data collection parameters as used in studying isolated triad junctions (Wagenknecht *et al.*, 2002). A tilt series (–60 to +60°) was collected at 2° intervals with an electron dose of 50–100 e/nm^2 per image (3000–8000 en/nm^2 total estimated dose per reconstruction). The images were collected using a JEOL JEM4000FX transmission electron microscope operated at 200 kV with objective lens underfocused to 10 μm. At this defocus, the first zero of the CTF is at 5 nm. Both the defocus level and the tilt angle increment limit the best attainable resolution to 5–6 nm. The magnification conditions from specimen to CCD camera corresponded to a pixel size of 1 nm. Alignment of the projections and 3D reconstruction were carried out using the SPIDER package (Frank *et al.*, 1996). A different instrumental set-up was used to collect data on the bacterium *Spiroplasma melliferum*. This structure was studied to ~4 nm resolution (Kurner *et al.*, 2005). Data collection was done with a defocus of 10 μm in a 300 kV microscope. The herpes simplex virus (Grunewald *et al.*, 2003) of ~200 nm in diameter was imaged at a defocus setting of 10 μm required to enhance the low contrast features (5 nm range) that would otherwise be undetectable in a noisy background.

3.5. Challenges Encountered in Cryotomography Data Collection

There are a number of technical issues that can make cryotomography a tedious process. Although ongoing developments make specimen preparation, data collection and image processing increasingly user-friendly, some technical hurdles are still present.

Imaging sectioned frozen-hydrated material for cryotomography is particularly problematic. First, the specimen is often not very level—it can be cracked and different areas on the section might therefore be at different heights (focus) and expose quite significant different orientations (tilted) (Al-Amoudi *et al.*, 2005; Hsieh *et al.*, 2002). In addition, the section is not always well adhered to the specimen support. These two characteristics can make data collection very problematic. For instance, when a conventional low-dose data collection scheme is used, the measured focus change and specimen movements may differ significantly from the movements at the area of exposure. Data collection on these specimens may require significant user interaction to keep the feature of interest centered. Moreover, since it is possible that the specimen is not attached to the specimen support, the distance between specimen and the fiducial markers (deposited on the specimen support for tracking and focusing) may in fact be a cause for inaccurate alignment of the tilt series. Developments on better data collection procedures will need to be developed. Another challenge is related to locating the feature of interest on the specimen within the TEM. Because of the high sensitivity of the specimen to the electron

dose, the search for the area of interest will have to be done at very low magnification (where hardly any details can be seen).

Conventional cryoholders are cooled with a liquid nitrogen dewar mounted on the end of the holder rod outside the microscope column. In older designs, due to the weight of the dewar, the specimen usually exhibits a continuous drift at specimen tilts higher than 45°. This movement is often due to a combination of mechanical sliding of the rod inside the microscope column and a slight change in temperature due to a varying liquid nitrogen level in the dewar. At these higher tilts, the specimen drift rate can become such that the automated procedures in the data collection algorithms have to include features to cope with it. When drift is present, the alignment of the imaging optics will be continuously adapted to compensate for this drift, and the accumulated amount of image shift will get larger and larger as the tilt series collection proceeds. Ultimately, the illumination and imaging alignment will be partly blocked by the (fixed) mechanical aperture of the objective lens. Moreover, when the alignment of the imaging optics is changed considerably, calibrations related to automatic focusing will become less valid and the automatic focusing step within the data collection procedure will be less accurate. Fortunately, most current side-entry cryoholders show a better performance in this regard than in the past (Fig. 11). In addition, novel designs of specimen holders have become available where the specimen is cooled inside the column and that do not exhibit these specimen movement problems.

Double-axis tilting has proven to be a powerful approach to minimize reconstruction artifacts due to the missing wedge in applications using room temperature electron tomography (Mastronarde, 1997; Penczek *et al.*, 1995). In many cases, double-axis tilting has been shown to be almost indispensable for obtaining tomograms of sufficient quality to allow tracking of 3D structures. Therefore, double-axis tilting for cryotomography is highly desirable to maximize the quality and interpretability of tomograms. Novel designs of specimen holders are capable of double-axis tilting at cryogenic temperatures, with a performance at least equal to conventional single-tilt room temperature specimen holders.

3.6. Setting up a Data Collection Experiment

3.6.1. Magnification

As stated in Section 3.3, the choice of magnification is an important parameter as it sets a limit to the highest resolution that can be obtained from the 2D images recorded and the largest field of view that can be observed. These two factors demand a compromise which depends on the specific biological question addressed. Because of the relatively small field of view of slow-scan CCD cameras, often the magnification is chosen to be as small as possible. However, it has to be kept in mind that for a subsequent image analysis steps, the magnification needs to be sufficiently high

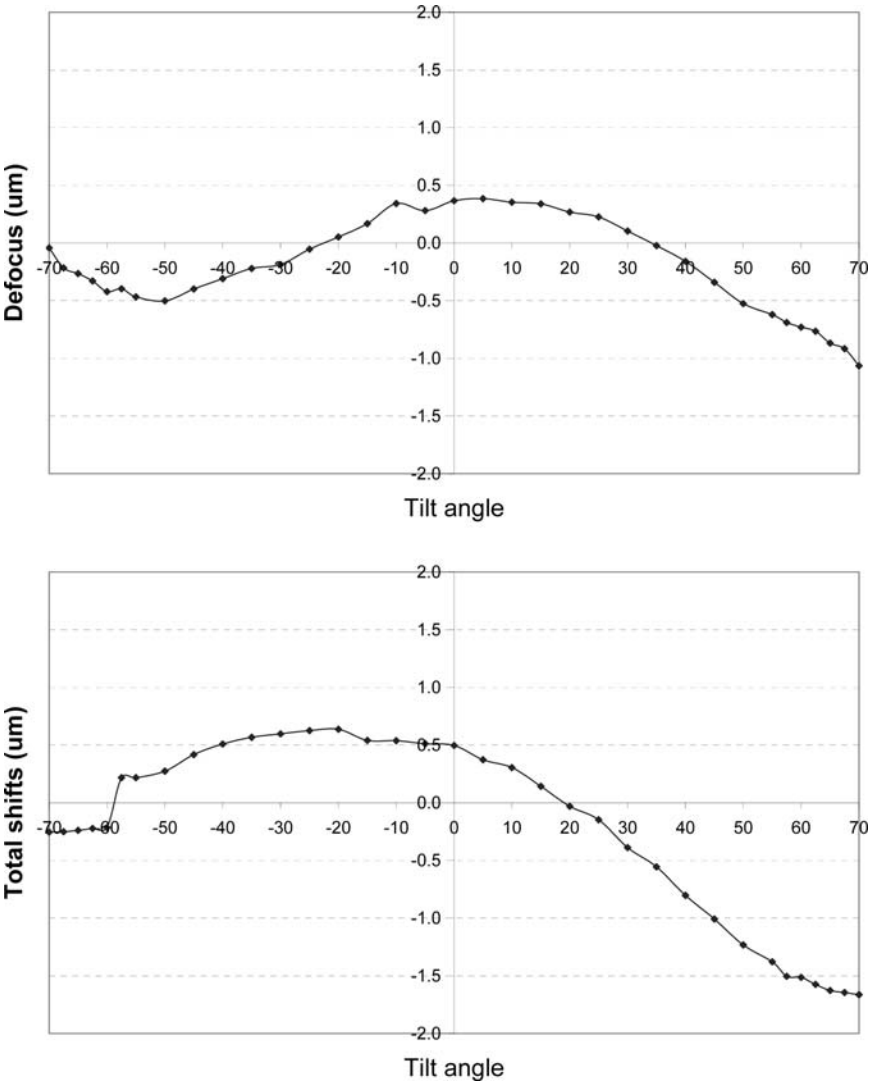


FIGURE. 11. Specimen movement (shifts in *xy*-direction) as a function of specimen tilt measured for a side-entry cryoholder close to the optimum alignment position for tomography. The shape of such a curve depends greatly on the goniometer, holder, eucentric height and alignment of the microscope, but also on possible specimen drift during data collection. In most cases, the total amount of shift in the plane of the specimen (*xy*) will be <2 μm.

to resolve the required image features measuring at least 3 pixels. For example, to resolve two membrane layers with 3 nm space in between, a pixel size of 1 nm would be a good choice. From these requirements on the pixel size together with the specifications of the camera set-up, a lower limit to the magnification can be determined. For a CCD camera with 24 μm (24,000 nm) sized pixels, the magnification required to image 6 nm features (diameter) with 3 pixels would be 24,000 nm divided by 2 nm which equals 12,000. At this magnification, and with a CCD camera that is composed of 1024^2 pixels of 24 μm, the field of view is then 2 μm.

3.6.2. Angular Range and Increment

A large angular tilt range is advantageous to minimize the missing wedge. However, the choice of the angular tilt range and tilt increments is mostly limited by practical circumstances such as the loss of image contrast due to inelastic scattering at high tilts. For example, a 200 nm thick sample at a 70° tilt will be 2.7 times thicker (540 nm). In most cases, a maximum tilt angle of 65° is chosen. Suppose the goal is to image a 200 nm thick section with a resolution in the z-direction of ~6 nm. For linear tilt increments, the resolution d is given by the relationship $d = \pi D/N$ (Crowther *et al.*, 1970; Grimm *et al.*, 1998 Hoppe, 1969), where D is the diameter of the object, and N the number of projections recorded at equally spaced tilt angles over a range of 180° also given by $180/\Delta\alpha$. To attain 6 nm resolution, the number of images we need to acquire is then ~100 ($3 \times 200/6$). The angular sampling than corresponds to 1.8° (180°/100 images). To cover the tilt range of 130°, ~66 images will need to be taken.

For specimens that are not spherical objects, such as isolated large macromolecular objects embedded in a section of vitrified ice, but a section (slab), an alternative data collection geometry will be more suitable. For these samples, where the thickness of the section (D) increases with the specimen tilt angle α following $D/\cos(\alpha)$ (Radermacher, 1992), the angular increments should be smaller to compensate for the increase in thickness to sample the object to the same resolution throughout the tilt series. This idea is the basis for the non-equidistant tilt increments scheme proposed by Saxton *et al.* (1984). The goal of the Saxton data collection scheme is to fill Fourier space more evenly by making the tilt increment smaller at high tilt angles, to compensate for the increase in specimen thickness. In this scheme, the tilt angle varies as $\alpha_{n+1} = \alpha_n + \arcsin(\sin\alpha_0 \cos\alpha_0)$. The number of images in a tomographic series (N) in terms of the initial tilt increment α_0 is then (approximately) given by $N = (2\ (\alpha_{max}/\alpha_0) + 1)\ (1/2\alpha_{max})\ \ln((1 + \sin\alpha_{max})/(1 - \sin\alpha_{max}))$. For a given initial angular increment α_0, the Saxton scheme will require more images than a constant angular increment strategy. For example, for a tilt range of –60°/+60° and an initial increment of 1.5°, there would be 102 images for the Saxton scheme compared to the

81 images of a linear scheme. Clearly, it is also possible to start from a larger initial angular increment to keep N the same or even lower compared with taking a tilt series with a constant increment.

3.6.3. Illumination conditions

For the illumination conditions, we take the allowable dose into account. Suppose we have measured in an independent experiment that the tolerable specimen damage occurs with a dose of ~5000 e/nm^2. From this number, we can compute the electron dose per image. If 61 images were needed, that would mean that the illumination intensity should be such that per exposure not more than 80 e/nm^2 (5000 e/nm^2/61) are used. With a magnification of 16,000 and a pixel size of 1.5 nm, 180 electrons are counted per pixel ($780 \times (1.5)^2$). For many camera systems, this intensity of illumination is sufficiently high for detection. Nevertheless, for much less intense illumination, the electronic readout noise of the camera may obscure the signal, and the individual 2D images will not contain sufficient information for subsequent alignment. In those circumstances, choosing a lower magnification may be a necessity.

In many cases, it can be advantageous to adapt the exposure time dynamically during data collection, maintaining the same total dose (Grimm *et al.*, 1998). The exposure time should be increased with the tilt angle to keep the SNR of the individual 2D images in a tilt series more or less constant. If this is not done, the images at high tilt angles may show an insufficient contrast to allow alignment of the data set. Several schemes for varying the exposure time can be used. An exponential increase in exposure time (t) can be used to maintain the count rate on the camera, and thereby the SNR, constant in all images of an evenly thick specimen. In the case of whole cells in thin ice, another scheme following an inverse cosine proved to be useful: $t = t_0/\cos\alpha$ (Grimm *et al.*, 1996). In practice, limits have to be set to the extent the exposure time is allowed to vary, for instance 100 ms for the shortest exposures time (to acquire sufficient signal for subsequent alignment of the data set) and 3 s for the longest (during which the specimen should not move significantly (e.g. move <0.5 pixel).

3.6.4. Defocus

As discussed above, it could be effective to choose the defocus such that the first zero of the CTF matches the target resolution, e.g. 3 nm, which corresponds to a defocus of ~3.5 μm for a 200 kV instrument (e.g. a Tecnai F20; Fig. 9). When a larger defocus is chosen, or when higher resolution is aimed for, contrast reversals in the image formation will require post-image processing to deconvolute with the appropriate point-spread function (Winkler and Taylor, 2003).

3.6.5. Search

It can be difficult to localize the area of interest on the microscope grid. Because of the low-dose requirements, one has to scan the specimen at very low magnification, and a very large defocus helps to maximize the contrast in the image. A novel development to improve the efficiency in searching an area or feature of interest on an electron microscope grid includes the combination of cryoelectron microscopy with fluorescence imaging. The fluorescence signal could potentially be used to indicate the area of interest for cryotomography (Biel *et al.*, 2003; Braet, 2004).

4. AUTOMATION

4.1. Automation Procedures

The steps involved in collecting a single-axis tilt tomographic data set are in principle straightforward. The specimen is tilted over a large angular range and images are recorded at a series of discrete tilt angles. However, because of the beam sensitivity of the specimen, data collection has to be performed under strict low-dose imaging conditions. Under these conditions, manual data collection is extremely tedious and prone to failure. Due to mechanical imperfections of the goniometer, the tilt axis is not completely stable. Therefore, even if the specimen has been carefully adjusted to the correct z-position (the eucentric height) and the tilt axis is well aligned to the optical axis of the microscope, the real tilt axis (and thus also the specimen) may move by several 100 nm, or even microns, when the specimen is tilted. This movement of the specimen produces image shifts as well as focus changes.

In automated electron tomography, all microscope operations required for recording data sets, including focusing and correction of image shifts, are done automatically. Essential in the data collection programs is the capability of low-dose data collection, e.g. recording of complete tilt series with a total dose of <5000 e/nm^2. Most automated programs that are set up for low-dose data collection can acquire tilt series for which only 3% of the electron irradiation is used for the overhead and 97% is actually used for the data themselves (Nickell *et al.*, 2005). The images are usually recorded by means of a slow-scan CCD camera in a digital format (Downing and Hendrickson, 1999; Faruqi *et al.*, 2003; Zhang *et al.*, 2003; Zuo, 2000). The image acquisition with a digital camera implies that changes in image position and defocus can be detected by on-line image processing and immediately be corrected for by computer control of the microscope. Moreover, the tilt series are directly available in digital format for subsequent processing and reconstruction. Since the original description of these systems (Dierksen *et al.*, 1992; Typke *et al.*, 1991), hardware and software

specifications have been improved: measures have been added to compensate more effectively for specimen drift during data collection, and procedures required for calibrating the microscope control at various magnifications and high tensions have been made more flexible and user-friendly.

4.2. First Generation: Image Shifts and Focus Tracking

In order not to lose the specimen area under investigation during the recording of a tilt series, image shifts and focus changes have to be measured and compensated for. To minimize the electron dose that is delivered to the specimen area of interest, the images required for focusing and tracking at the higher magnification are recorded on areas which do not coincide with the area of interest (Fig. 12).

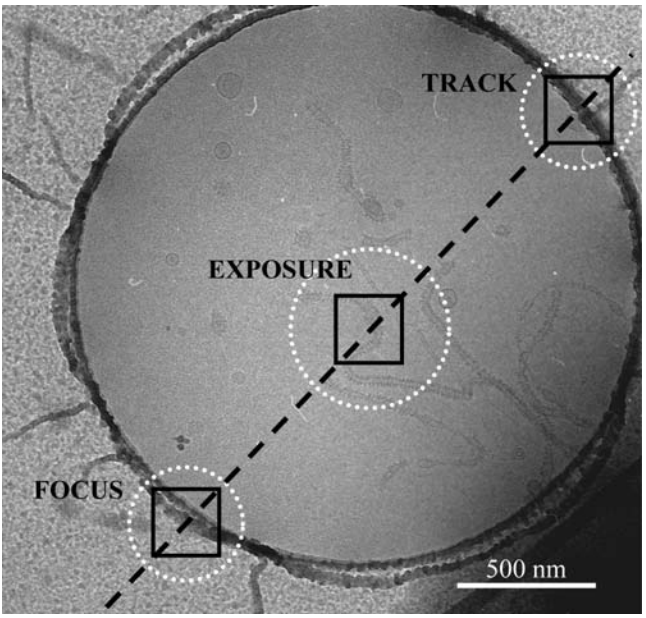


FIGURE 12. The first generation automated data collection approach. In this approach, three steps are carried out automatically at every tilt angle: TRACK, FOCUS and EXPOSURE. In order not to lose the specimen area under investigation during recording of a tilt series, image shifts and focus changes have to be measured and compensated for. These steps (TRACKING and FOCUS, respectively) are carried out at some distance from the field of interest, in non-overlapping areas situated along the tilt axis (dashed line). Magnifications closer to the one used for EXPOSURE are generally selected and the images can be binned (e.g. 4×4) to increase the readout of the CCD. Finally, the image is recorded in the area of interest, with such low-intensity illumination conditions that the dose applied is the corresponding fraction of the total allowed dose.

For the compensation of specimen movement, either the electron optical controls, i.e. the deflection coils and the objective lens current, or, alternatively, the mechanical specimen stage controls (x–y–z controls) are used. The electron optical control is fast, convenient, accurate and reproducible, and thus preferable to the control of the mechanical stage as long as the image shift is within the accessible range. The amount of shift and focus change that occurs during data collection depends on the goniometer and the specimen holder. Figure 11 shows an example of a specimen movement curve (shifts and focus changes). In most experiments, accumulated specimen shifts of only 1–3 μm occur during a tilt series, provided that the specimen has carefully been set to the eucentric height.

Image shifts are measured by determining the peak position in the cross-correlation function of an image recorded after setting a new tilt angle with a reference image that was recorded previously. The two correlated images are recorded at different tilt angles. Therefore, the two images will not only appear be shifted but will also differ to some extent due to the different projection angles. To compensate for this effect prior to the computation of the cross-correlation function, the image recorded at higher absolute tilt angles α is stretched perpendicular to the tilt axis by a factor according to $1/\cos(\alpha)$ (Dierksen *et al.*, 1992). To compensate for the image shift, the current in the image shift coils is changed (using an appropriate calibration) so that the mid-point of the feature of interest prior to the specimen tilt is centered back with this compensation. The above data collection scheme will perform worse at high magnifications, e.g. higher than 50,000, because the image shift per tilt increment may be larger than half the field of view of the CCD camera. Therefore, at high magnifications, the field of view will be too small to measure the displacement and a more elaborate data collection scheme has to be used (Dierksen *et al.*, 1992, 1993; Koster *et al.*, 1997; Rath *et al.*, 1997; Zhang *et al.*, 2001; Zheng *et al.*, 2004).

In one of these alternatives schemes, the shift measurements are done at a lower magnification, for instance at 4000, where the field of view increases to several μm^2. In some cases, it is optimal to choose the options for data collection such that after incrementing the tilt angle, the image shift is first measured and corrected for at low magnification (the SEARCH position). Next, at a higher magnification, a more accurate image centering is performed (at the TRACK position). The accuracy in compensating the image shift is better than 0.5% of the full image size. The accuracy in obtaining a tilt series data set is less accurate than 0.5%, usually due to sudden specimen movements of the holder ('jumps') that can occur when the tilt angle is incremented. In practice, an accuracy in the range of 2% of the full image size is reached.

After TRACKING, automatic focusing can be done with the technique of measuring the displacement of the image upon tilting the beam at the FOCUS position (Koster *et al.*, 1987, 1989). The autofocusing procedure determines the defocus by comparing two images acquired with different

(opposite) beam tilts. As an option to compensate for specimen drift, it is possible to record a third image and use this image to estimate and compensate for the drift between the images, thus separating beam tilt-induced from drift-induced shift. The amount of image shift, measured with cross-correlation, is proportional to the defocus and can thus be used for focus correction by changing the objective lens current. Although the accuracy in measuring the defocus can be very high, ~5 nm for thin samples (Koster *et al.*, 1987), the actual setting of the defocus will have an error in the order of 100 nm for the somewhat thicker typical specimens studied with electron tomography (Dierksen *et al.*, 1995). An increase in reliability for the autofocusing procedure can be obtained by taking into account that the specimen is tilted and therefore the defocus in the image will vary across the image (Ziese *et al.*, 2003). As stated above, the amount of image shift that will occur on invoking a beam tilt will be proportional to the defocus. Therefore, supposing that the center of the image is exactly in focus, for a tilted specimen the situation will be that the further the feature is away from the center of the image—and thus away from the focus plane—the larger the image shift will be. To ensure that all areas within the image of a tilted specimen will give rise to the same cross-correlation peak (which provides the average defocus), it is necessary to squeeze/stretch the two images used for defocus determination with a factor that depends on the direction of the tilt axis as well as on the amount of specimen tilt. The possible error in focus determination will be larger at low magnification (where the defocus ramp within the image can be larger).

Finally, after compensation for the shift and focus changes, a 2D image (part of the tilt series) is recorded (EXPOSURE). These four steps (SEARCH, TRACK, FOCUS and EXPOSURE) are repeated until the full angular tilt range is covered.

4.3. Second Generation: Pre-calibration

In the late 1990s, the mechanical construction of specimen holders and TEM goniometers was improved considerably, and a novel approach was evaluated to collect tilt series automatically (Ziese *et al.*, 2002a). Compared with the previous generation of holders, the novel holders exhibited reproducible specimen movement. The basic idea was to measure, prior to data acquisition, how the specimen moves, and extract from the movement parameters a mathematical model describing this behavior. The approach resulted in a 5-fold increase in speed, enabling the acquisition of 151 images in <20 min.

Experimentally, it was found that the actual shape of the *xyz* movement of the specimen (the specimen movement calibration curve) depends on the displacement of the tilt axis of the goniometer from the optical axis (Fig. 13), whether or not the specimen holder was removed/inserted, the *x*/*y* of the stage position, the *x*/*y*/*z* change in stage position, the *z*-height of an

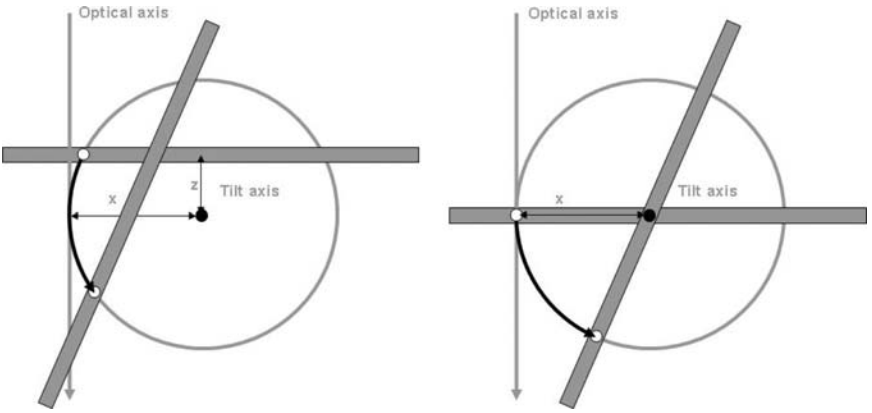


FIGURE 13. Observed movements when a specimen holder is tilted. The image shifts and the changes in defocus depend on both the distance of the tilt axis (*x*) of the goniometer from the optical axis and on the distance from the tilt axis to the eucentric height position (*z*). The measurement of shifts and changes in defocus as a function of specimen tilt can be used for calibration purposes as well as for predicting specimen movement.

image feature and the image shift settings. Only the last two parameters, the *z*-height of the specimen and the electron optical image shift settings, can be adjusted by the user. Note, however, that the displacement of the tilt axis of the goniometer from the optical axis is influenced by the electron optical image shift settings. With the best settings, there will be hardly any movements; otherwise, shifts can be positive or negative and can take values up to several micrometers.

Therefore, in its most basic application, the procedure for automated data collection would consist of two steps: first, a pre-calibration step in which the movement of the stage is measured in both the *xy*-plane (image shifts) and the *z*-direction (defocus change) for the range of tilt angles needed to acquire a tilt series. It is sufficient to record these measurements at low magnification with 5° tilt increments. Next, the actual data sets with smaller tilt increments and higher magnifications are recorded, interpolating the values of image shift and defocus change obtained in the first step. The modeling used for the pre-calibration method showed that the image displacements will be smaller when the optical axis of the microscope is aligned to the tilt axis of the specimen stage. This alignment can be done with the image shift controls inside the microscope. Furthermore, using the mathematical model, it was shown to be possible to predict the overall image movements after a few measurements (three) of image shift and defocus change (Ziese *et al.*, 2002a). This ability to make predictions was further developed and exploited in the third generation of data collection programs.

The approach based on changing the alignment to match the position of the specimen tilt axis enhanced the quality of a tilt series. For example, the amount of required defocus change compensation was reduced by an order of magnitude (from 10 μm to 1 μm), which made the problem of magnification change during data collection negligible. Another advantage of the approach is that the required computer-controlled steps for tracking image shifts and defocus changes are uncoupled from the acquisition of the tilt series. The uncoupling provides opportunities for acquisition modes other than bright-field imaging and the use of detectors other than CCD/film. For instance, a tilt series could be taken in STEM mode (Midgley and Weyland, 2003; Ziese *et al.*, 2002a,b) and, possibly simultaneously, (3D) element-specific information about the sample could be collected with an energy-dispersive X-ray (EDX) detector.

4.4. Third Generation: Prediction

The ongoing developments in specimen cryoholders made it possible to model the observations of image shifts and focus changes and to predict the overall image movements after a few measurements of image shift and defocus change (Zheng *et al.*, 2004; Ziese *et al.*, 2002a). For instance, by assuming that the sample follows a geometric rotation and that the optical system can be characterized in terms of an offset between the optical and mechanical axes, it was found that the image movement in the x, y and z directions due to stage tilt can be dynamically predicted with desired accuracy (15 nm in the x–y position and 100 nm in focus). Thus, the microscope optical system (beam/image shift and focus) can be automatically adjusted to compensate for the predicted image movement prior to taking the projected image at each tilt angle. As a consequence, it became no longer necessary either to record additional images for tracking and focusing during the course of data collections or to spend valuable set-up time in a pre-calibration of stage motions. This scheme was found to tolerate a significant degree of non-eucentricity and to be quite robust in the collection of regular and cryo low-dose images on thin or thick samples even at magnifications >62,000× and angular step as large as 10°. More elaborate schemes were proposed that included several steps to make the procedure more robust to unmodeled events (Nickell *et al.*, 2005).

Though the first generation of the automated data collection procedure was already very efficient by spending >95% of the electron dose on the actual data (and not on the overhead related to automation), the second and third generation of programs made data collection faster (5-fold) and more convenient. Current automation data collection developments are aimed at making the tomography procedure part of a larger automation process that includes automated (2D) scanning of the grid, as well as automated alignment and reconstruction (Carragher *et al.*, 2000). For applica-

tions of cryotomography where high resolution is an important factor, the requirements for the accuracy of defocus determination, as well as of obtaining quantitative information on the amount of phase/amplitude contrast, may become more stringent possibly to allow (automated) corrections for the CTF. Although the approach has been described (Winkler and Taylor, 2003), it is still unclear how corrections for the CTF can be effectuated on cryo-tomograms.

5. CONCLUDING REMARKS

Regardless of the specimen and the type of data recording, the series of 2D images acquired with the TEM need to be processed to obtain a 3D image: the tomogram. First, the images have to be aligned with respect to each other. For an extensive description on the alignment procedure, see Chapters 5 and 6 of this volume. Briefly, several alignment procedures are available. Because of the very low SNR in the images, the alignment step can pose problems. During alignment, the relative translation between the individual images will need to be determined, a possible gradual change in scaling (magnification) and possibly a change in image rotation. Some of the procedures available make use of fiducial markers present within or on the sample, whereas other methods rely fully on the contrast generated by the material within the section and do not need markers (Liu *et al.*, 1995) (see also Section 2.5).

Next, after alignment, the tomogram can be computed using either resolution-weighted back-projection (Radermacher, 1992) or iterative refinement methods. For an extensive description on the reconstruction phase, see Chapters 7, 8 and 9 of this volume. Briefly, we like to point out that the iterative methods, e.g. maximum entropy (Barth *et al.*, 1988; Skoglund *et al.*, 1996), ART (Marabini *et al.*, 1997) or SIRT (Sorzano *et al.*, 2001), can produce reconstructions that appear smoother than tomograms produced with weighted back-projection, which can considerably help the subsequent visualization.

One of the major problems is that cryotomograms suffer from substantial residual noise. The low SNR is particularly unavoidable at higher resolutions because of the dose sensitivity of the specimen. The individual images within a tilt series are very noisy and, as a result, the tomogram generated is also noisy, in spite of optimized data collection schemes and optimal reconstruction approaches. In specific applications, such as cryotomography on whole cells, an extra problem is due to the fact that the cytoplasm within the specimen is densely packed with molecules connecting to each other (Grunewald *et al.*, 2003). Within cryotomograms, it is therefore virtually impossible to apply volume segmentation and feature extraction based upon visual inspection of the tomogram, which is an approach often

successful with other types of samples which are less low-dose demanding and with much higher contrast as a result of staining with metal salts. Other characteristics inherent to the technique complicate the tomogram segmentation and the analysis of the results. For example, for correct interpretation of the tomograms, the geometry of data collection has to be taken into account (double- or single-axis tilting, the tilt range and tilt increments) as well as special characteristics that are linked to image formation (contrast enhancement by zero-loss filtering and large defocus to enhance contrast).

In most 3D visualization approaches, isosurface representation or volume rendering is used to display the 3D volume. However, the inherent noise and low contrast of cryotomograms make such an approach in most cases not effective. Special means need to be used to smooth the appearance of the reconstruction in order to extract morphological information. To enhance the SNR, spatial filtering, such as median or Gaussian filtering, can be applied. A characteristic of these approaches is that the filtering enhances the SNR in the image at the cost of resolution. The high-resolution information is hidden in the noise, and straightforward application of spatial filters will eliminate the high-resolution information.

In order to segment and analyze 3D subtomograms from the tomograms or to extract quantitative information, dedicated strategies and innovative image analysis techniques are needed. Special algorithms were developed to minimize the resolution-lowering effects of isotropic low-pass filters. Depending on the particular question at hand, other algorithms might be effective in extracting the information. For instance, segmenting a volume in different objects can be quite effective using eigenvector analysis (Frangakis and Hegerl, 2002), or using wathershed transformations (Volkmann, 2002). To filter out objects with a particular shape, wavelet transformation might be useful (Frangakis *et al.*, 2001). One type of filter that turned out to be a powerful tool for visualization uses non-linear anisotropic diffusion to reduce noise in electron tomograms (Fernandez and Li, 2003; Frangakis and Hegerl, 2001).

Though spatial filtering techniques can be very useful to visualize the overall shape of structures within tomograms, they remove the high-resolution information, and the presence of individual molecular structures can be concealed. However, it is not necessarily required to identify the 3D molecular structures by direct visual inspection. The identification of a given molecular structure can be based on 3D correlation, or on other pattern recognition methods. A 3D cross-correlation peak may be significant even if the particle is hardly visually recognizable. Suitable reference structures can be derived from data obtained with other high-resolution techniques such as X-ray crystallography, NMR or electron crystallography. The pattern recognition approach, as developed for cryotomography, is known as template matching and is an algorithm capable of detecting and identifying macromolecules in tomographic volumes in a fully automated manner

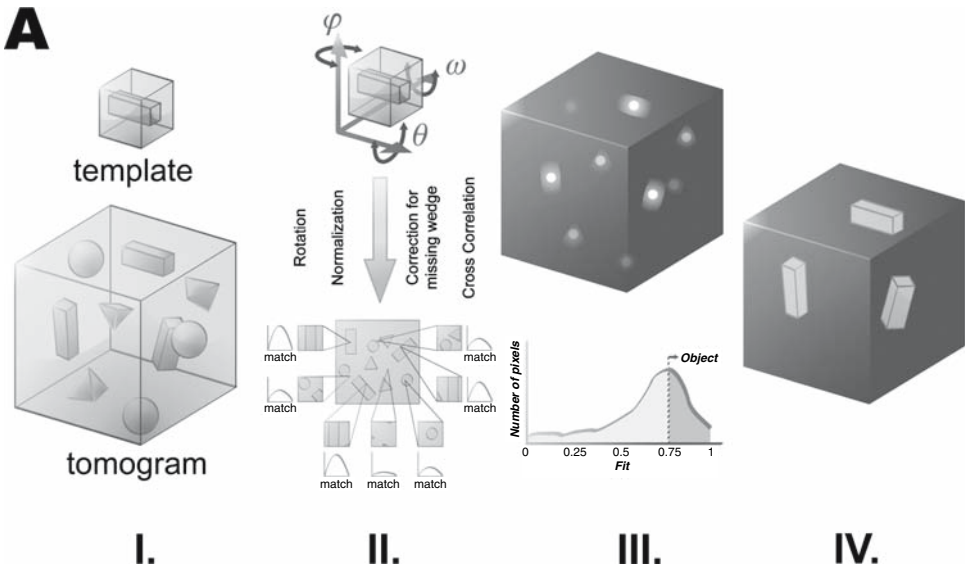


FIGURE 14. Identification of individual macromolecules based on their high-resolution structure. The pattern recognition approach as developed for cryotomography is known as template matching and is an algorithm capable of detecting and identifying macromolecules in tomographic volumes in a fully automated manner (Böhm *et al.* 2000; Frangakis *et al.* 2002). The algorithm is based on non-linear cross-correlation and incorporates elements of multivariate statistical analysis.

(Böhm *et al.*, 2000; Frangakis *et al.*, 2002; Rath *et al.*, 2003) (Fig. 14). The algorithm is based on non-linear cross-correlation and incorporates elements of multivariate statistical analysis. This approach can be suitable to identify macromolecules in the size range of 0.5–1 MDa with good fidelity in a cryotomogram with a resolution of 4–6 nm (Forster *et al.*, 2005). Methods for multivariate statistical analysis of sets of tomograms allow the classification of these volumes into homogenous sets. In this manner, only identical units are averaged (Frangakis *et al.*, 2002; Walz *et al.*, 1997b). The method is applicable not only to identical units within a tomogram but also to identical units coming from different tomograms. This type of strategy constitutes a way to enhance the resolution significantly.

Cryotomography is a technique capable of bridging the gap between high-resolution (nanometer resolution scale) imaging and molecular imaging (atomic resolution) but also between nanometer scale and fluorescence light microscopic imaging as used for live cell imaging (micrometer resolution scale). Though technical developments in bridging this gap are ongoing, solutions have not matured yet.

In order to bridge the resolution gap with techniques achieving atomic resolution, the information provided by macromolecule structures resolved

by high-resolution methods (e.g. single-particle reconstruction methods with electron microscopy, X-ray diffraction methods or NMR methods) needs to be put into the relatively low resolution cryotomograms of cell organelles or cells (Sali *et al.*, 2003). Cryotomography, by providing 3D images at moderate resolution recorded in a non-invasive mode with minimal perturbations of the biological systems, could be a powerful approach to determine the relative location of molecular structures using pattern recognition techniques. Such an approach attempts to identify macromolecular complexes as a function of their structural signature in cryotomograms of cryofixed cells.

In this chapter, we zoomed into some of the theoretical and practical issues related to cryotomography. Reading this chapter will make it clear that the electron microscopist performing cryotomography will be confronted with various challenges related to specimen preparation, data collection, image processing and aspects of pattern recognition to exploit the tomograms. In spite of these hurdles, we hope to have shown that the outcome, the cryotomogram, can be very gratifying. In spite of their moderate resolution (5–10 nm), cryotomograms of frozen-hydrated biological structures reveal an astonishing amount of information about cell architecture as a network of molecular interactions.

ACKNOWLEDGEMENTS

We are very grateful to Carmen Mannella (Albany) for the critically reading of the manuscript and for the many helpful suggestions to improve the correctness and readability of this chapter. Figure 14 was kindly provided by Misjael Lebbink (Utrecht University).

REFERENCES

Adrian, M., Dubochet, J. *et al.* (1998). Cryo-negative staining. *Micron* **29**:145–160.

Adrian, M., ten Heggeler-Bordier, B. *et al.* (1990). Direct visualization of supercoiled DNA molecules in solution. *EMBO J.* **9**:4551–4554.

Ahting, U., Thun, C. *et al.* (1999). The TOM core complex: the general protein import pore of the outer membrane of mitochondria. *J. Cell Biol.* **147**:959–968.

Al-Amoudi, A., Chang, J. J. *et al.* (2004). Cryo-electron microscopy of vitreous sections. *EMBO J*. **23**:3583–3588.

Al-Amoudi, A., Dubochet, J. *et al.* (2003). An oscillating cryo-knife reduces cutting-induced deformation of vitreous ultrathin sections. *J. Microsc.* **212**:26–33.

Al-Amoudi, A., Studer, D. *et al.* (2005). Cutting artefacts and cutting process in vitreous sections for cryo-electron microscopy. *J. Struct. Biol.* **150**:109–121.

Angert, I., Majorovits, E. *et al.* (2000). Zero-loss image formation and modified contrast transfer theory in EFTEM. *Ultramicroscopy* **81**:203–222.

Bajaj, C., Yu, Z. *et al.* (2003). Volumetric feature extraction and visualization of tomographic molecular imaging. *J. Struct. Biol.* **144**:132–143.

Barth, M., Bryan, R. K. *et al.* (1988). Estimation of missing cone data in three-dimensional electron microscopy. *Scanning Microsc. Suppl.* **2**:277–284.

Baumeister, W. (2002). Electron tomography: towards visualizing the molecular organization of the cytoplasm. *Curr. Opin. Struct. Biol.* **12**:679–684.

Baumeister, W. (2005). From proteomic inventory to architecture. *FEBS Lett.* **579**:933–937.

Baumeister, W., Grimm, R. *et al.* (1999). Electron tomography of molecules and cells. *Trends Cell Biol.* **9**:81–85.

Baumeister, W. and Steven, A. C. (2000). Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem. Sci.* **25**:624–631.

Beck, M., Forster, F. *et al.* (2004). Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. *Science* **306**:1387–1390.

Benjamin, J., Ganser-Pornillos, B. K. *et al.* (2005). Three-dimensional structure of HIV-1 virus-like particles by electron cryotomography. *J. Mol. Biol.* **346**:577–588.

Biel, S. S., K. Kawaschinski, *et al.* (2003). From tissue to cellular ultrastructure: closing the gap between micro- and nanostructural imaging. *J Microsc* **212**:91–99.

Böhm, J., Frangakis, A. S. *et al.* (2000). Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proc. Natl Acad. Sci. USA* **97**:14245–14250.

Böhm, J., Lambert, O. *et al.* (2001). FhuA-mediated phage genome transfer into liposomes: a cryo-electron tomography study. *Curr. Biol.* **11**:1168–1175.

Bongini, L., Fanelli, D. *et al.* (2004). Freezing immunoglobulins to see them move. *Proc. Natl Acad. Sci. USA* **101**:6466–6471.

Bouwer, J. C., Mackey,M. R. *et al.* (2004). Automated most-probable loss tomography of thick selectively stained biological specimens with quantitative measurement of resolution improvement. *J. Struct. Biol.* **148**:297–306.

Braet, F. (2004). How molecular microscopy revealed new insights into the dynamics of hepatic endothelial fenestrae in the past decade. *Liver Int.* **24**:532–539.

Brandt, S., Heikkonen, J. *et al.* (2001). Automatic alignment of transmission electron microscope tilt series without fiducial markers. *J. Struct. Biol.* **136**:201–213.

Braunfeld, M. B., Koster, A. J. *et al.* (1994). Cryo automated electron tomography: towards high-resolution reconstructions of plastic-embedded structures. *J Microsc* **174**:75–84.

Bullitt, E., Rout, M. P. *et al.* (1997). The yeast spindle pole body is assembled around a central crystal of Spc42p. *Cell* **89**:1077–1086.

Burge, R. E., Dainty, J. C. *et al.* (1977). Optical and digital image processing in high-resolution electron microscopy. *Ultramicroscopy* **2**:169–178.

Cardone, G., Grünewald, K., Steven, A.C. (2005). A resolution criterion for electron tomography based on cross-validation. *J. Struct. Biol.* **151**:117–119.

Carragher, B., Kisseberth, N. *et al.* (2000). Leginon: an automated system for acquisition of images from vitreous ice specimens. *J. Struct. Biol.* **132**:33–45.

Chestnut, M. H., Siegel, D. P. *et al.* (1992). A temperature-jump device for time-resolved cryo-transmission electron microscopy. *Microsc. Res. Tech*. **20**:95–101.

Coene, W., Janssen, G. *et al.* (1992). Phase retrieval through focus variation for ultra-resolution in field-emission transmission electron microscopy. *Phys. Rev. Lett.* **69**:3743–3746.

Crowther, R. A., Amos, L. A. *et al.* (1970). Three dimensional reconstructions of spherical viruses by fourier synthesis from electron micrographs. *Nature* **226**:421–425.

Cyrklaff, M., M. Adrian, *et al.* (1990). Evaporation during preparation of unsupported thin vitrified aqueous layers for cryo-electron microscopy. *J. Electron Microsc. Tech.* **16**:351–355.

Cyrklaff, M., C. Risco, *et al.* (2005). Cryo-electron tomography of vaccinia virus. *Proc. Natl Acad. Sci. USA* **102**:2772–2777.

De Carlo, S., El-Bez, C. *et al.* (2002). Cryo-negative staining reduces electron-beam sensitivity of vitrified biological particles. *J. Struct. Biol.* **138**:216–226.

DeRosier, D. J. and Klug, A. (1968). Reconstruction of three-dimensional structures from electron micrographs. *Nature* **217**:130–134.

Dierksen, K., Typke, D. *et al.* (1993). Implementation of autofocus and low-dose procedures. *Ultramicroscopy* **49**:109–120.

Dierksen, K., Typke, D. *et al.* (1992). Towards automatic electron tomography. *Ultramicroscopy* **40**:71–87.

Dierksen, K., Typke, D. *et al.* (1995). Three-dimensional structure of lipid vesicles embedded in vitreous ice and investigated by automated electron tomography. *Biophys. J.* **68**:1416–1422.

Downing, K. H. and Hendrickson, F. M. (1999). Performance of a 2k CCD camera designed for electron crystallography at 400 kV. *Ultramicroscopy* **75**:215–233.

Dubochet, J., Adrian, M. *et al.* (1988). Cryo-electron microscopy of vitrified specimens. *Q. Rev. Biophys.* **21**:129–228.

Dubochet, J. and Sartori Blanc, N. (2001). The cell in absence of aggregation artifacts. *Micron* **32**:91–99.

Erk, I., Nicolas, G. *et al.* (1998). Electron microscopy of frozen biological objects: a study using cryosectioning and cryosubstitution. *J. Microsc.* **189**:236–248.

Faruqi, A. R., Cattermole, D. M. *et al.* (2003). Evaluation of a hybrid pixel detector for electron microscopy. *Ultramicroscopy* **94**:263–276.

Feja, B. and Aebi, U. (1999). Determination of the inelastic mean free path of electrons in vitrified ice layers for on-line thickness measurements by zero-loss imaging. *J. Microsc.* **193**:15–19.

Fernandez, J. J. and Li, S. (2003). An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms. *J. Struct. Biol.* **144**:152–161.

Forster, F., Medalia, O. *et al.* (2005). Retrovirus envelope protein complex structure in situ studied by cryo-electron tomography. *Proc. Natl Acad. Sci. USA* **102**:4729–4734.

Frangakis, A. S., Bohm, J. *et al.* (2002). Identification of macromolecular complexes in cryo-electron tomograms of phantom cells. *Proc. Natl Acad. Sci. USA* **99**:14153–14158.

Frangakis, A. S. and Hegerl, R. (2001). Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. *J. Struct. Biol.* **135**:239–250.

Frangakis, A. S. and Hegerl, R. (2002). Segmentation of two- and three-dimensional data from electron microscopy using eigenvector analysis. *J. Struct. Biol.* **138**:105–113.

Frangakis, A. S., Stoschek, A. *et al.* (2001). Wavelet transform filtering and nonlinear anisotropic diffusion assessed for signal reconstruction performance on multidimensional biomedical data. *IEEE Trans. Biomed. Eng.* **48**:213–222.

Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **31**: 303–319.

Frank, J., Radermacher, M. *et al.* (1996). SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**:190–199.

Frank, J., Wagenknecht, T. *et al.* (2002). Three-dimensional imaging of biological complexity. *J. Struct. Biol.* **138**:85–91.

Frederik, P. M., Busing, W. M. *et al.* (1982). Concerning the nature of the cryosectioning process. *J. Microsc.* **125**:167–175.

Freitag, B., Kujawa, S. *et al.* (2005). Breaking the spherical and chromatic aberration barrier in transmission electron microscopy. *Ultramicroscopy* **102**:209–214.

Fung, J. C., Liu, W. *et al.* (1996). Toward fully automated high-resolution electron tomography. *J. Struct. Biol.* **116**:181–189.

Grimm, R., Barmann, M. *et al.* (1997). Energy filtered electron tomography of ice-embedded actin and vesicles. *Biophys. J.* **72**:482–489.

Grimm, R., Singh, H. *et al.* (1998). Electron tomography of ice-embedded prokaryotic cells. *Biophys. J.* **74**:1031–1042.

Grimm, R., Typke, D. *et al.* (1996). Determination of the inelastic mean free path in ice by examination of tilted vesicles and automated most probable loss imaging. *Ultramicroscopy* **63**:169–179.

Grunewald, K., Desai, P. *et al.* (2003). Three-dimensional structure of herpes simplex virus from cryo-electron tomography. *Science* **302**:1396–1398.

Grunewald, K., Medalia, O. *et al.* (2003). Prospects of electron cryotomography to visualize macromolecular complexes inside cellular compartments: implications of crowding. *Biophys. Chem.* **100**:577–591.

Han, K. F., Sedat, J. W. *et al.* (1995). Mechanism of image formation for thick biological specimens: exit wavefront reconstruction and electron energy-loss spectroscopic imaging. *J. Microsc.* **178**:107–119.

Han, K. F., Sedat, J. W. *et al.* (1997). Practical image restoration of thick biological specimens using multiple focus levels in transmission electron microscopy. *J. Struct. Biol.* **120**:237–244.

Hart, R. G. (1968). Electron microscopy of unstained biuological material: The polytropic montage. *Science* **159**:1464–1467.

He, W., Cowin, P. *et al.* (2003). Untangling desmosomal knots with electron tomography. *Science* **302**:109–113.

Hegerl, R. and Hoppe, W. (1976). Influence of electron noise on three-dimensional image reconstruction. *Z. Naturforsch.* **31a**:1717–1721.

Henderson, R. (1995). The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**: 171–193.

Heymann, J. B., Cheng, N. *et al.* (2003). Dynamics of herpes simplex virus capsid maturation visualized by time-lapse cryo-electron microscopy. *Nat. Struct. Biol.* **10**:334–341.

Hoppe, W. (1969). Das Endlichkeitspostulat und das Interpolationstheorem der dreidimensionalen electronenmikroscopischen Analyse aperiodischer Strukturen. *Optik* **29**: 617–621.

Horowitz, R. A., Koster, A. J. *et al.* (1997). Automated electron microscope tomography of frozen-hydrated chromatin: the irregular three-dimensional zigzag architecture persists in compact, isolated fibers. *J. Struct. Biol.* **120**:353–362.

Hsieh, C. E., Marko, M. *et al.* (2002). Electron tomographic analysis of frozen-hydrated tissue sections. *J. Struct. Biol.* **138**:63–73.

Humbel, B. M., Weber, K. *et al.* (1991). Versatile controlling system for cryopreparation techniques in electron microscopy. *J. Electron Microsc. Tech.* **17**:450–455.

Koster, A. J., A. van der Bos, *et al.* (1987). An autofocus method for a TEM. *Ultramicroscopy* **21**: 209–222.

Koster, A. J., Chen, H. *et al.* (1992). Automated microscopy for electron tomography. *Ultramicroscopy* **46**:207–227.

Koster, A. J., Grimm, R. *et al.* (1997). Perspectives of molecular and cellular electron tomography. *J. Struct. Biol.* **120**:276–308.

Koster, A. J., der Ruijter, W. J. *et al.* (1989). Autofocus method for a TEM using minimum electron dose. *Ultramicroscopy* **27**:251–272.

Kremer, J. R., Mastronarde, D. N. *et al.* (1996). Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**:71–76.

Kurner, J., Frangakis, A. S. *et al.* (2005). Cryo-electron tomography reveals the cytoskeletal structure of *Spiroplasma melliferum*. *Science* **307**:436–448.

Kurner, J., Medalia, O. *et al.* (2004). New insights into the structural organization of eukaryotic and prokaryotic cytoskeletons using cryo-electron tomography. *Exp. Cell Res.* **301**:38–42.

Leapman, R. D. and Sun, S. (1995). Cryo-electron energy loss spectroscopy: observations on vitrified hydrated specimens and radiation damage. *Ultramicroscopy* **59**:71–79.

Lepault, J., Booy, F. P. *et al.* (1983). Electron microscopy of frozen biological suspensions. *J. Microsc.* **129**: 89–102.

Lepault, J., Erk, I. *et al.* (1991). Time-resolved cryo-electron microscopy of vitrified muscular components. *J. Microsc.* **161**:47–57.

Liu, Y., Penczek, P. A. *et al.* (1995). A marker-free alignment method for electron tomography. *Ultramicroscopy* **58**:393–402.

Lucic, V., Yang, T. *et al.* (2005). Morphological characterization of molecular complexes present in the synaptic cleft. *Structure (Camb.)* **13**:423–334.

Mannella, C. A. (2006). The relevance of mitochondrial membrane topology to mitochondrial function. *Biochim Biophys Acta* **1762**:140–147.

Mannella, C. A., Pfeiffer, D. R. *et al.* (2001). Topology of the mitochondrial inner membrane: dynamics and bioenergetic implications. *IUBMB Life* **52**:93–100.

Marabini, R., Rietzel, E. *et al.* (1997). Three-dimensional reconstruction from reduced sets of very noisy images acquired following a single-axis tilt schema: application of a new three-dimensional reconstruction algorithm and objective comparison with weighted backprojection. *J. Struct. Biol.* **120**:363–371.

Marco, S., Boudier, T. *et al.* (2004). Electron tomography of biological samples. *Biochemistry* **69**:1219–1225.

Marsh, B. J., Volkmann, N. *et al.* (2004). Direct continuities between cisternae at different levels of the Golgi complex in glucose-stimulated mouse islet beta cells. *Proc. Natl Acad. Sci. USA* **101**:5565–5570.

Mastronarde, D. N. (1997). Dual-axis tomography: an approach with alignment methods that preserve resolution. *J. Struct. Biol.* **120**:343–352.

Matias, V. R., Al-Amoudi, A. *et al.* (2003). Cryo-transmission electron microscopy of frozen-hydrated sections of *Escherichia coli* and *Pseudomonas aeruginosa*. *J. Bacteriol.* **185**:6112–6118.

McEwen, B. F., Downing, K. H. *et al.* (1995). The relevance of dose-fractionation in tomography of radiation-sensitive specimens. *Ultramicroscopy* **60**:357–373.

McEwen, B. F. and Frank, J. (2001). Electron tomographic and other approaches for imaging molecular machines. *Curr. Opin. Neurobiol.* **11**:594–600.

McEwen, B. F., Marko, M. *et al.* (2002). Use of frozen-hydrated axonemes to assess imaging parameters and resolution limits in cryoelectron tomography. *J. Struct. Biol.* **138**:47–57.

McIntosh, R., Nicastro, D. *et al.* (2005). New views of cells in 3D: an introduction to electron tomography. *Trends Cell Biol.* **15**:43–51.

Medalia, O., Typke, D. *et al.* (2002a). Cryoelectron microscopy and cryoelectron tomography of the nuclear pre-mRNA processing machine. *J. Struct. Biol.* **138**:74–84.

Medalia, O., Weber, I. *et al.* (2002b). Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* **298**:1209–1213.

Messaoudi, C., Boudier, T. *et al.* (2003). Use of cryo-negative staining in tomographic reconstruction of biological objects: application to T4 bacteriophage. *Biol. Cell* **95**:393–398.

Midgley, P. A. and Weyland M. (2003). 3D electron microscopy in the physical sciences: the development of Z-contrast and EFTEM tomography. *Ultramicroscopy* **96**:413–431.

Moritz, M., Braunfeld,M. B. *et al.* (1995). Three-dimensional structural characterization of centrosomes from early *Drosophila* embryos. *J. Cell Biol.* **130**:1149–1159.

Muller, E. G., Snydsman, B. E. *et al.* (2005). The organization of the core proteins of the yeast spindle pole body. *Mol. Biol. Cell*. **16**:3341–3352.

Murk, J. L., Humbel, B. M. *et al.* (2003). Endosomal compartmentalization in three dimensions: implications for membrane fusion. *Proc. Natl Acad. Sci. USA* **100**:13332–13337.

Nicastro, D., Frangakis, A. S. *et al.* (2000). Cryo-electron tomography of *Neurospora* mitochondria. *J. Struct. Biol.* **129**:48–56.

Nickell, S., Forster, F. *et al.* (2005). TOM software toolbox: acquisition and analysis for electron tomography. *J. Struct. Biol.* **149**:227–234.

Nickell, S., Hegerl, R. *et al.* (2003). *Pyrodictium* cannulae enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *J. Struct. Biol.* **141**:34–42.

Nitsch, M., Walz, J. *et al.* (1998). Group II chaperonin in an open conformation examined by electron tomography. *Nat. Struct. Biol.* **5**:855–857.

Owen, C. H. and Landis, W. J. (1996). Alignment of electron tomographic series by correlation without the use of gold particles. *Ultramicroscopy* **63**:27–38.

Patwardhan, A. (2003). Transmission electron microscopy of weakly scattering objects described by operator algebra. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **20**:1210–1222.

Penczek, P., Marko, M. *et al.* (1995). Double-tilt electron tomography. *Ultramicroscopy* **60**:393–410.

Penczek, P. A. (2002). Three-dimensional spectral signal-to-noise ratio for a class of reconstruction algorithms. *J. Struct. Biol.* **138**:34–46.

Perkins, G. A., Renken, C. W. *et al.* (1997). Electron tomography of large, multicomponent biological structures. *J. Struct. Biol.* **120**:219–27.

Radermacher, M. (1992). *Weighted Back-projection Methods*. Plenum Press, New York.

Radermacher, M. (1994). Three-dimensional reconstruction from random projections: orientational alignment via Radon transforms. *Ultramicroscopy* **53**:121–136.

Radon, J. (1917). Über die Bertimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Ber. Verh. König. Sächs. Ges. Wiss. Math. Phs. Klasse* **69**:262–277.

Rath, B. K., Hegerl, R. *et al.* (2003). Fast 3D motif search of EM density maps using a locally normalized cross-correlation function. *J. Struct. Biol.* **144**:95–103.

Rath, B. K., Marko, M. *et al.* (1997). Low-dose automated electron tomography: a recent implementation. *J. Struct. Biol.* **120**:210–218.

Ress, D. B., Harlow,M. L. *et al.* (2004). Methods for generating high-resolution structural models from electron microscope tomography data. *Structure (Camb.)* **12**:1763–1774.

Rockel, B., Walz, J. *et al.* (1999). Structure of VAT, a CDC48/p97 ATPase homologue from the archaeon *Thermoplasma acidophilum* as studied by electron tomography. *FEBS Lett.* **451**:27–32.

Rose, H. (2005). Prospects for aberration-free electron microscopy. *Ultramicroscopy* **103**:1–6.

Russell, R. B., Alber, F. *et al.* (2004). A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.* **14**:313–324.

Sali, A., Glaeser, R. *et al.* (2003). From words to literature in structural proteomics. *Nature* **422**:216–225.

Sandin, S., Ofverstedt, L. G. *et al.* (2004). Structure and flexibility of individual immunoglobulin G molecules in solution. *Structure (Camb.)* **12**:409–415.

Sawada, H., Tomita, T., Naruse, M., Honda, T., Hambridge, P., Hartel, P., Haider, M., Hetherington, C., Doole, R., Kirkland, A., Hutchison, J., Titchmarsh, J. and Cockayne, D. (2005) Experimental evaluation of a spherical aberration-corrected TEM and STEM. *J. Electron Microsc. (Tokyo)* **54**:119–121.

Saxton, W. O. and Baumeister, W. (1982). The correlation averaging of a regularly arranged bacterial cell envelope protein. *J. Microsc.* **127**:127–138.

Saxton, W. O., Baumeister,W. *et al.* (1984). Three-dimensional reconstruction of imperfect two-dimensional crystals. *Ultramicroscopy* **13**(1-2):57–70.

Sherman, M. B., Brink, J. *et al.* (1996). Performance of a slow-scan CCD camera for macromolecular imaging in a 400 kV electron cryomicroscope. *Micron* **27**:129–139.

Sherman, M. B., Jakana, J. *et al.* (1997). A strategy for electron tomographic data collection and crystallographic reconstruction of biological bundles. *J. Struct. Biol.* **120**:245–256.

Sitte, H. (1996). Advanced instrumentation and methodology related to cryoultramicrotomy: a review. *Scanning Microsc. Suppl.* **10**:387–463; discussion 463–466.

Skoglund, U., Ofverstedt, L. G. *et al.* (1996). Maximum-entropy three-dimensional reconstruction with deconvolution of the contrast transfer function: a test application with adenovirus. *J. Struct. Biol.* **117**:173–188.

Somlyo, A. P. and Shuman,H. (1982). Electron probe and electron energy loss analysis in biology. *Ultramicroscopy* **8**:219–233.

Sorzano, C. O., Marabini, R. *et al.* (2001). The effect of overabundant projection directions on 3D reconstruction algorithms. *J. Struct. Biol.* **133**:108–118.

Steven, A. C. and Aebi, U. (2003). The next ice age: cryo-electron tomography of intact cells. *Trends Cell Biol.* **13**:107–110.

Steven, A. C., Heymann, J. B. *et al.* (2005). Virus maturation: dynamics and mechanism of a stabilizing structural transition that leads to infectivity. *Curr. Opin. Struct. Biol.* **15**:227–236.

Stoffler, D., Feja, B. *et al.* (2003). Cryo-electron tomography provides novel insights into nuclear pore architecture: implications for nucleocytoplasmic transport. *J. Mol. Biol.* **328**:119–130.

Studer, D. and Gnaegi, H. (2000). Minimal compression of ultrathin sections with use of an oscillating diamond knife. *J. Microsc.* **197**:94–100.

Studer, D., Michel, M. *et al.* (1989). High pressure freezing comes of age. *Scanning Microsc Suppl.* **3**: 253–268; discussion 268–269.

Subramaniam, S. and Milne, J. L. (2004). Three-dimensional electron microscopy at molecular resolution. *Annu. Rev. Biophys. Biomol. Struct.* **33**:141–155.

Talmon, Y. (1987). Cryotechniques in biological electron microscopy. In *Electron beam radiation damage to organic and biological cryospecimens* (R. A. Steinbrecht and K. Zierold, eds). Berlin, Springer-Verlag, pp. 64–86.

Talmon, Y., Burns, J. L. *et al.* (1990). Time-resolved cryotransmission electron microscopy. *J. Electron Microsc. Tech* **14**:6–12.

Taylor, K. A., Schmitz, H. *et al.* (1999). Tomographic 3D reconstruction of quick-frozen, Ca2+-activated contracting insect flight muscle. *Cell* **99**:421–431.

Toyoshima, C. and Unwin, N. (1988). Contrast transfer for frozen-hydrated specimens: determination from pairs of defocused images. *Ultramicroscopy* **25**:279–291.

Typke, D., Downing, K. H. *et al.* (2004). Electron microscopy of biological macromolecules: bridging the gap between what physics allows and what we currently can get. *Microsc. Microanal.* **10**:21–27.

Typke, D., Nordmeyer, R. A. *et al.* (2005). High-throughput film-densitometry: an efficient approach to generate large data sets. *J. Struct. Biol.* **149**:17–29.

Unser, M., Sorzano, C. O. *et al.* (2005). Spectral signal-to-noise ratio and resolution assessment of 3D reconstructions. *J. Struct. Biol.* **149**:243–255.

Van Aert, S., den Dekker, A. J. *et al.* (2002). High-resolution electron microscopy and electron tomography: resolution versus precision. *J. Struct. Biol.* **138**:21–33.

Volkmann, N. (2002). A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. *J. Struct. Biol.* **138**:123–129.

von Schack, M. L., Fakan, S. *et al.* (1993). Cryofixation and cryosubstitution: a useful alternative in the analyses of cellular fine structure. *Eur. J. Histochem.* **37**:5–18.

Wagenknecht, T., Hsieh, C. E. *et al.* (2002). Electron tomography of frozen-hydrated isolated triad junctions. *Biophys. J.* **83**:2491–2501.

Walz, J., Koster, A. J. *et al.* (1999). Capsids of tricorn protease studied by electron cryomicroscopy. *J. Struct. Biol.* **128**:65–68.

Walz, J., Tamura, T. *et al.* (1997a). Tricorn protease exists as an icosahedral supermolecule *in vivo*. *Mol. Cell* **1**:59–65.

Walz, J., Typke, D. *et al.* (1997b). Electron tomography of single ice-embedded macromolecules: three-dimensional alignment and classification. *J. Struct. Biol.* **120**:387–395.

White, H. D., Thirumurugan, K. *et al.* (2003). A second generation apparatus for time-resolved electron cryo-microscopy using stepper motors and electrospray. *J. Struct. Biol.* **144**:246–252.

Winkler, H. and Taylor, K. A. (2003). Focus gradient correction applied to tilt series image data used in electron tomography. *J. Struct. Biol.* **143**:24–32.

Zhang, P., Beatty, A. *et al.* (2001). Automated data collection with a Tecnai 12 electron microscope: applications for molecular imaging by cryomicroscopy. *J. Struct. Biol.* **135**:251–261.

Zhang, P., Borgnia, M. J. *et al.* (2003). Automated image acquisition and processing using a new generation of 4K x 4K CCD cameras for cryo electron microscopic studies of macromolecular assemblies. *J. Struct. Biol.* **143**:135–144.

Zhang, P., Bos, E. *et al.* (2004). Direct visualization of receptor arrays in frozen-hydrated sections and plunge-frozen specimens of *E. coli* engineered to overproduce the chemotaxis receptor Tsr. *J. Microsc.* **216**:76–83.

Zhang, P., Land, W. *et al.* (2005). Electron tomography of degenerating neurons in mice with abnormal regulation of iron metabolism. *J. Struct. Biol.* **150**:144–153.

Zhao, Q., Ofverstedt, L. G. *et al.* (2004). Morphological variation of individual *Escherichia coli* 50S ribosomal subunits in situ, as revealed by cryo-electron tomography. *Exp. Cell Res.* **300**:190–201.

Zheng, Q. S., Braunfeld, M. B. *et al.* (2004). An improved strategy for automated electron microscopic tomography. *J. Struct. Biol.* **147**:91–101.

Zhu, J., Penczek, P. A. *et al.* (1997). Three-dimensional reconstruction with contrast transfer function correction from energy-filtered cryoelectron micrographs: procedure and application to the 70S *Escherichia coli* ribosome. *J. Struct. Biol.* **118**:197–219.

Ziese, U., Geerts,W. J. *et al.* (2003). Correction of autofocusing errors due to specimen tilt for automated electron tomography. *J. Microsc.* **211**:179–85.

Ziese, U., Janssen, A. H.*et al.* (2002). Automated high-throughput electron tomography by precalibration of image shifts. *J. Microsc.* **205**:187–200.

Ziese, U., C. Kubel, *et al.* (2002). Three-dimensional localization of ultrasmall immuno-gold labels by HAADF-STEM tomography. *J. Struct. Biol.* **138**:58–62.

Zuo, J. M. (2000). Electron detection characteristics of a slow-scan CCD camera, imaging plates and film, and electron image restoration. *Microsc. Res. Tech.* **49**:245–268.

5

Fiducial Marker and Hybrid Alignment Methods for Single- and Double-axis Tomography

David N. Mastronarde

1. INTRODUCTION

Accurate alignment of projection images is an important step in obtaining a high-quality tomographic reconstruction. Ideally, all images should be aligned so that each represents a projection of the same 3D object at a known projection angle. Inadequate image alignment will result in blurring or smearing of features in the reconstruction. The problem is made more difficult because exposure to the electron beam during the acquisition of a tilt series induces geometric changes in many samples (primarily plastic-embedded material). For any sample, but particularly for ones that are not

David N. Mastronarde • The Boulder Laboratory for 3D Electron Microscopy of Cells, Department of Molecular, Cellular, and Developmental Biology, Campus Box 347, University of Colorado, Boulder CO 80309, USA

rigid during imaging, a powerful method of image alignment uses the measured coordinates of fiducial markers through the series of images; these can be fit to equations that describe the image projection. Alternative methods are based on cross-correlation of images. The fiducial marker method has the advantage that it guarantees a consistent alignment among the images from the full range of tilt angles. It is also more easily adapted to correct for changes in the sample that occur during imaging.

Various formulations of fiducial alignment have been described; they differ in their levels of complexity and methods of solution and in whether they require each marker to be measured on all projections (Berriman *et al.*, 1984; Jing and Sachs, 1991; Lawrence, 1992; Penczek *et al.*, 1995). The formulation presented below reflects the implementation of fiducial alignment in the IMOD package (Kremer *et al.*, 1996) (http://bio3d.colorado.edu/imod), which provides a superset of features described elsewhere. This chapter will describe the basic method as well as some practical aspects of fiducial alignment, drawing on extensive experience with the method in the Boulder Laboratory for 3D Electron Microscopy of Cells. It shows how the method can be extended to preserve resolution in reconstructions of large, heterogeneous volumes (e.g. Marsh *et al.*, 2001). It also describes the role and limitations of fiducial markers in making reconstructions from tilt series taken around two orthogonal axes (Mastronarde, 1997; Penczek *et al.*, 1995) and the complementary role that cross-correlation can play in this process.

2. DEFINITION OF VARIABLES AND PROJECTION EQUATIONS

Figure 1A depicts the geometry of the projection during a tilt series. There are three coordinates systems to consider: that of the microscope, the specimen and the projection images. The coordinate system of the microscope coincides with that of the specimen before it is tilted; the axes in both will be referred to as x, y and z. The axes of the projection images are designated u and v. The fiducial points are described by the following:

n_T is the total number of fiducial points

$\mathbf{r}_j = (x_j, y_j, z_j)$ are the coordinates in the specimen of the jth fiducial point, $j = 1, \ldots, n_T$

$\mathbf{p}_{ij} = (u_{ij}, v_{ij})$ are the measured projection coordinates of the jth point in the ith view

n_i is the number of points measured in the ith view (not all need to be measured)

V_i is the set of points measured in the ith view

$j \in V_i$ means that the jth point was measured in the ith view

The equations presented next represent a relatively complete model of the projection process, but in practice only a subset of the variables

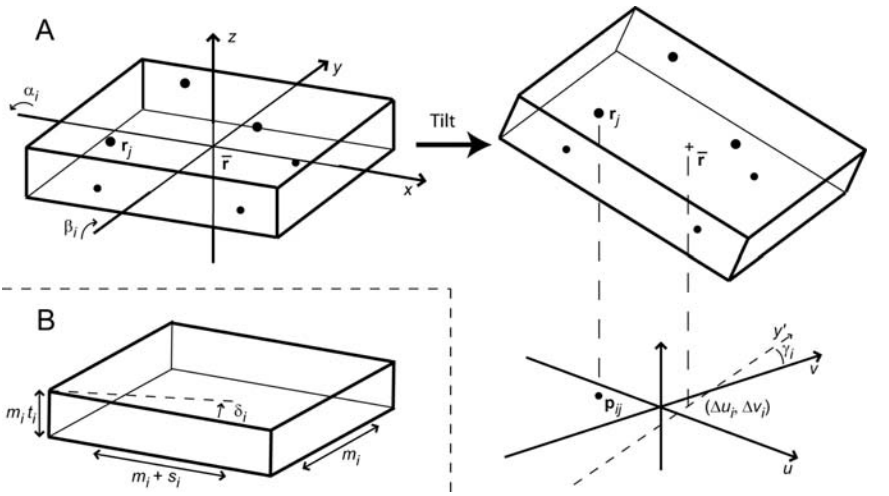


FIGURE 1. (A) Projection of a specimen. The coordinate system of the untilted specimen in the microscope is (x, y, z) while the coordinate system of the digitized image is (u, v). For the ith projection, the specimen may be tilted slightly by angle α_i about the x-axis as well as tilted by angle β_i around the y-axis. The projection image is rotated by angle γ_i and shifted by (Δu_i, Δv_i). (B) Specimen changes included in the model: m_i provides an overall scale change, t_i represents additional thinning, s_i provides for stretch along the x-axis relative to the y-axis, and δ_i represents skewing of the axes.

included in the model can be derived from measurements of fiducial positions in the projection images (see below). The model for the projection is that the specimen is tilted about the y-axis, the tilt axis of the goniometer, and may also be tilted somewhat about the x-axis. In the course of projection onto a detector, the image is also generally rotated and shifted. For the ith view:

α_i is the angle of tilt around the microscope x-axis, and $\mathbf{X}_i$ is the rotation matrix for this tilt:

$$\mathbf{X}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_i & -\sin\alpha_i \\ 0 & \sin\alpha_i & \cos\alpha_i \end{bmatrix} \tag{1}$$

β_i is the angle of tilt around the microscope y-axis, and $\mathbf{Y}_i$ is the product of a rotation matrix for this tilt and a matrix for projection into the u, v plane:

$$\mathbf{Y}_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} \cos\beta_i & 0 & \sin\beta_i \\ 0 & 1 & 0 \\ -\sin\beta_i & 0 & \cos\beta_i \end{bmatrix} = \begin{bmatrix} \cos\beta_i & 0 & \sin\beta_i \\ 0 & 1 & 0 \end{bmatrix} \tag{2}$$

γ_i is the angle of rotation around the microscope z-axis after projection into the u, v plane, with rotation matrix $\mathbf{R}_i$:

$$\mathbf{R}_i = \begin{bmatrix} \cos\gamma_i & -\sin\gamma_i \\ \sin\gamma_i & \cos\gamma_i \end{bmatrix} \tag{3}$$

$\mathbf{d}_i = (\Delta u_i, \Delta v_i)$ is the translation after projection into the u, v plane

Changes in the specimen during the recording of the tilt series can be modeled by three possible transformations: an overall size change, which is indistinguishable from a change in microscope magnification; an additional shrinkage along the specimen z-axis, referred to as thinning; and an additional anisotropic shrinkage along an axis in the x–y plane of the specimen. Such anisotropic shrinkage is the same as a negative stretch, so it will be referred to as stretch. There are several equivalent ways of representing this change; Fig. 1B shows one way in which it is expressed in terms of changes in the length and orientation of the specimen x-axis relative to the y-axis. For the ith view:

m_i is the overall scaling factor
s_i is an additive factor such that $m_i + s_i$ is the scaling along the specimen x-axis
δ_i is the skew angle, a rotation of the specimen x-axis toward the y-axis
t_i is a multiplicative factor for the thinning along the specimen z-axis
$\mathbf{D}_i$ is the matrix representing all of these specimen changes:

$$\mathbf{D}_i = \begin{bmatrix} (m_i + s_i)\cos\delta_i & 0 & 0 \\ (m_i + s_i)\sin\delta_i & m_i & 0 \\ 0 & 0 & m_i t_i \end{bmatrix} \tag{4}$$

The overall projection matrix, $\mathbf{A}_i$, is thus the following product:

$$\mathbf{A}_i = \mathbf{R}_i \mathbf{Y}_i \mathbf{X}_i \mathbf{D}_i \equiv \begin{bmatrix} a_i & b_i & c_i \\ d_i & e_i & f_i \end{bmatrix} \tag{5}$$

where it is sometimes convenient below to refer to the six elements of $\mathbf{A}_i$ as a_i, b_i, etc. Given values for all of the unknown parameters just described, the fiducial coordinates implied by the projection model, $\mathbf{p}'_{ij} = (u'_{ij}, v'_{ij})$, are calculated from:

$$\mathbf{p}'_{ij} = \mathbf{A}_i \mathbf{r}_j + \mathbf{d}_i \tag{6}$$

From this, it is evident that $\mathbf{r}_j$ can be arbitrarily shifted, and $\mathbf{d}_i$ can be adjusted to give the same projection values. This degeneracy can be eliminated by fixing the origin of the specimen coordinate system at the centroid

of the fiducial points, which is implemented by making the last point not a variable but a sum of the others:

$$\mathbf{r}_{n_T} = -\sum_{j=1}^{n_T-1} \mathbf{r}_j \tag{7}$$

When the systems of equations is sufficiently constrained, as described in the next section, an appropriate subset of the unknown parameters can be found by minimizing the total error E, which is the sum of squares of the distances between projected and measured positions:

$$E = \sum_i \sum_{j \in V_i} \left((u'_{ij} - u_{ij})^2 + (v'_{ij} - v_{ij})^2 \right) \tag{8}$$

The displacements $\mathbf{d}_i$ can be found and eliminated from the equations by taking the derivatives of E with respect to Δu_i and Δv_i and setting them to zero. For Δu_i:

$$\frac{dE}{d\Delta u_i} = \sum_{j \in V_i} (2(a_{11i}x_j + a_{12i}y_j + a_{13i}z_j - u_{ij}) + 2\Delta u_i) = 0 \tag{9}$$

$$\text{Since } \sum_{j \in V_i} \Delta u_i = n_i \Delta u_i,$$

$$\Delta u_i = \frac{1}{n_i} \sum_{j \in V_i} (u_{ij} - a_{11i}x_j - a_{12i}y_j - a_{13i}z_j) \tag{10}$$

Now define:

$\bar{\mathbf{p}}_i = \sum_{j \in V_i} \frac{\mathbf{p}_{ij}}{n_i}$, the centroid of fiducial points measured in the ith view, a constant, and

$\bar{\mathbf{r}}_i = \sum_{j \in V_i} \frac{\mathbf{r}_j}{n_i}$, the centroid of points projecting into the ith view, a sum of variables.

Then equation (10) and a similar expression for Δv_i can be written:

$$\mathbf{d}_i = \bar{\mathbf{p}}_i - \mathbf{A}_i \bar{\mathbf{r}}_i \tag{11}$$

Substitution into equation (6) then gives

$$\mathbf{p}'_{ij} = \mathbf{A}_i (\mathbf{r}_j - \bar{\mathbf{r}}_i) + \bar{\mathbf{p}}_i \tag{12}$$

Given values for the projection variables governing $\mathbf{A}_i$, the 3D coordinates $\mathbf{r}_j$ can be solved from linear equations such as equation (12). When the shrinkage is assumed to be isotropic, a closed-form solution is then available for each projection variable (Penczek *et al*., 1995), and a solution

can be obtained by iterating until convergence is achieved. In the more general formulation here, the remaining variables must be found by numerical methods for minimizing E. Since it is possible to take the derivative of E with respect to each variable, methods such as variable metric minimization (see Koval, 1997) or conjugate gradients can be used. Such minimization methods require adequate initial values of the variables to avoid being trapped in local minima. Tilt angles can be initialized to the nominal angles from the goniometer; values representing specimen changes can be set to 1 or 0 as appropriate. An initial value for the rotation after projection can be estimated from the apparent angle of the tilt axis in the images and does not need to be very accurate. Given these projection variables, initial values of the 3D coordinates and displacements can be estimated, which can be done by solving a single large set of linear equations (Brandt *et al*., 2001a). An alternative approach, used in IMOD, is to start with initial estimates of the $\mathbf{d}_i$ and solve for each $\mathbf{r}_j$ separately by a least-squares fit to equation 6, then compute $\mathbf{d}_i$ from equation 11, and iterate to convergence.

3. OBTAINING UNIQUE SOLUTIONS

In order to find a unique alignment solution, it is necessary to choose an appropriate subset of variables, which involves constraining the minimization by giving some variables fixed values, and avoiding combinations of related variables that can be covaried to give identical or near-identical solutions. All variables except the rotation angle after projection must be fixed to an appropriate value on one view in order to constrain the 3D fiducial coordinates. For example, if tilt angle were allowed to vary on all views, then there would be an infinite number of equivalent solutions with the tilt angles changed by a constant and the fiducials rotated about the y-axis by the same amount.

There are several relationships among the variables; the easiest one to see is the equivalence between thinning and a combination of stretch along the x-axis and a change of tilt angle. Just considering the product of distortion and y-axis tilt matrices, $\mathbf{X}_i\mathbf{D}_i$, the projected x-coordinate on one view is given by:

$$u_i = (m_i + s_i)\cos\beta_i x + m_i t_i \sin\beta_i z \tag{13}$$

However, for a given thinning t_i, if we change the tilt angle to:

$$\beta_i' = \sin^{-1}(t_i \sin\beta_i) \tag{14}$$

and change the x-axis stretch to:

$$s_i' = (m_i + s_i)\frac{\cos\beta_i}{\cos\beta_i'} - m_i \tag{15}$$

then it is equally true for all x and z that

$$u_i' = (m_i + s_i')\cos\beta_i' x + m_i \sin\beta_i' z \tag{16}$$

Thus, thinning can always be expressed with these equations as an x-stretch and a change in tilt angle, which means that it is impossible to solve for all three of these variables. This interplay between thickness, tilt angle and x-stretch also causes a problem when solving for tilt angle and x-stretch. If the tilt angle is fixed for the zero-tilt view, then the 3D fiducial coordinates can be scaled in z and the tilt angle and x-stretch values for all of the other views can be changed to be consistent with the scaled z values. This degeneracy can be avoided by constraining the x-stretch to zero in two views instead of just one.

When fiducials are confined to one plane in z, as happens when colloidal gold is deposited on only one surface of a section, even fewer variables can be solved for. Thinning cannot be detected at all and a change in tilt angle cannot be distinguished from a stretch along the x-axis.

In addition to these relationships, which can generate many exactly equivalent solutions, experience has shown that some variables are very difficult to solve for together because they can be varied jointly so as to generate nearly equivalent solutions. It is difficult to obtain an accurate solution for tilt around the x-axis when the rotation after projection is free to vary as well. When the fiducials are in only one z-plane, inappropriate solutions can occur if this rotation and the skew angle are both allowed to vary.

4. TRANSFORMATIONS TO ALIGN PROJECTION IMAGES

Once a solution is obtained specifying how each point in the specimen projects to each view, it is possible to reconstruct the volume, for example by back-projection. However, if the original images can be transformed so that they represent projections of a rigid body (i.e. an unchanging object) tilting around the y-axis (and possibly the x-axis also), then the computations can be much more efficient, because lines parallel to the x-axis can be back-projected from horizontal lines in the transformed images. If there is no tilt around the x-axis, then the computations are further simplified and reduced to a 2D task because each x–z plane of the reconstruction is computed from a set of lines at the same y value in the transformed images.

The image transformations to be applied are linear ones, of the form

$$\mathbf{p'} = \mathbf{Gp} + \mathbf{s} \tag{17}$$

where $$\mathbf{G} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \text{ and } \mathbf{s} = (\delta u, \delta v).$$

In fully aligned projections of a rigid body, the projection equation into the *i*th view is simply:

$$\mathbf{p} = \mathbf{Y}_i\mathbf{X}_i\mathbf{r} \tag{18}$$

In order for a transformation of the projections into a raw image (as given by equation 6) to be the same as projections from a rigid body,

$$\mathbf{G}_i(\mathbf{A}_i\mathbf{r} + \mathbf{d}_i) + \mathbf{s}_i = \mathbf{Y}_i\mathbf{X}_i\mathbf{r} \tag{19}$$

should be true for all 3D positions **r**, which requires that the matrices multiplying **r** be the same on both sides, and that the constant terms on the left sum to 0:

$$\mathbf{G}_i\mathbf{A}_i = \mathbf{Y}_i\mathbf{X}_i \tag{20}$$

$$\mathbf{s}_i = -\mathbf{G}_i\mathbf{d}_i \tag{21}$$

If changes in the section are isotropic, involving only the scale factor m_i, then $\mathbf{A} = m_i\mathbf{R}_i\mathbf{Y}_i\mathbf{X}_i$ and $\mathbf{G}_i$ is just $\mathbf{R}_i^{-1}/m_i$. However, in the general case, equation (20) cannot be solved for $\mathbf{G}_i$ because it involves 3×2 matrices and is thus equivalent to six equations with four unknowns. A solution that is valid for points in the plane $z = 0$ can be obtained by multiplying the matrix products on each side by $\mathbf{Z}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$ to convert them to 2×2 matrices. Thus

$$\mathbf{G}_i\mathbf{A}_i\mathbf{Z}_0 = \mathbf{X}_i\mathbf{Y}_i\mathbf{Z}_0 \Rightarrow \mathbf{G}_i = \mathbf{X}_i\mathbf{Y}_i\mathbf{Z}_0(\mathbf{A}_i\mathbf{Z}_0)^{-1} \tag{22}$$

For the case of tilt only around the *y*-axis, the transformation is:

$$\mathbf{G}_i = \begin{bmatrix} -e_i\cos\beta_i/(b_id_i - a_ie_i) & b_i\cos\beta_i/(b_id_i - a_ie_i) \\ d_i/(b_id_i - a_ie_i) & -a_i/(b_id_i - a_ie_i) \end{bmatrix} \tag{23}$$

Once $\mathbf{G}_i$ is found, equation (21) yields the displacements δu_i and δv_i that put the centroid of the fiducials at the origin. However, these displacements are generally not optimal for reconstruction. Typically, images need to be shifted to minimize the amount of area lost upon transformation and to place the material of interest in the center of the reconstruction in *z* (Ress *et al*., 1999). To shift the zero-degree view in *u* by u_s and to shift the reconstruction in *z* by z_s, the δu_i would be offset by $-u_s\cos\beta_i - z_s\sin\beta_i$.

5. LIMITATIONS OF IMAGE TRANSFORMATIONS

When there are anisotropic changes in the specimen, these transformations are accurate only for data from one *z*-plane of the specimen. This

problem is depicted in Fig. 2, which shows the projection of three lines that lie parallel to the x-axis and have the same y coordinate, but are placed at the top, middle and bottom of the specimen. In the absence of shrinkage, the projections of the lines are always superimposed at the same v coordinate, although their end-points are staggered in u, depending on the tilt angle (Fig. 2A and B). When the specimen shrinks along an oblique axis (Fig. 2C), the lines project onto oblique lines in the u–v plane, and the effect of specimen tilt is to separate their projections in u because of their different z-heights (Fig. 2D and E). Transforming the images according to equation (23) makes all of the lines horizontal, but they are now separated in v (Fig. 2D′ and E′). Notice that the separation in v changes sign between positive and negative tilt angles (compare Fig. 2D′ with E′). Thus, in aligned views of a specimen that has experienced such oblique shrinkage during the tilt series, fiducials on one surface will tend to move downward, and fiducials on the other surface will tend to move upward. Such effects are seen in some tilt series.

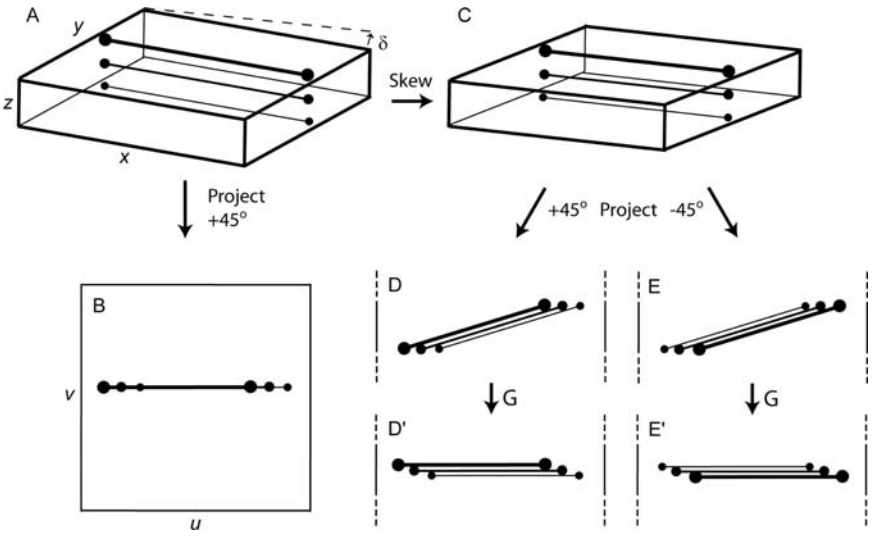


FIGURE 2. Illustration of the limitations of transforming projection images into alignment when there is anisotropic specimen shrinkage. (A) View of the original untilted specimen, with three lines located at different heights in one x–z plane. (B) Projection image from the original specimen at one tilt angle; the three lines are at the same v value at all projection angles. (C) View of the specimen after a skew of the x-axis relative to the y-axis. (D and E) Portions of the projection images from the skewed specimen at +45° and –45°, respectively. The three lines are offset from each other in u by tilt. (D′ and E′) Projection images after transforming images into alignment. The three lines are horizontal but at different v values.

To see how the projection position varies as a function of z in the specimen, consider a simplified situation where the images have already been transformed to account for translations, rotations and magnification changes. The matrix for projection into these images is $\mathbf{Y}_i\mathbf{D}_i/m_i$ to take into account the anisotropic changes, i.e.

$$\begin{aligned} a_i &= (m_i + s_i)\cos\delta_i \cos\beta_i / m_i \quad & b_i &= 0 \quad & c_i &= t_i \sin\beta_i \\ d_i &= (m_i + s_i)\sin\delta_i / m_i & e_i &= 1 \quad & f_i &= 0 \end{aligned} \tag{24}$$

The point (x, y, z) projects to $(a_i x + c_i z, d_i x + e_i y)$ in these images. The transformation from these images to fully aligned images is given by substituting the expressions from equation (24) into equation (23). When this transformation is applied to the images, this projection point is transformed to

$$(g_{11i}(a_i x + c_i z) + g_{12i}(d_i x + e_i y),\ g_{21i}(a_i x + c_i z) + g_{22i}(d_i x + e_i y)) \tag{25}$$

Substituting equations (23) and (24) into equation (25) and simplifying gives the projection position in the transformed images:

$$\begin{aligned} u' &= u + t_i \sin\beta_i (m_i / ((m_i + s_i)\cos\delta_i) - 1) z \\ v' &= v + t_i \tan\delta_i \tan\beta_i z \end{aligned} \tag{26}$$

where (u, v) is the rigid body projection position from equation (18). This result confirms the suggestion from Fig. 2 that when there are anisotropic changes in the specimen, it is not possible to transform images so that they can be treated like projections of a rigid body when reconstructing by back-projection. Rather, it is necessary to modify the back-projection position as a function of z in the reconstruction, using the coefficients of z in equation (26). Since a line of the reconstruction at one value of z and y back-projects from a single horizontal line of an aligned image, the computation including this modification can still be fairly efficient.

The effects of this correction are subtle for typical data sets, in which the coefficient of z in the factor added to v in equation (26) has a maximum value at high tilts of 0.005–0.02. The latter value means that in a reconstruction 200 pixels thick, the back-projection position needs to be adjusted by 2 pixels at each surface of the reconstruction. Without this adjustment, the point-spread function can show effects like those of Fig. 3, which presents three x–z slices through the average of 12 gold particles from one region of a reconstruction. Without the correction (upper row), the gold particle itself appears less symmetric, and the diagonal artifactual rays are strong in one direction on one side of the center (Fig. 3A) and strong along the other diagonal on the other side of the center (Fig. 3C). However, the typical effect of the adjustment is much less than shown here.

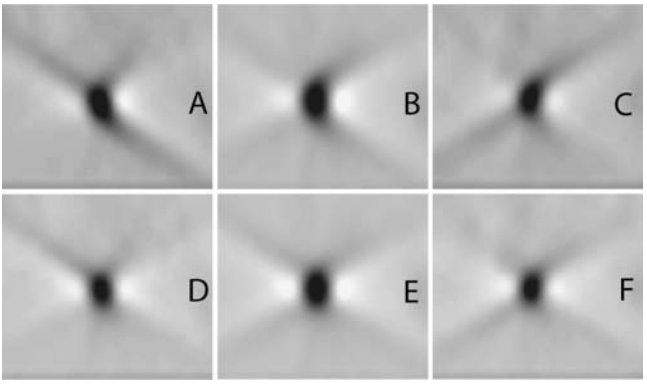


FIGURE 3. Slices in the *x*–*z* plane through an average of 12 gold particles from one region of a reconstruction. (A), (B) and (C) are from the reconstruction computed without the adjustments in equation (26); (D), (E) and (F) include the adjustments. (B) and (E) are through the center of the particle; (A) and (D) are slices 4.5 nm in front of the center, and (C) and (F) are slices 3 nm behind the center slice. The particles are from a reconstruction of a 250 nm section of a T-cell lymphocyte, courtesy of Mary Morphew.

6. CONSTRAINING VARIABLES ON ADJACENT VIEWS

Although the fiducial alignment method allows one to solve for an independent value of each variable for each view, there are circumstances where this is not desirable. When the ratio of the number of fiducial positions measured to the number of variables being solved for is not high enough, random errors in the fiducial positions can have an excessive effect on the solution and lead to a poor alignment. Moreover, some variables are intrinsically harder to solve for and are prone to this problem. Rotation and magnification are relatively easy to solve for, since a change affects the projection position substantially and unambiguously at any tilt angle. Tilt angle, thinning and the stretch variables are more difficult to solve for, especially at low tilt angles where changes may not have much effect on projection positions. Figure 4A shows an extreme case where solving for an independent tilt angle on every view yielded aberrant values near zero tilt. In this case, the tilt series itself showed no hint of a discontinuity in tilt angle or other aspects of its alignment.

The solution to this problem is to constrain variables that are expected to change slowly through the tilt series to have the same or similar values on a set of adjacent views. The simplest way to accomplish this constraint is to make a set of views have identical values for a parameter by assigning them all to one underlying variable to be solved. A smoother solution is obtained by setting up an underlying variable for only a subset of the views, such as every fifth view. For each of the remaining views, the

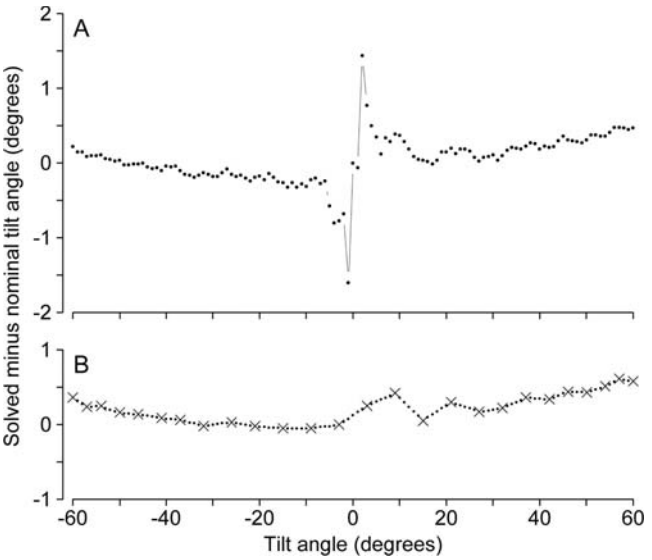


FIGURE 4. Example showing the value of constraining adjacent views to have similar values for some variables. (A) Solution for tilt angle with an independent variable for each view. (B) Solution using independent variables for one-fifth of the views, with the tilt angle constrained to change linearly between those views. The anomalous changes near zero tilt are prevented. The alignment included 15 markers confined to one surface of the specimen. The data set was from a 300 nm section of a *Caenorhabditis elegans* meiotic spindle and was acquired and aligned by Eileen O'Toole.

parameter is taken to be a linear combination of the values for the two surrounding views that have associated variables. The latter method is illustrated in Fig. 4B, where the crosses indicate views with associated variables. Either method will reduce the number of variables being solved for and will average over fiducial positions on several views in solving for each individual variable, thus reducing the effect of random errors on the solution. In Fig. 4B, the solution not only eliminates the aberrant values near zero tilt, but it also has reduced the view-to-view variability seen over most of the tilt range in Fig. 4A.

7. LOCAL ALIGNMENTS FOR LARGE AREAS

The solution from a fiducial alignment is based on the assumption that changes in the specimen during the tilt series occur uniformly across the area being analyzed. Non-uniform changes in the specimen violate this assumption and will result in a poorer fit to the fiducial positions, especially in the periphery of the area. These effects will become greater, the larger the area being reconstructed. To obtain the same quality of alignment in a

large reconstruction, it is necessary to break the reconstruction into smaller pieces in some way, where each smaller piece uses an alignment based primarily on the fiducials in that area. This could be accomplished in a number of ways; the procedures followed in the IMOD package are described here to illustrate some of the required and desirable features of an implementation.

1. A global alignment is obtained from all of the fiducials.
2. The area is divided into an array of local areas that overlap, typically by ~50%. Unless the fiducials are either very numerous or very evenly distributed, some areas may need to be expanded about their centers to encompass the minimum useful number of points. When stretching variables are to be solved for, the fiducials in each area must be distributed in z; this is implemented by requiring a minimum number of points from each of two surfaces.
3. The variables to be solved for are expressed as increments to the global solution. Thus, all variables are expected to change slowly and need to be solved on only a small subset of views (see Section 6). This feature allows a high ratio of measurements to variables even with a small number of fiducials in a given local area.
4. Geometric consistency with the global solution is desirable, particularly if the reconstruction is to be merged with one from a tilt series around an orthogonal axis. Consistency can be guaranteed by fixing the 3D coordinates of the fiducials at their global values, but adequate solutions may be obtained without this constraint.
5. The solution in a local area is used to derive a set of refining image transformations, which would be applied after aligning images with the transformations from the global solution.

A seamless reconstruction can be built from a set of local alignments by the procedure depicted in Fig. 5. The back-projection program takes as its input the globally aligned images, and the refining transformations and tilt angles for each local area. In Fig. 5, the crosses mark the centers of the local areas. Each column of points in the reconstruction is located at a particular position with respect to up to four nearby local areas. The point in the slice in Fig. 5 is located in a column that is close to two local area centers and farther from two others. This point is reconstructed by back-projecting from the appropriate position in each of the aligned images. The projection ray and filled circles indicate the projection position at one tilt angle before modification by the local transformations. The refining transformations for each of the four local areas are used to find modified projection positions (open circles). These positions are resolved by interpolation using weights based on the distance of the point's column from the centers of the local areas.

Local alignments can be helpful for reconstructing areas greater than ~1 μm in extent. For a 2 × 2 μm reconstruction of plastic-embedded,

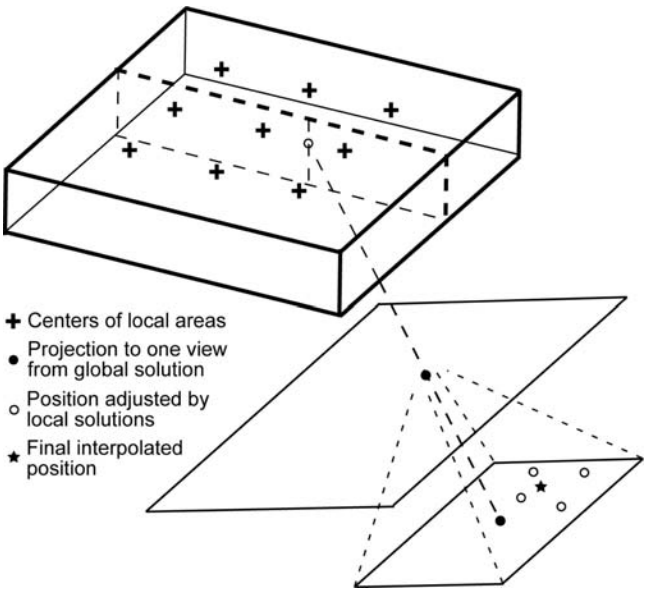


FIGURE 5. Illustration of how a seamless reconstruction can be generated using a set of alignments based on fiducials in local areas. The diagram indicates how a position in the volume being reconstructed is mapped to a single point in the projection at one tilt angle. See text for details.

sectioned material with ~1 nm pixel size, a typical result is for the global alignment to have a mean error of ~1 pixel and for local alignments to give average errors of ~0.6 pixel. The most easily observed effect on the reconstruction is an improvement in the point-spread function, as revealed by the reconstructed gold markers in an x–z view having a more compact and symmetric shape (Fig. 6). Improvements are less apparent in an x–y view but can sometimes be seen, such as in the case of Fig. 6G and H.

Note that the use of local alignments, such as the correction for anistropic shrinkage in equation (26), eliminates the one-to-one relationship between x–z planes in the reconstruction and sets of lines at the same y value in the aligned images. This complication makes it difficult to apply algebraic reconstruction techniques (Gilbert, 1972; Gordon and Herman, 1971), which attempt to solve a set of linear equations relating the measured projection pixels to the unknown voxels in the volume. When each x–z slice projects to data at the same y value in all of the views, there is a separate set of equations for each slice. With the more complex alignments described here, nearby slices are linked and there is a single set of equations to be solved for the whole volume. While the problem is soluble in principle, it is now much less tractable.

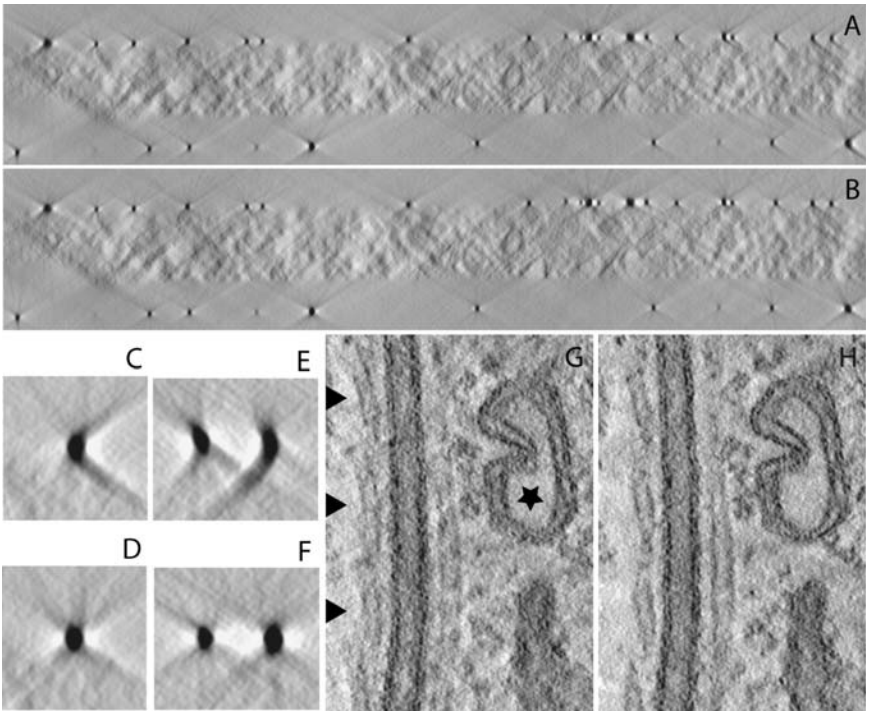


FIGURE 6. Effects of local alignments on reconstruction quality. The images are based on two reconstructions of trypanosomes in 250 nm plastic sections (±60° tilt series at 1° intervals, 2.3 × 6.5 μm area acquired at a 1.14 nm pixel size). Local alignments reduced the mean error from 1.50 to 0.90 for the reconstruction shown in (A–D), and from 2.68 to 0.74 for the reconstruction shown in (E–H). (A and B) Projections through 57 nm of reconstruction in the *x–z* plane, without and with local alignments, respectively. Note that the gold particles are crescent shaped in (A) and have strong artifacts facing to the left and right on the left and right sides, respectively. (C and D) One *x–z* slice through the center of a gold particle without and with local alignments, respectively. (E and F) From the reconstruction with greater global error, one *x–z* slice passing through the center of the particle on the right and off-center through the other particle. (G and H) An *x–z* slice showing that local alignment (in H) gives much better delineation of the separate microtubules (arrowheads) and better definition of the membranes around the structure marked with the star. The data set was acquired and reconstructed by Mary Morphew.

8. PRACTICAL CONSIDERATIONS

The error in an alignment solution can be considered to consist of two components, random and systematic. Random errors derive from inaccuracy in measuring the positions of the fiducial markers in the projection images. Two signs that random error is having an excessive effect on the solution are visible jitter from view to view when stepping through the

aligned images, and variation from view to view in solved values for variables such as magnification and rotation angle. For example, variations in solved magnification by >0.0005 from the long-term trend might indicate excessive random error. The effects of random error can be reduced by marking the fiducial positions more accurately or by adding more fiducials. The latter remedy averages the errors over more fiducials on a given view and keeps the solution from accommodating to the errors in a few fiducial positions.

Systematic errors arise when the fiducial positions are not adequately described by the alignment model being fit. The primary signs of systematic error are asymmetries in the gold particles in a reconstruction. When viewed with high contrast in the x–z plane, a sphere may appear crescent shaped rather than appearing elliptical with symmetric tails. When viewed in x–y planes above or below a particle, the artifactual tails may run diagonally rather than horizontally to either side of the particle, with the slope of the diagonal reversed between the top and the bottom of the particle. Another sign of systematic errors is a mean residual error higher than the best that can be achieved for the particular kind of specimen and imaging conditions. (A residual is the distance between a measured fiducial position and the projection position implied by the alignment solution.) In such cases, the residuals tend to be higher at high tilt. The remedies for systematic error are to fit the data to a more complex alignment model and to fit over smaller areas (e.g. with local alignments as described in Section 7).

The best alignment that can be expected will depend on several factors: the pixel size, the size of the gold fiducials, the thickness of the sample, the precision with which fiducial positions can be measured automatically and the amount of effort expended to reduce errors in their positions. Some generalizations based upon experience in the Boulder Laboratory, using the automatic fiducial tracking tools in IMOD, may be helpful. With relatively large pixel sizes (~2 nm) and images with information about spatial frequencies near the Nyquist limit (e.g. images digitized from film), mean residuals as low as 0.25–0.35 pixels may be achieved. For data acquired from 200–300 nm plastic sections on CCD cameras with a pixel size of ~1 nm, as is currently typical, mean residuals can be as low as 0.4–0.5 pixels. If data are taken with a smaller pixel size, but the specimen is still relatively thick, the mean residual in pixels can be expected to be higher, although perhaps smaller in nanometers. Tomography on cryospecimens also tends to give a higher mean residual (0.5–0.7 pixels).

Which parameters should be solved for depends on what kinds of changes might be occurring in the specimen and on the number and distribution of available fiducial markers. Magnification and rotation are probably the most important variables to solve for, if the specimen is at all susceptible to shrinkage. Solving for tilt angles can overcome inaccuracies in the goniometer readout and can compensate for a change in the angle of the irradiated part of the specimen relative to the rest of the grid. If fiducials are

well distributed in z, either specimen thinning or stretch can be solved for; a solution for stretch together with tilt angle is preferable because it can correct for thinning (see Section 3) as well for anisotropic shrinkage.

The inclusion of stretching variables usually gives only a modest improvement in the quality of the fit. For 26 recent tilt series from the Boulder Laboratory, adding the stretching variables reduced the mean residual by a factor of 1.07 on average (range 1.01–1.22). Most of the reduction was for points measured at high tilt angles; for example, for points measured at 50–60° the mean residual improved by a factor of 1.15 on average (range 1.02–1.50). Accordingly, comparison of reconstructions based on alignments with and without stretching variables show subtle improvements in the back-projection of high-angle information, similar in kind to the differences illustrated in Fig. 3. The improvement in the fit from adding stretching variables is partially masked by the fact that the image rotation variables alone can compensate for much of the misalignment caused by skew (shrinkage in the x–y plane along an oblique axis). Tests with model data indicate that the rotation can be varied to account for about two-thirds of the error caused by a skew. The error introduced by specimen stretching is thus somewhat greater than is indicated by the small reductions in mean residual achieved by explicitly accounting for this effect. Nevertheless, including stretching variables is not nearly as valuable as using local alignments for large areas, which can substantially reduce the residual error and noticeably improve the reconstruction.

It is preferable, but not always necessary to have fiducials distributed in z, e.g. on both surfaces of a section. The primary benefit of such a distribution is that the alignment should be accurate throughout the region being reconstructed because it is based on fiducial positions that span the region. With fiducials on one surface, the alignment must be extrapolated from that plane into the specimen, and systematic alignment errors will make this extrapolation progressively less valid the farther one is from that plane. While reconstructions can often be done successfully with fiducials on one surface, there is a risk that the alignment will not be adequate and that the quality of the reconstruction will visibly deteriorate near the other surface of the specimen. In addition to avoiding this risk, having fiducials distributed in z gives the secondary benefit of allowing more complex specimen changes to be corrected.

The number of fiducials required for alignment depends on the changes that need to be solved for. This requirement is best characterized by considering the ratio of the number of measurements to the number of variables being solved for, where each fiducial point contributes two measurements, an x and a y position, on each view. Unless fiducial positions are measured with a higher degree of precision than provided by the tracking tools in IMOD, a ratio of at least 4–5 seems to be required for a solution that is not impaired by random errors in the particle positions. This requirement translates to ~5 fiducials when solving for magnifications and

rotations only, or ~10 fiducials when solving for tilt angles and stretching variables, provided that the latter are solved on only a subset of views, as described in Section 6.

9. USING FIDUCIAL MARKERS IN DOUBLE-AXIS TOMOGRAPHY

Taking tilt series around two orthogonal axes and combining the results can significantly improve the definition of some features compared with those seen in a single-axis tomogram (Mastronarde, 1997; Penczek *et al*., 1995). The equations and procedures described above could be used to analyze all of the marker positions from the two tilt series together, but most implementations of double-axis tomography have aligned the two tilt series separately, then used the solved 3D fiducial coordinates to derive a transformation relating the two tilt series. One advantage of this approach is that inhomogeneous specimen changes accumulating over two tilt series will have less effect on the fit of the data and the accuracy of the alignment. When each tilt series is fit separately, each fit gives a lower error, in the same way that a piece-wise linear fit to a curving set of points will give less error than a single linear fit to all points. Penczek *et al*. (1995) used the 3D fiducial coordinates to derive a rigid rotation about the three axes, then applied this rotation to modify the projection angles for the second tilt series and computed a single reconstruction from the two series with a generalized weighted back-projection (Radermacher, 1992). The approach in IMOD is to reconstruct two separate volumes from the tilt series, transform one to match the other, then combine them. The 3D fiducial coordinates are used in only the first stage of a two-stage alignment procedure, which takes advantage of information in the specimen itself to achieve a better registration between the two reconstructed volumes.

The first stage in this procedure derives a general linear transformation from the 3D marker coordinates:

$$\mathbf{r}'_j = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \mathbf{r}_j + \mathbf{t} \tag{27}$$

where $\mathbf{r}_j$ and $\mathbf{r}'_j$ are coordinates of corresponding points in the second and first volumes, respectively, and $\mathbf{t} = (\Delta x, \Delta y, \Delta z)$ is the translation after transformation. Given a set of corresponding fiducial positions in the two volumes, the transformation can be found by three separate multiple linear regressions; e.g. a_{11}, a_{12}, a_{13}, and Δx are found from a least-squares fit of x'_j as a linear function of x_j, y_j and z_j, and similar fits yield the other components. At least four points are required for each fit, and at least twice that many are desirable to provide a more robust solution. Transformations found in this way generally reveal that a simple rigid rotation, even com-

bined with a uniform scaling, does not give a sufficiently accurate registration between the volumes (Mastronarde, 1997).

If fiducial points are confined to one surface, they will not provide sufficient information to solve for the coefficients of z in equation (27). In this case, one can solve for a transformation consisting of a linear transformation in the plane of the specimen (which includes the 90° rotation) and small rotations around the x- and y-axes:

$$\mathbf{r}'_j = \mathbf{XY}\begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & 1 \end{bmatrix}\mathbf{r}_j + \mathbf{t} \tag{28}$$

where **X** and **Y** are matrices for rotation about the x- and y-axes like those in equations (1) and (2). Here the equations are not linear, and a minimization search can be used to find the transformation.

These transformations, which are based on the fiducial coordinates, are not able to compensate for any non-linear changes in the volume between the two tilt series. For large volumes (e.g. >1 μm in extent), the fit between fiducial coordinates deteriorates, often giving a mean error of >1 pixel and larger errors in the periphery. This problem could be alleviated by fitting to local sets of fiducials, just as in the case of local alignments. However, it is preferable to use cross-correlation between the two reconstructed volumes, for two reasons: (i) the material in the volumes is generally better distributed than the gold fiducials, so displacements between the volumes can be measured at many more positions; and (ii) these correlations may give a better registration between the structures of interest than the fiducials, particularly if some of the latter are located on a support film that changes differently under the beam in comparison wih the embedded sample. The procedure for using correlations is:

1. Obtain a transformation from the fiducial coordinates by fitting to either equation (27) or equation (28).
2. Transform the second reconstruction into preliminary alignment with the first.
3. Cross-correlate corresponding subvolumes between the first volume and the transformed second volume at a 3D array of positions. For example, the subvolume size might be 80 × 80 × 40 pixels, and there might be three layers of subvolumes in z at 100 pixel intervals in x and y. An example of the displacement vectors found by correlation appears in Fig. 7.
4. The correlations yield a set of corresponding positions in the two volumes, which can then be fit according to equation (27) to obtain one refining transformation. For large volumes, this global fit is still not very good: for volumes ~2 μm in extent processed with a 1 nm pixel size, the mean error of the fit is typically 0.5–2.5 pixels. In these cases, it is necessary to fit the same equations to subsets of the corresponding positions that are

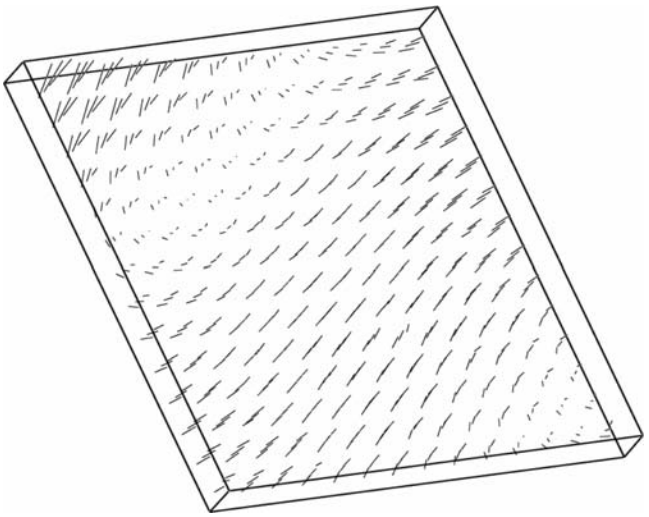


FIGURE 7. Vectors representing the local displacements between the two single-axis reconstructions of a double-axis data set, after the reconstructions have been registered using corresponding fiducial markers. The vectors are exaggerated in length by a factor of 15. For clarity, only one-fourth of the data used for obtaining local transformations is shown here. The reconstruction was a 4.2×4.2 μm area of a pancreatic beta cell obtained by Brad Marsh.

small enough to give a good fit. For example, equation (27) (or equation (28) if appropriate) would be fit to every set of $4 \times 4 \times 3$ corresponding positions. With this method of local fitting, the mean error can generally be reduced to 0.18–0.35 pixel. This procedure yields an array of refining transformations.

5. To produce the final transformed volume, each position in the volume is located relative to up to four transformations centered at surrounding locations, and a transformation is derived by interpolation from those surrounding transformations.

This warping procedure can produce a seamless transformation of the second tomogram, essentially undoing the specimen distortions that occurred between the first and second reconstructions. Provided that data are transformed only once, directly from the original volume, warping does not degrade the image quality any more than using a single linear transformation or rigid rotations, since data must be interpolated in any case.

One might wonder how the fact that the two volumes are derived from tilt series around orthogonal axes affects the correlations between them. A local cross-correlation can be computed by taking the 3D Fourier transform of each of the two subvolumes, forming the product of one with the complex conjugate of the other, and taking the inverse Fourier transform. The Fourier transform of each subvolume is missing a wedge of data due to the restricted

tilt range, and these wedges are orthogonal to each other, so the conjugate product of the transforms has two missing wedges. The primary effect of a missing wedge is to elongate structures in z. By extension from the fact that a missing pyramid or cone gives less elongation than a missing wedge (Mastronarde, 1997; Radermacher, 1992), one would expect a double missing wedge to produce greater elongation than a single wedge. The peak in these cross-correlations is thus probably substantially elongated in z. Despite this elongation, the highest point of the peak can still be localized accurately enough to give transformation fits with mean errors as low as 0.2 nm.

This method does have some disadvantages. The local correlation will not work in areas without sufficient material; such areas must be excluded from the analysis and they will consequently not align as well. Fortunately, features of interest are rarely located in these areas. The size of the subvolumes needed for reliable alignment depends on the density of features in the specimen and the resolution of the reconstruction, so it is difficult to make this procedure completely automatic. Nevertheless, when tools are provided for overcoming these drawbacks, local correlations give an intrinsically better registration between the two reconstructions from a dual-axis tilt series than do the fiducial marker positions alone.

10. CONCLUSIONS AND COMPARISONS

Alignment with fiducial markers has two main strengths. Because fiducial positions measured over the full range of tilt angles are fit to a single set of projection equations, the alignment of the images is guaranteed to be globally consistent. The method can be used to correct for anisotropic and non-uniform changes of the specimen during the tilt series, thus allowing reconstruction of large areas where such changes have a significant impact on the quality of the reconstruction.

The fiducial marker method has several practical disadvantages. It can be difficult to obtain an appropriate distribution of markers on the specimen. Colloidal gold particles are usually the densest material in a sample, so they introduce artifacts characteristic of back-projection into the body of the reconstruction, adding noise in the regions of interest. This effect can be alleviated by performing double-axis tomography, which reduces the artifacts by making them more circularly symmetric (Mastronarde, 1997). Another disadvantage of fiducials is the need to track their positions accurately, a potentially labor-intensive step. Several methods for automatically tracking fiducials have been developed (Brandt *et al.*, 2001b; Fung *et al.*, 1996; Ress *et al.*, 1999), suggesting that with concerted effort the labor involved in this step can be reduced to a reasonable level or nearly eliminated. Finally, depending on the complexity of the alignment, the use of fiducial markers requires some skill in the choice of parameters and the interpretation of results.

The main alternative to alignment by fiducials is alignment by cross-correlating successive projections (Frank and McEwen, 1992; Guckenberger, 1982), which can be refined by generating a tomogram and correlating with reprojections from it (Owen and Landis, 1996; Taylor *et al*., 1997; Winkler and Taylor, 2006). These methods face two major challenges. First, the location of the tilt axis in z after alignment by correlation is determined by the average z height of the features that contribute to the correlation, which can change through the tilt series. For example, suppose the main structure in the volume is a curving sheet similar to a 120° segment of a cylinder with its central axis parallel to the tilt axis, oriented so that the sheet is vertical in the middle of the volume and oriented at 60° near the top and –60° near the bottom of the volume. The projections of this sheet include a sharp line at all angles up to ±60°, but this sharp line will originate from the middle of the volume at zero tilt and from near the top or the bottom of the volume at the high tilt angles. With a correlation alignment, the tilt axis will shift progressively in z through the series. The second challenge is that a reprojection contains information from only a limited range (~15°) of the views contributing to a tomogram (a consequence of the central section theorem). It will therefore not provide significant information for correcting the aliynment of views that are systematically out of register with ones many degrees away. In the example of the curving sheet, correlating with reprojections will, to first order, simply recreate the misalignment.

The main advantage of alignment with fiducials is that it guarantees a globally consistent alignment that correlation methods have difficulty achieving. Indeed, a different approach to markerless alignment analyzes images to find localizable features that may then be fit to the alignment equations, potentially providing the same quality of global alignment as with fiducial markers (Brandt *et al*., 2001a; Chapter 6 of this volume).

ACKNOWLEDGEMENTS

This work was supported by NIH/NCRR grant RR00592 and NIH Program Project grant P01GM61306 to J. Richard McIntosh. I thank Pawel Penczek and J. Richard Mclntosh for helpful suggestions on the manuscript.

REFERENCES

Berriman, J., Bryan, R. K., Freeman, R. and Leonard, K. R. (1984). Methods for specimen thickness determination in electron microscopy. *Ultramicroscopy* **13**:351–364.

Brandt, S., Heikkonen, J. and Engelhardt, P. (2001a). Automatic alignment of transmission electron microscope tilt series without fiducial markers. *J. Struct. Biol*. **136**:201–213.

Brandt, S., Heikkonen, J. and Engelhardt, P. (2001b). Multiphase method for automatic alignment of transmission electron microscope images using markers. *J. Struct. Biol*. **133**:10–22.

Fung, J. C., Liu, W., de Ruijter, W. J., Chen, H., Abbey, C. K., Sedat, J. W. and Agard, D. A. (1996). Toward fully automated high-resolution electron tomography. *J. Struct. Biol.* **116**: 181–189.

Frank, J. and McEwen, B. F. (1992). Alignment by cross-correlation. In *Electron Tomography* (J. Frank, ed.). Plenum Press, New York, pp. 205–213.

Gilbert, P. (1972). Iterative methods for the three-dimensional reconstruction of an object from projections. *J. Theor. Biol*. **36**:105–117.

Guckenberger, R. (1982). Determination of a common origin in the micrographs of tilt series in three-dimensional microscopy. *Ultramicroscopy* **9**:167–174.

Gordon, R. and Herman, G. T. (1971). Reconstruction of pictures from their projections. *Commun. ACM* **14**:759–768.

Jing, Z. and Sachs, F. (1991). Alignment of tomographic projection using an incomplete set of fiducial markers. *Ultramicroscopy* **35**:37–43.

Koval, J. J. (1997). Variable metric function minimization. *Appl. Stat.* **46**:515–540.

Kremer, J. R., Mastronarde, D. N. and McIntosh, J. R. (1996). Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**:71–76.

Lawrence, M. C. (1992). Least-squares method of alignment using markers. In *Electron Tomography* (J. Frank, ed.). Plenum Press, New York, pp. 197–204.

Marsh, B. J., Mastronarde, D. N., Buttle, K. F., Howell, K. E. and McIntosh, J. R. (2001). Organellar relationships in the Golgi region of the pancreatic beta cell line, HIT-T15, visualized by high resolution electron tomography. *Proc. Natl Acad. Sci. USA* **98**:2399–2406.

Mastronarde, D. N. (1997). Dual-axis tomography: an approach with alignment methods that preserve resolution. *J. Struct. Biol.* **120**:343–352.

Owen, C. H. and Landis, W. J. (1996). Alignment of electron tomographic series by correlation without the use of gold particles. *Ultramicroscopy* **63**:27–38.

Penczek, P., Marko, M., Buttle, K. and Frank, J. (1995). Double-tilt electron tomography. *Ultramicroscopy* **60**:393–410.

Radermacher, M. (1992). Weighted back-projection methods. In *Electron Tomography* (J. Frank, ed.). Plenum Press, New York, pp. 91–115.

Ress, D., Harlow, M. L., Schwarz, M., Marshall, R. M. and McMahan, U. J. (1999). Automatic acquisition of fiducial markers and alignment of images in tilt series for electron tomography. *J Electron Microsc. (Tokyo)* **48**:277–287.

Taylor, K. A., Tang, J., Cheng, Y. and Winkler, H. (1997). The use of electron tomography for structural analysis of disordered protein arrays. *J. Struct. Biol*. **120**:372–386.

Winkler, H. and Taylor, K. A. (2006). Accurate marker-free alignment with simultaneous geometry determination and reconstruction of tilt series in electron tomography. *Ultramicroscopy* **106**:240–254.

6

Markerless Alignment in Electron Tomography

Sami S. Brandt

1. INTRODUCTION

In computing high-accuracy reconstructions from transmission electron microscope (TEM) tilt series, image alignment currently has an important role. Though most are automated devices today, the imaging systems have certain non-idealities which give rise to abrupt shifts, rotations and magnification changes in the images. Thus, the geometric relationships between

Sami S. Brandt • Helsinki University of Technology, Laboratory of Computational Engineering, PO Box 9203, FI-02015 TKK, Finland

the object and the obtained projections are not precisely known initially. In this chapter, *image alignment* refers to the computation of the projection geometry of the tilt series so that most of the above deviations from the assumed ideal projection geometry could be rectified by using simple 2D geometric transformations for the images before computing a tomographic reconstruction.

An accurate way to compute the image alignment is by using the conventional colloidal fiducial gold markers. Marker-based alignment methods are accurate, first, because gold beads can be localized very accurately due to their spherical shape and typically high contrast to the background and, secondly, the localization errors are averaged out if a large number of beads is used (see Brandt *et al.*, 2001b). However, it is not always possible to use markers. The markers, being electron-opaque, can interfere with the reconstruction (Frank *et al.*, 1987). Moreover, even if markers were sprinkled on the preparation, they might not be visible at all in an interesting part of the preparation, or there might be too few of them to compute the alignment parameters (Brandt *et al.*, 2001a). Also, with freely supported objects, the use of markers may be problematic (Liu *et al*., 1995).

In this chapter, we will consider automatic, marker-free image alignment methods designed to avoid the above problems of the marker-based alignment methods. In fact, different kinds of markerless approaches have been proposed. For instance, the cross-correlation-based alignment has been used for a long time to compute the translational alignment for tilt series (see, for example, Frank and McEwen, 1992; Frank *et al*., 1987; Guckenberger, 1982). A modern, completely different approach for the problem is feature-based alignment (Brandt *et al.*, 2001a; Brandt and Ziese, 2006), where automatically extracted points of interest are automatically tracked through the image series to solve the alignment parameters. In our taxonomy, the third category is formed by the 3D model-based methods where the projection geometry is refined with the help of an intermediate reconstruction (Brandt and Kolehmainen, 2004; Brandt *et. al*, in preparation; Dengler, 1989; Owen and Landis, 1996).

The organization of this chapter is as follows. The geometry of the alignment problems is first briefly described in Section 2. In Section 3, we review the cross-correlation-based alignment and the alignment by the common line. Section 4 is devoted to feature-based methods and Section 5 to 3D model-based alignment. In Section 6, we present some alignment examples, and concluding remarks are found in Section 7.

2. THE GEOMETRIC PROBLEM SETTING

For the TEM, a parallel projection model is usually assumed. The general parallel projection model is also known as an *affine camera model* and it includes the orthographic, scaled orthographic and para-perspective

projection models as special cases (see, for example, Xu and Zhang, 1996). A general affine projection is an affine mapping of a 3D coordinate vector **x** onto the image plane. The general affine projection model can be written in the form

$$\hat{\mathbf{m}} = \mathbf{A}\mathbf{x} + \mathbf{t}, \tag{1}$$

where $\hat{\mathbf{m}}$ is a 2×1 vector corresponding to the measured coordinates **m** on the image plane, **A** is a 3×2 matrix and **t** is a 2×1 translation vector.

Assuming that the rows of **A** are linearly independent, **A** can be decomposed into the form $\mathbf{A} = \mathbf{KPR}$ with the help of the RQ-decomposition (Hartley and Zisserman, 2000), where **K** is a non-singular upper triangular 2×2 matrix containing the *intrinsic* parameters of the affine camera (Faugeras and Luong, 2001), **P** is the 2×3 orthographic projection matrix and **R** is an orthonormal, 3×3 rotation matrix. Moreover, if we parameterize the rotation matrix by Euler angles, i.e. as a product of three basic rotations by the angles α, β and γ over the z-, y- and z-axes (or over the z-, x- and z-axes if the second tilt projections are considered), respectively, the model (1) takes the form

$$\hat{\mathbf{m}} = \mathbf{K}\mathbf{R}_{\alpha}\mathbf{P}\mathbf{R}_{\beta}\mathbf{R}_{\gamma}\mathbf{x} + \mathbf{t}, \tag{2}$$

where $\mathbf{R}_{\alpha}$ is a 2×2 rotation matrix, and $\mathbf{R}_{\beta}$ and $\mathbf{R}_{\gamma}$ are 3×3 rotation matrices (note, however, the discussion of the parameterization at the end of this section).

The intrinsic camera parameter matrix **K** generally has three parameters, which can be defined as

$$\mathbf{K} = s\begin{pmatrix} \xi & \tau \\ 0 & 1 \end{pmatrix} \tag{3}$$

where s is the scaling factor, ξ the aspect ratio and τ the skew parameter (Faugeras and Luong, 2001). The skew parameter allows a possible non-orthogonality of the axes of the image plane to be taken into account. The aspect ratio describes the possible difference in scaling along the x- and y-axes of the image plane. However, for a unique solution of the general form, a sufficient number of intrinsic parameters must be fixed (see Brandt and Palander, 2005; Quan, 1996). To the author's knowledge, skew has been formerly neglected in TEM imaging and, for many alignment schemes (both marker and markerless), the aspect ratio is set to 1, thus, the camera parameter matrix has the form $\mathbf{K} = s\mathbf{I}_{2\times 2}$.

We wrote the imaging model in the form (2) to illustrate the general nature of the alignment problem and how different alignment approaches relate to it. Using this construction, we consider the alignment problem as finding the *unknown* parameters of the imaging model for each projection image. Once the parameters are estimated, we may apply the following 2D

geometric transformation for the projection images, i.e. the coordinates of the points in the projection images are transformed by the formula

$$\mathbf{m}_{\text{aligned}} = \mathbf{R}_{\alpha}^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{m} - \mathbf{t}), \tag{4}$$

where $\mathbf{m}_{\text{aligned}}$ is the measured projection, corresponding to the 3D point $\mathbf{x}$, in the *aligned* coordinate frame, whose origin is the projection of the 3D origin. By using this 2D geometric transformation for the observed images, we 'undo' the 2D transformation part of the projection equation, i.e. the effect of the intrinsic parameters of the 'camera' and the 2D rotation.

The correction model above has practical implications. In single-axis tilting, for instance, one has used the projection model in a form where the tilt axis is the y-axis in the 3D coordinate frame (Lawrence, 1992), which corresponds to $\mathbf{R}_{\gamma}$ being the identity matrix in equation (2). This transforms the 3D reconstruction problem into a stack of 2D reconstruction problems, making the problem easily parallizable. The approach to alignment by cross-correlation has often considered an even more simplified setting such as only correcting the translational term $\mathbf{t}$, i.e. the rotation $\mathbf{R}_{\alpha}$ is assumed to be the identity matrix, and the rotation parameter is assumed to be fixed for all the images, or the rotations are solved by different means (in the single-particle reconstruction problem, 2D rotation alignment has been computed, e.g. by cross-correlation in the polar coordinate plane, see Section 3.4). Likewise, the magnification change and the other two intrinsic parameters are usually neglected in the cross-correlation alignment for tilt series (see, for example, Frank and McEwen, 1992), i.e. it is approximated that $\mathbf{K} = \mathbf{I}_{2\times 2}$.

On the other hand, for a double-tilt series, if we could assume that the tilt axes are precisely the y- and x-axes of the 3D coordinate frame, one could set $\mathbf{R}_{\gamma} = \mathbf{I}$, if we define $\mathbf{R}_{\beta}$ as the 3D rotation over the y-axis for the images in the first tilt and as the 3D rotation over the x-axis for the second tilt, where β is the corresponding tilt angle parameter. In practice, to take into account the possible non-orthogonality of the tilt axes (and the projection direction), it should be recommendable to use the complete 3D rotation model in the computations (for more discussion about the dual-axis tilting, see Section 4.3). By using the complete 3D rotation model in equation (2), some deviations from the assumption of a perfect tilt axis could also be taken into account in the single-axis geometry—however, if the complete, three-parameter rotation model is needed in either a single- or dual-axis alignment problem, it is *not* recommended to use the Euler angle parameterization, as in equation (2). This parameterization for 3D rotations causes singularities for certain configurations that may cause numerical problems in the parameter optimization (see Triggs *et al.*, 2000); a better way should be to use the quaternion parameterization (Kuipers, 2002) or some other parameterizations where these problems are avoided.

3. ALIGNMENT BY CROSS-CORRELATION AND THE COMMON LINE

The emphasis of this section is on the standard, cross-correlation-based alignment (Frank and McEwen, 1992; Guckenberger, 1982). Hence, Section 3.1 presents the background and principle of the cross-correlation method, and Section 3.2 considers its implementation in detail. In Section 3.3, we review the principle of the common lines method as it has been proposed for the alignment of TEM tilt series (Liu *et al.*, 1995). In Section 3.4, we discuss the problems and deficiencies of the cross-correlation and the common lines alignment methods.

3.1. The Basic Idea of Cross-correlation Alignment

The cross-correlation alignment is based upon the use of the discrete 2D cross-correlation function that we here define in the form (see, for example, Gonzalez and Woods, 1993)

$$h(m,n) = \frac{1}{MN} \sum_{j=0}^{M-1} \sum_{k=0}^{N-1} f(j,k) g(j+m, k+n), \tag{5}$$

where f and g denote the optical density measurements of two images and M and N denote their width and height (see the discussion below about zero-padding). If f and g are similar to each other up to a shift of $\mathbf{r}_0 = (m_0, n_0)$, the cross-correlation array h will have a maximum at that location. The cross-correlation alignment problem principally means identifying these maximum locations pairwise for the images in the tilt series, and these locations are taken as the estimates of the relative shift parameters. Once the relative shifts have been found for a sufficient number of image pairs, the image set can be aligned by identifying the relative shift between the reference view and each image, and then translating the images by the corresponding negative shift values so that the relative shifts become zero.

3.2. Implementation of Cross-correlation Alignment

In practical implementations, certain image processing tasks must be used to obtain robust and fast computations. Such tasks include, for example, image pre-processing, computation of the Fourier transform and peak detection. The basic algorithm can be modified in many ways, so in the following we try to discuss the general idea of the algorithm and the related implementation issues from our perspective rather than trying to describe it exactly in the form it has been proposed in the literature. A flow chart of the idea of the cross-correlation-based alignment is shown in Fig. 1.

Cross-correlation alignment is based on pairwise processing of the images to compute the relative shift between each image pair. In order to

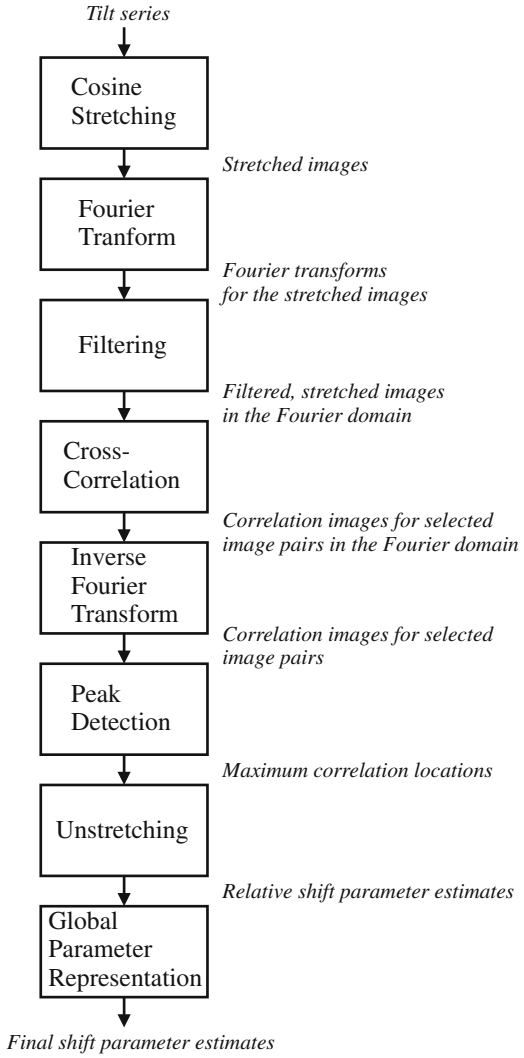


FIGURE 1. A schematic flow chart of cross-correlation-based image alignment. The algorithm is modified from the one presented in Frank and McEwen (1992). Note, however, that in practical implementations, it is worth processing the images pairwise to compute the relative shifts between the images pairs, not all together as the chart might suggest. The final shift estimates, with respect to the reference view, are then formed from the relative shift between the selected image pairs.

obtain a robust and accurate estimate for the relative shift parameter, i.e. a strong and narrow correlation peak, often only the neighboring images are compared in the tilt series. Alternatively, one could set one image as a fixed reference to which the relative shifts are measured (Saxton *et al.*, 1984)

or use additional shift measurements rather than those of only the nearest image pairs (Saxton, 1994). As soon as all the relative shifts between the neighboring images have been estimated, the overall shifts, with respect to the reference view, can be computed. In the following, it is described in detail how the relative shift between two images is computed.

To avoid wraparound and boundary effects, the representation of the images should be first considered. In signal processing, a conventional way would be to mean-correct the images (i.e. shift the gray values so that the mean gray value becomes zero), use a suitable window function and extend the modified images by zero-padding. If the original images are large enough, the wraparound error can also be avoided by selecting a smaller rectangular window, which is correlated with the other image, while those regions where wraparound occurs are finally neglected. Alternatively, other versions of the correlation such as the mutual or phase-doubled correlation function (Saxton, 1994; Taylor *et al.*, 1997; van Heel *et al.*, 1992; Winkler and Taylor, 2003) could be used. As a pre-processing technique, low-pass and high-pass filtering have also been used to reduce noise sensitivity (Frank and McEwen, 1992; Frank *et al.*, 1987) and to suppress the effect of the shape of the specimen (referred to as 'rooftop' effect, see Saxton and Frank, 1977).

A common practice in the cross-correlation alignment is to stretch the images by the factor of $1/\cos(\beta)$ in the perpendicular direction of the tilt axis, where β is the assumed tilt angle of the projection in question (Guckenberger, 1982). The stretching is performed to compensate for the fore-shortening of an object lying on the preparation, when viewed from different directions (see Fig. 2). The compensation thus assumes that the imaged object is flat and the viewing direction corresponding to the tilt of 0° is parallel to the sample plane normal.

Depending on the quality of the image data, the stretched images are then filtered by a suitable digital filter. The filtering is meant to enhance the relevant information of the images, to remove scanning artifacts, as well as to reduce noise sensitivity of the approach. It is worth performing the filtration in the Fourier domain as is done for the cross-correlation; see below.

Instead of computing the cross-correlation directly based on the definition given in equation (5), it is computationally more efficient to employ the correlation theorem. In Fourier space, we get the cross-correlation image for the pre-processed images from

$$H(u,v) = F(u,v)G^*(u,v), \tag{6}$$

where H, F and G are the discrete Fourier transforms, typically computed by the fast Fourier transform (FFT) algorithm. The correlation image h is obtained by the discrete inverse Fourier transform of $H(u,v)$.

To find estimates for the shifts, the peak detection is performed for the correlation image $h(m,n)$. The precise peak location can be computed in various ways but, in principle, the maximum location is searched, after

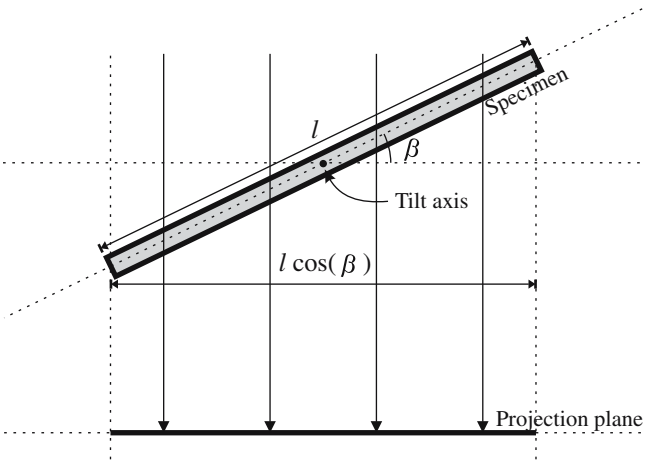


FIGURE 2. Schematic illustration of the image stretching to compensate for perspective distortions of a planar object when viewed from different directions.

which the center of gravity of the neighborhood of the maximum is computed to obtain a subpixel level estimate for the location. To compute the center of gravity, the neighborhood must be segmented from the correlation image. In practice, one may use circular or elliptical masks using the user-specified radius or lengths for the principal axes (Frank and McEwen, 1992). Alternatively, the segmented area could be found by thresholding the correlation image. After estimating the peak location, its coordinate values need to be unstretched by the factor of $\cos(\beta)$ in the perpendicular direction of the tilt axis, to invert the cosine stretching specified above.

3.3. Alignment by the Common Line

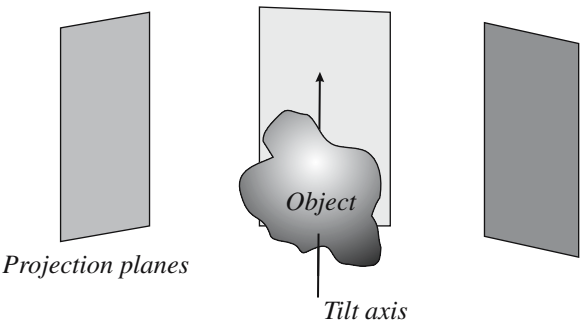
Another technique for determining the alignment for electron microscope tilt series is based on the concept of common lines (Crowther, 1971; Crowther *et al.*, 1970). The cross-correlation has been reported to be better suited for translational alignment than rotational alignment (Liu *et al.*, 1995); the common line approach can be used as an alternative way of determining the rotational alignment (see also Owen and Landis, 1996). Common lines techniques have been used frequently in estimating the projection directions to facilitate singe-particle reconstructions (see, for example, Crowther *et al.*, 1970; Kenney *et al.*, 1997; Lauren and Nandhakumar, 1997; Lindahl, 2001; van Heel, 1987), but this section will focus on how the common line technique was proposed (Liu *et al.*, 1995) to be utilized in the alignment of a single-axis tilt series.

The common line technique is based on the Fourier slice theorem (see, for example, Kak and Slaney, 1988), which states that each 2D projection

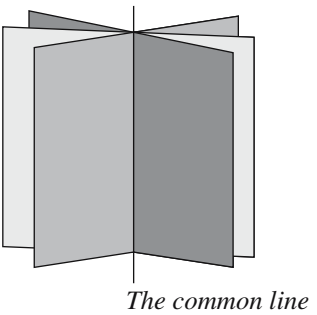
image corresponds to a 2D central section in 3D Fourier space. When the projections have been aligned, i.e. they have been put into the same spatial coordinate frame, the corresponding sections intersect on a single line in Fourier space, which is called the common line of the tilt series (Fig. 3). The idea of the common line alignment of Liu *et al.* (1996) is hence the common line in the projections, by using the projection theorem once again, and using the implied property that there is a 1D profile in the 2D projection images that is common to all the projections in the tilt series (Fig. 4).

Let $f(x,y)$ represent the piecewise continuous 2D projection image representing the ray attenuation measurement. The Radon transform R of the image is represented by

$$(Rf)(\rho,\varphi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(\rho - x\cos\varphi - y\sin\varphi)dxdy, \tag{7}$$



(a) Real space object with projection planes



(b) The Fourier space

FIGURE 3. Illustration of the relationship between the tilt series and the Fourier transform of the 3D object. The 2D Fourier transforms of the projection images correspond to a pencil of planes in the Fourier space that intersect at a single line called the common line.

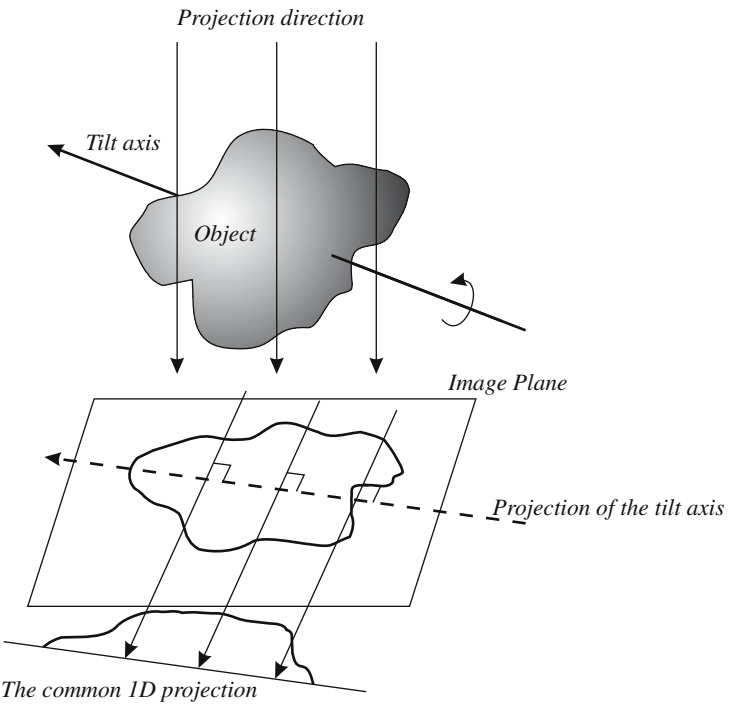


FIGURE 4. Sketch of the interpretation of the common line on the image planes of a tilt series: there is a 1D projection that is common for all projections in the tilt series. The 1D projection in the perpendicular direction to the projection of the tilt axis is the real space interpretation for the common line.

where δ denotes the Dirac delta function. The function obtained, Rf, is also called the *sinogram* of f. Consider the 1D density $g^i(\rho) = (Rf)(\rho - \rho^i, \varphi^i)$ corresponding to the image i. The common 1D projection for each view is defined by the parameter pairs (ρ^i, φ^i), for which ideally the relationship holds

$$g^1(\rho) = g^2(\rho) = \ldots = g^N(\rho) \text{ for all } \rho; \tag{8}$$

hence, the alignment problem is posed as finding these parameter pairs for each view.

In a practical implementation, one has to use discrete sinograms and take into account the fact that the realizations for the common 1D projection will not be identical due to measurement errors. Liu *et al.* (1995) proposed an iterative scheme that alternates between the shift and rotation refinement. The relative shift of the first view is fixed as the reference ($\rho^1 = 0$), and the remaining shift parameters are updated by such estimates that maximize the correlation coefficient between the current estimate of the

1D reference profile and 1D profiles of the other views. Correspondingly, the rotation estimates are updated for each view by determining the angle that maximizes the correlation coefficient of the 1D common profile of a view and the mean profile of the other views. More details of the algorithm can be found in Liu *et al.* (1995).

3.4. Discussion of the Cross-correlation and Common Line Approaches

The primary advantage of the cross-correlation-based alignment is that it is fast and straightforward to implement. The correlation theorem and FFT are the algorithmic tools that speed up the practical computations. In addition, binning the images lightens the computational load, although it compromises on the achievable accuracy level. In general, the cross-correlation-based alignment provides moderate accuracy for thin samples when only translational alignment is needed.

Naturally, problems occur when the assumptions of the method are not met. The cross-correlation alignment implicitly assumes a planar object, and the accuracy is adversely affected, if the sample thickness is not negligible. Usually there are also other inaccurate parameters, in addition to the shift parameters, that need to be corrected. Not only in-plane rotations and magnifications but also the tilt angles may be inaccurately known due to mechanical and optical inaccuracies of the imaging system. Moreover, since the images are compared pairwise in the correlation-based alignment, the accumulation of the error in the estimated parameters has been another problem in the cross-correlation-based alignment.

In principle, one could extend the cross-correlation alignment to handle rotation and magnification changes, but a direct extension results in the addition of the dimensionality of the search space. Therefore, iterative correlation-based translation and rotation alignment has been also considered (Frank, 1980) to take small rotational deviations into account. Alternatively, one could perhaps employ the 2D correlation-based alignment methods, used in the single particle analysis (see, for example, Cong *et al.*, 2003; Joyeux and Penczek, 2002; Penczek *et al.*, 1992); however, the 2D rotation corrections would be only relative between image pairs, i.e. it seems to remain problematic how to solve the absolute orientation of the tilt axis.

The common line alignment method for tilt series was reviewed above as an alternative method for the cross-correlation alignment. It should be noted, however, that the common line method reveals, in addition to the in-plane rotation parameter, only one shift parameter in the direction of the tilt axis for each image. The other shift parameter has to be known or found by different means. In the experiments by Liu *et al.* (1995), the other shift parameter was revealed by computing the center of mass of the object in the direction perpendicular to the tilt axis. The common line method for alignment of tilt series was unfortunately also reported to be unreliable

unless the specimen is freely supported so that the object has sharp boundaries and only faint background visible on the projections (Liu *et al.*, 1995).

To deal with the problems with the alignment methods above, several modifications have been proposed. Certain modifications to the correlation function were used, e.g. by Taylor *et al.* (1997) and Winkler and Taylor (2003); their alignment approach was based on area matching, however. A hybrid method for first solving rotational misalignment by fiducials was proposed by Frank *et al.* (1987). Owen and Landis (1996) combined the cross-correlation and the common line method, where they first solved the translational alignment by cross-correlation and then the rotational alignment by the common line alignment. However, Owen and Landis (1996) computed their final alignment using a different approach, based on the idea originally proposed by Dengler (1989) (see Section 5). In that iterative method, the alignment is refined by alternating the computation of reconstruction and correlation between the modeled projections and measured projections.

For general 3D objects and sample motions, the cross-correlation-based alignment is unfortunately invalid (see, for example, Brandt and Ziese, 2006). Likewise, the common line method also has problems as described above. Proper automatic alignment without fiducial markers may, however, be achieved by a completely different approach, based on feature tracking, that is the subject of the following section.

4. FEATURE-BASED ALIGNMENT

In this section, we discuss a modern approach to image alignment for electron tomography. Since the approach is based on feature tracking, it avoids most of the problems described above. On the other hand, the price is paid in the computational cost of the method.

4.1. Background

As accurate alignment can be achieved by using colloidal gold fiducial markers, a natural question is if there are some other points in the images that could be used as markers if conventional gold markers cannot be used. The answer is affirmative, and, in fact, the alignment problem is one form of the classical problem known as the *structure-from-motion problem* in the computer vision discipline, which has been intensively researched during the last two decades. The principal idea of such marker-free, feature-based alignment is to track certain points of interest in the images, points that can be automatically extracted from the images. A feature-based approach for TEM image alignment was proposed by Brandt *et al.* (2001a), and the work was further developed by Brandt and Ziese (2006). Flow graphs of these two feature-based approaches are displayed in Fig. 5.

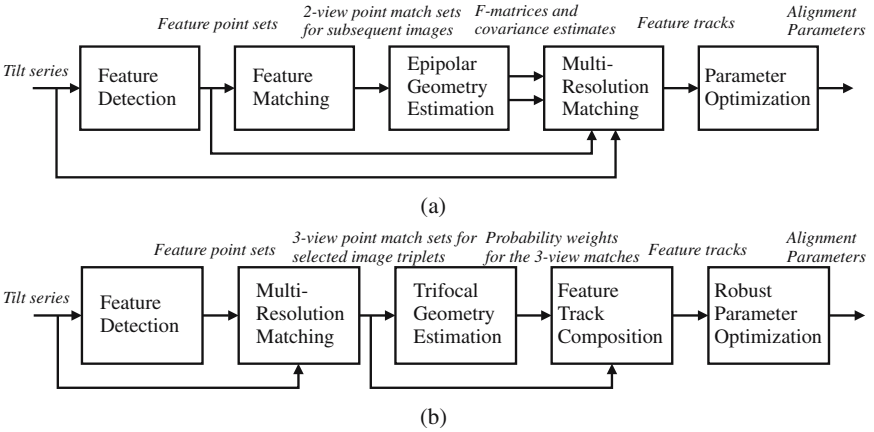


FIGURE 5. Flow graphs of the two feature-based image alignment methods. (a) Epipolar alignment, adapted from Brandt (2002). (b) Trifocal alignment.

The most demanding problem in marker-free, feature-based alignment is, however, finding correspondences between the extracted points such that the correspondence sets would be at least nearly free from mismatches. The previous approaches (Brandt and Ziese, 2006; Brandt *et al.*, 2001a) therefore have the following characteristics: first, initial feature-point correspondences between close-by image pairs or triplets are established by using the correlation of intensity neighborhood of the points, perhaps on several resolution scales, and certain constraints in the location where the corresponding feature point must lie in the other view(s). Secondly, the geometric constraints—either the epipolar or the trifocal constraints of close-by views—are estimated that are used to constrain the final correspondences, and hence, reliable, mostly correct point tracks should be obtained. As soon as reliable point tracks are found in this way, parameter optimization follows, also known as *bundle adjustment*, where the actual alignment parameters are estimated.

4.2. Description of the Method

Let us now take a closer look at the feature-based alignment method. In this section, we will, however, emphasize the overall idea, which we consider the most important, rather than trying to report all the details that can be found from the original works (Brandt and Ziese, 2006; Brandt *et al.*, 2001a) and the references therein.

4.2.1. Feature Point Extraction

The first step in the feature-based alignment is to extract points of interest from the images. Suitable points of interest should have repeatable

occurrence and be well localizable in the images. Standard features in geometric computer vision are the *Harris corner features* (Harris and Stephens, 1988; Schmid *et al.*, 2000), which are computed from the local autocorrelation function of the image intensity gradient. The cornerness measure is here defined as

$$C = \det \mathbf{R}_{xy} - k \operatorname{trace} \mathbf{R}_{xy} \tag{9}$$

where $\mathbf{R}_{xy}$ is the smoothed autocorrelation matrix

$$\mathbf{R}_{xy}(x,y) = \begin{pmatrix} (h * I_x^2)(x,y) & (h * I_x I_y)(x,y) \\ (h * I_x I_y)(x,y) & (h * I_y^2)(x,y) \end{pmatrix} \tag{10}$$

I_x and I_y are the x and y components of the image intensity gradient, and h is a 2D Gaussian smoothing kernel. The local maxima of the cornerness measure are taken as the points of interest, while the sensitivity of the operator can be adjusted by the parameter k.

4.2.2. Feature Matching

After the features have been extracted, the correspondences between the independently computed point sets of image pairs have to be found. In Brandt *et al.* (2001a,b), the Harris features were matched comparing the intensity neighborhood of the feature points by correlation, and a more or less consistent set was found by employing the relaxation algorithm of Zhang *et al.* (1994). However, to our present knowledge (Brandt and Ziese, 2006), a much better way is to use multiresolution cues of the intensity neighborhood of feature points. The principle of the latter algorithm is hence summarized below.

Having an image pair and a set of feature points for both images, a multiresolution pyramid is formed by down-sampling the images by the factors of 2, 4, 8 and 16. For each feature point in one image, only those feature points in the other view are selected as possible matching candidates that lie inside a selected search area. The intensity neighborhoods of the point in the first view and all the matching candidate points in the second are compared by computing the normalized correlation coefficients (Gonzalez and Woods, 1993) on all the resolution scales. Those point pairs are taken as the putative correspondences that have the maximum mean correlation scores over all the resolution scales and for which the correlation scores exceed certain threshold values on each scale. It is additionally required that the selected point pair must also be found if the order of the matched image pair is swapped.

4.2.3. Estimation of the Geometric Constraints

In previous works, we have utilized the two-view, epipolar constraint (Brandt *et al.*, 2001a) and the three-view, trifocal constraint (Brandt and

Ziese, 2006) to create reliable point tracks from point correspondences. These geometric constraints can be represented by a tensor, which is a 3×3 matrix $\mathbf{F}$ called the *fundamental matrix* in the two-view case, and a $3 \times 3 \times 3$ tensor $\mathbf{T}$ in the three-view case, called the *trifocal tensor* (see, for example, Hartley and Zisserman, 2000). As soon as the geometric constraint is known, the projection matrices of the two or three views can be recovered, yet only up to a projective transformation, or affine transformation in the case of an affine camera.

In practice, knowing the two-view constraint tells the corresponding line in the other view where the correspondence for a point in the first view must be. In fact, the epipolar constraint can be represented by the bilinear equation $\mathbf{m}'^{\mathrm{T}}\mathbf{F}\mathbf{m} = 0$ which holds for any matching point pairs $\mathbf{m}$ and $\mathbf{m}'$ (here in the homogeneous form) in the two views. The three-view constraint is even stronger than the two-view constraint: for instance, the knowledge of the projections of a 3D point in two views and the trifocal constraint generally imply the location of the projection in the third view. As the trifocal constraint is richer that the epipolar constraint, there are several algebraic incidence or trilinear relations related to it (see Faugeras and Luong, 2001; Hartley and Zisserman, 2000).

Since the geometric constraints are initially unknown, they must be estimated from the views. The estimation is typically performed using putative point correspondences between the views (see also Brandt *et al.*, 2001; and Brandt and Ziese, 2006), but other types of correspondences, such as curve matches, could be additionally used. Since automatically determined correspondences almost unavoidably contain mismatches, the constraint estimation method should be robust against the occurrence of such mismatches. In the computer vision community, popular robust estimation methods are, for example, Random Sample Consensus (RANSAC) (Fishler and Bolles, 1981) and Least Median of Squares (LMedS) (see, for example, Xu and Zhang, 1996). In the two-view and three-view geometry estimation problem, we have used the Maximum Likelihood Robust Estimator (MLRE) (Brandt, 2002).

As knowing the two- or three-view affine tensor is equivalent to knowing the affine projection matrices of the views, up to an affine ambiguity, the least-squares solution for the affine geometry can be recovered by the classic factorization algorithm of Tomasi and Kanade (1992). In our recent work (Brandt and Ziese, 2006), we therefore used our robust extension (MLRE) of the factorization algorithm, utilizing the equilibration principle (Irani and Anadan, 2000; Mühlich and Mester, 2001), to recover the affine three-view geometry. The epipolar geometry could also be estimated in this way. In fact, we did not use the trifocal tensors directly after the estimation but used only the match probability estimates, which are given as a by-product of MLRE, for selecting those putative point match triplets that were consistent with the estimated trifocal geometries.

4.2.4. Feature Tracking

After the geometry constraints have been estimated, the feature tracks are formed from point correspondences that are consistent with the estimated geometric constraints. In Brandt *et al.* (2001a), a multiresolution algorithm was proposed that utilized the recovered epipolar geometry in establishing the final correspondences. The most recent work by Brandt and Ziese (2006) used another, relatively reliable multiresolution algorithm for computing the *initial* correspondences, so no further matching needed to be performed after the trifocal geometry estimation and the match selection.

In the epipolar alignment (Brandt *et al.*, 2001a), the images were pairwise, sequentially matched. However, to obtain more matches and longer tracks, one should not only process the images sequentially starting from the image pairs 1–2, 2–3, 3–4, . . . or image triplets 1–2–3, 2–3–4, 3–4–5, . . . but also process longer steps such as 1–3, 2–4, . . . or 1–3–5, 2–4–6 . . . In this way, as was done by Brandt and Ziese, a larger amount of reliable tracks may be established with a wide angular coverage. After matching all the selected view pairs or triplets, the obtained matches were only rearranged into point tracks from which the alignment parameters were finally optimized.

4.2.5. Parameter Optimization

The final phase in the feature-based alignment is the step where the actual alignment parameters are estimated; the problem is also known as *bundle adjustment* (see Triggs *et al.*, 2000). The same model $\hat{\mathbf{m}}^i_j$ for the measured feature point coordinates $\hat{\mathbf{m}}^i_j$ can be used as have been used with fiducial markers (Lawrence, 1983, 1992). We have used it in the form

$$\hat{\mathbf{m}}^i_j = s^i \mathbf{R}^i_\alpha \mathbf{P} \mathbf{R}^i_\beta \mathbf{x}_j + \mathbf{t}^i, \tag{11}$$

where s^i is the scaling parameter of the view i, $\mathbf{R}^i_\alpha$ is a 2×2 rotation matrix parameterized by the rotation α^i, $\mathbf{P}$ is the orthographic projection matrix, $\mathbf{R}^i_\beta$ is a 3×3 rotation matrix (around the y-axis) depending on the tilt parameter β^i, $\mathbf{x}_j$ is the 3D coordinate vector of the point j, and $\mathbf{t}^i$ is the translation vector.

We first compute the least-squares estimates for the unknown parameters by minimizing

$$Q(\mathbf{m}) = \sum_{i,j} \left(\mathbf{m}^i_j - \hat{\mathbf{m}}^i_j\right)^{\mathrm{T}} \left(\mathbf{m}^i_j - \hat{\mathbf{m}}^i_j\right), \tag{12}$$

by using standard optimization tools for non-linear least-squares. Recently (Brandt and Ziese, 2006), we have used the trust region method (Coleman and Li, 1996) as it has been implemented in Matlab Optimization Toolbox (Matlab, The MathWorks Inc., Natick, MA, USA). The least-squares cost

function is, however, not robust, i.e. for a gross error in the measurements such as false point matches, the alignment results easily deteriorate. Therefore, we refine the least-squares estimates by robust estimators.

A robust solution can be obtained, for instance, by M-estimators, where the squared cost function in equation (12) is replaced by another function of the residual that is less sensitive to outliers. Let us first rewrite the cost function in the form

$$Q(\mathbf{m}) = \sum_k \varepsilon_k^2 = \sum_k \rho(\varepsilon_k), \tag{13}$$

where ε_k denotes a scalar residual value. We replace the squared cost function $\rho(x) = x^2$ by another function of the residuals such as the Huber M-estimator (Huber, 1981; Xu and Zhang, 1996)

$$\rho(x) = \begin{cases} \dfrac{x^2}{2\sigma^2}, & \text{if } |x| \le k\sigma \\ k\left(\dfrac{|x|}{\sigma} - \dfrac{k}{2}\right), & \text{otherwise} \end{cases}, \tag{14}$$

where we have used the value 1.345, corresponding to the 95% asymptotic efficiency, for the tuning constant k. We have used the same optimization algorithm in minimizing the robust cost function as we used for the least-squares minimization. In Brandt and Ziese (2006), we additionally chose the heuristic to refine the solution by a few iterations with Cauchy, Welsch and Tukey M-estimators.

In the least squares optimization, we suggest initializing the scale parameters to unity, rotations to zeros, tilts to the assumed values, translations to the center of the images and the 3D coordinate vectors such that their x- and y-coordinates represent the centroids of the found projections and the z-coordinates are initialized to zero. As far as the parameterization is concerned, since we may set the location of the 3D origin to an arbitrary position, we fix it to the 3D location of the first point because this selection has certain advantages in the optimization (see Brandt and Ziese, 2006). Likewise, we suggest fixing the reference tilt to zero.

4.3. Discussion of the Feature-based Alignment

Feature-based methods provide an alternative approach for image alignment to facilitate electron tomographic reconstruction. The feature-based alignment is technically similar to the gold marker alignment, and therefore a good level of accuracy should be obtained. In both approaches, certain points are tracked through the image set and, once the correspondences have been established, the alignment parameters are optimized for the image set. The principal difference between the marker- and

feature-based methods is that the localization and matching of the features are more difficult in the latter case.

Features such as Harris corners normally localized less accurately than gold beads, but the decrease in accuracy can be compensated by using a much larger number of feature points than there would be markers available. Naturally, fully automatic feature matching makes mistakes, although the probability for a false match might be small. Therefore, it is important that the cost function (equation 13) is robust in the alignment parameter optimization. In the latest development, the trifocal alignment has shown even better results than were achieved by manual marker picking in some of our experiments. This is encouraging, especially considering the cases where the use of markers is neither possible nor preferable.

It might also be a good idea to use a robust cost function if object deformations are expected in the image series. In this way, small non-rigidity of the specimen or the specimen support might not cause appreciable inaccuracy, if the deformation is not on the area of interest, and the deformation could be automatically handled (Brandt *et al.*, 2001b). One could also introduce additional weights for certain features, stressing the most important area of the images so that the objects inside the area would be most accurately aligned.

Marker- and feature-based alignments could also be combined in a trivial way. In this case, since marker localization is normally better than that of feature points, one should replace the distance function in equation (12) by the Mahalanobis distance to put more weight onto the markers. In addition, types of features other than Harris corners could be considered. For instance, line and curve matches could be searched over multiple views, as in the way proposed by Schmid and Zisserman (2000) or Shan and Zhang (2002). However, in this case, the alignment method may become more complicated and more difficult to handle.

Compared with the cross-correlation alignment, there is greater flexibility in feature-based methods since it is straightforward to take almost arbitrary 3D geometries and projection models into account by modifying equation (11). For instance, this classic model (Lawrence, 1983) assumes that there are only two unknown degrees of freedom in the rotation matrix corresponding to one view. Due to non-ideality of the imaging device, one could instead optimize over all the three parameters of a general 3×3 rotation matrix. In this case, one should, however, be careful in the rotation parameterization, as discussed earlier in this chapter (see Triggs *et al.*, 2000).

Double-tilt tomography (Mastronarde, 1997; Penczek *et al.*, 1995) could also be implemented by feature-based methods. Instead of aligning the two tilts separately and solving the relative 3D geometries of the two tilt series in a later stage, as in Penczek *et al.* (1995) and in Mastronarde (1997), a direct way would be to optimize the alignments of the two tilt series simultaneously as one optimization problem. In practice, the correspondences between the two tilt series should be additionally established. These corre-

spondences could be found by using a similar matching procedure, as described in this section, for image pairs being in different tilt series but having close projection directions.

The computational cost of the feature-based methods is relatively high when compared with the traditional cross-correlation-based alignment methods. The bottleneck is in the multiresolution feature matching since a large number of point pairs must be compared, on several resolution scales, using the normalized correlation of the image intensity neighborhood. The computation can be reduced by using cross-correlation *pre-alignment*. If the images are pre-aligned and one additionally uses, at least implicitly, the cosine stretching, a smaller search window could be used, i.e. the region in the image where the matching point correspondence is searched could be set smaller.

Although the feature-based methods should in many cases provide an accurate alignment of conventional electron microscope tilt series of critical-point-dried samples or sections, there are cases where they may fail. Probable failure occurs, for instance, when the signal-to-noise-ratio of the image is poor. This is due to the fact that the Harris corner detector is not tolerant to a high noise level simply because image noise not only causes false responses but also devastates the good ones. Effectively, a large noise level may occur in negatively stained preparations and cryoimage series. Some of the problems can be avoided by using noise removal such as Wiener filtering (see Lim, 1990), but, as a rule of thumb, there should always be some detailed structure visible by eye in the images, to ensure that feature-based methods could work with the image set.

5. 3D MODEL-BASED APPROACHES

As discussed above, there are cases where neither cross-correlation nor feature-based methods give satisfactory alignment results. For some of these cases, the third family of solutions, according to our categorization, might provide a solution. In this section, we discuss methods that are based on the idea of building an intermediate 3D model of the object as a reference that is then reprojected onto the image plane to which the original projection images are aligned. At the end of the section, we discuss a unification of the approaches.

The idea of using the reconstruction as a model to refine the projection alignment is, in fact, relatively old. For instance, Saxton *et al.* (1984) used the estimated 3D model of a crystal to refine the cross-correlation alignment of the tilt series. In their approach, the initial alignment was obtained by cross-correlation, and the computation of the alignment refinement took place in Fourier space. The shift parameters were refined by comparing the plane section of the accumulated 3D model, at the appropriate angle in the Fourier space, with each projection image and finding the location of the correlation peak.

Dengler (1989) proposed a multiresolution or coarse-to-fine approach where the scale and rotation parameters, in addition to the translation parameters, were also estimated. The 3D models were created by using an extended filtered back-projection algorithm starting from the lowest resolution images from the Gaussian pyramid representation of the projections. The alignment parameters were refined by minimizing the approximate sum of squared errors between the modeled and the measured projections over the unknown alignment parameters, starting from the lowest resolution scale. Because the method is based on the coarse-to-fine approach, prealignment is not as crucial as it would be for an algorithm that used only the finest resolution.

The alignment method of Owen and Landis (1996) is based on a similar principle to that of Dengler (1989); their implementation was reported to be similar to those used in single-particle analysis. However, Owen and Landis (1996) computed their initial translation parameter estimates by cross-correlation (Guckenberger, 1982) and the rotation parameters by the common line method (Liu *et al.*, 1995). Moreover, the 3D reconstruction was computed using the simultaneous iterative reconstruction technique (SIRT) (see, for example, Kak and Slaney, 1988), as it has been implemented in SPIDER (Frank *et al.*, 1981a, 1996), and the alignment parameters were finally refined in an interative way using the 3D model and the correlation techniques used in single-particle analysis (Frank *et al.*, 1981b).

Winkler and Taylor (2003) also considered an iterative refinement of the alignment that is based on references that are computed form back-projected images. Their work was principally focused on thin specimens such as sections and 2D crystals, and the initial alignment is based on area matching, as proposed by Taylor *et al.* (1997). The refinement was carried out by maximizing the cross-correlation coefficient between the alignment reference and the aligned image, and different modifications of the correlation functions were considered in the work (Winkler and Taylor 2003).

Brandt and Kolehmainen (2004) proposed yet another principle for imaging geometry estimation from tomographic projections that can be seen to belong to the category of this section. The approach is based on the utilization of the Bayesian, statistical inversion theory to solve the motion parameters so that the result would be consistent with the imaging model and the reconstructed structure. In Brandt and Kolehmainen (2004), a simplified setting of 1D projections of a 2D object was considered. However, we are currently investigating an extension of this approach to electron tomography to solve the image alignment with 2D projections of a 3D object. The preliminary results are encouraging and they are to be published soon (Brandt *et al.*, in preparation).

Finally, we already see a certain unification in the development of alignment. The 3D model-based approaches have been used widely in single-particle reconstruction, also referred to as projection matching (Penczek *et al.*, 1994). In fact, Yang *et al.* (2005) recently proposed using the

quasi-Newton algorithm to optimize the reconstruction and orientation of the object simultaneously by constructing the gradient of the objective function, though partly numerically. Their approach is somewhat close to that of Brandt and Kolehmainen (2004), although the works by Brandt and Kolehmainen (2004) and Brandt *et al.* (in preparation) provide more general statistical treatment for the 3D model-based alignment problem.

To summarize, we believe that the model-based approaches for image alignment are promising because the image intensity measurements are all taken as they are and the unknown parameters including the 3D model are just fitted to conform, possibly with a statistically sound criterion, with the observations. There are at least two drawbacks to this approach: that they are relatively heavy computationally and that there is generally no guarantee for finding the global minimum of the cost function; the geometry parameter initialization therefore has an important role.

6. ALIGNMENT EXAMPLES

In this section, we show some alignment examples by using some of the methods introduced above. In the experiments, we use the same critical-point-dried microvillus as used in the experiments of Brandt *et al.* (2001a), which is a freely dangling detail of a whole-mounted critical-point-dried cell, prepared as described in Engelhardt (2000). An sample image from the tilt series is shown in Fig. 6. The tilt series was taken in 3° increments from –60° to + 6° with a voltage of 120 kV and a magnification of 50,000× on a film, and manually digitized by a scanner. The image size was 1200×2300 pixels, with the pixel dimension, on the object scale, of 5.1 Å.

The series is an example of a situation where an automated image alignment procedure is needed. The microvillus had been immunologically gold labeled by 3 nm (ezrin) and 6 nm (mucin) beads, but these were so densely clustered that marker alignment would have been very difficult and inaccurate in practice since, at best, only three markers could be manually picked from the set (Peter Engelhardt, personal communication). The set contained some additional gold particles of 1.4 nm, which were hardly visible in an individual projection image, but they provided an additional tool for us to inspect the quality of the reconstruction.

In Brandt *et al.* (2001a), it was reported that the epipolar alignment, in the form as it was proposed at that time, is not as accurate as marker-based methods. This was due to the fact that, in the epipolar alignment, the feature tracks were relatively short, the localization of the Harris corner points was less accurate than markers, and a robust cost function was not used in the parameter optimization. The alignment inaccuracy was visible in the microvillus reconstruction, shown in Brandt *et al.* (2001a), in that the reconstruction was clearly stretched in the depth axis direction. Since the epipolar alignment was not completely satisfactory at that time, it is interesting

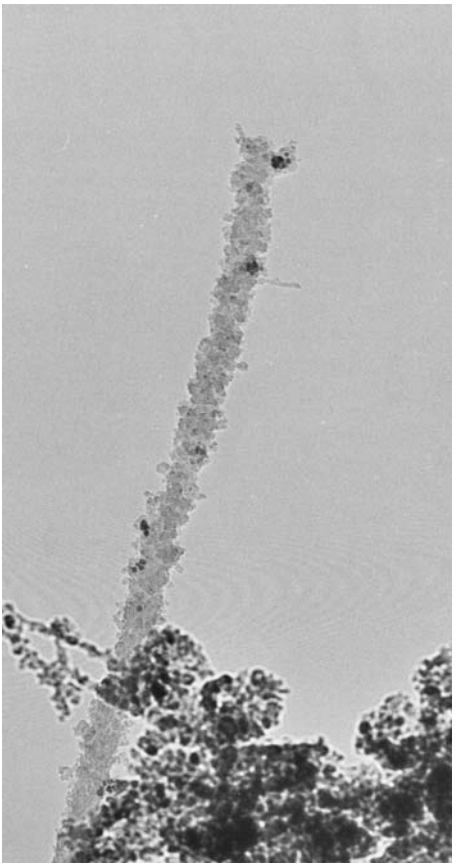


FIGURE 6. An sample image from the tilt series. The tilt axis approximately coincided with the vertical axis of the image. Courtesy of Dr Peter Engelhardt.

to experiment with the same tilt series again to compare the cross-correlation-based alignment, epipolar alignment as well as our feature-based state-of-the-art method, the trifocal alignment.

We first aligned the set by cross-correlation using the IMOD software package (Kremer *et al.*, 1996). We used the basic version of the algorithm to find the translational shifts, but manually hand-tuned the orientation of the tilt axis (one parameter, common for all the images in the set). Additional hand-tuning might have produced a better alignment, but we did not go in that direction since our preference is for fully automatic procedures. However, because the microvillus set had been taken on a film, which was manually digitized by a scanner, there was additionally relatively large deviation in tilt axis orientation, if compared with the rotation deviation one would typically expect with a CCD camera.

By both feature-based methods, more than 30,000 Harris feature points were successfully matched with both the epipolar and trifocal alignment

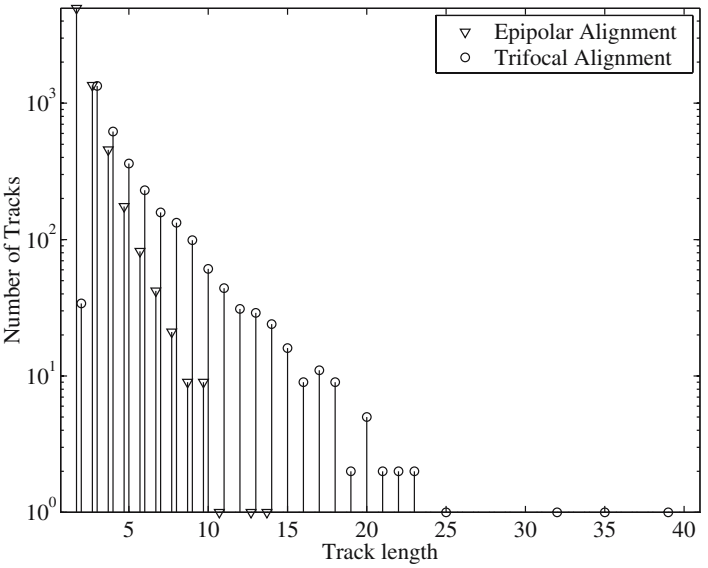


FIGURE 7. Histogram of the point track lengths for the microvillus image set using the epipolar alignment (triangle) and trifocal alignment (circle). In the trifocal alignment, the longest point track covers 39 of the 41 images in the set, whereas the length of the longest track is only 14 in the epipolar alignment. The two histograms have been slightly shifted away from each other to facilitate inspection.

methods. However, as the histogram of the track lengths shows in Fig. 7, the established tracks were much shorter with the epipolar alignment than with the trifocal alignment method. In addition, due to the earlier optimization setting, the shortest tracks with the lengths 2 and 3 had to be omitted in the epipolar alignment computations, so there were only 7716 measurements in the epipolar alignment that were finally utilized. In contrast, 32,242 measurements were utilized in the trifocal alignment where they satisfy a stronger geometric constraint; hence a much more accurate alignment could be expected.

We evaluated the three alignment methods by computing the filtered back-projection reconstructions of the microvillus by IMOD (Kremer *et al.*, 1996). In Fig. 8, there are six slice planes representing the reconstructed cross-sections from top to bottom of the microvillus. As can be seen, the feature-based alignment was able to give better reconstructions, of which the one obtained by the trifocal alignment is clearly the best. The alignments tended to be most inaccurate for the top and bottom slices, probably due to the fact that the measurements, on average, mostly represent the middle part of the object. In addition, in the third row of Fig. 8, the 1.4 nm gold particle is well distinguishable only at the trifocal reconstruction. In fact, our current experience with trifocal alignment is that the level of

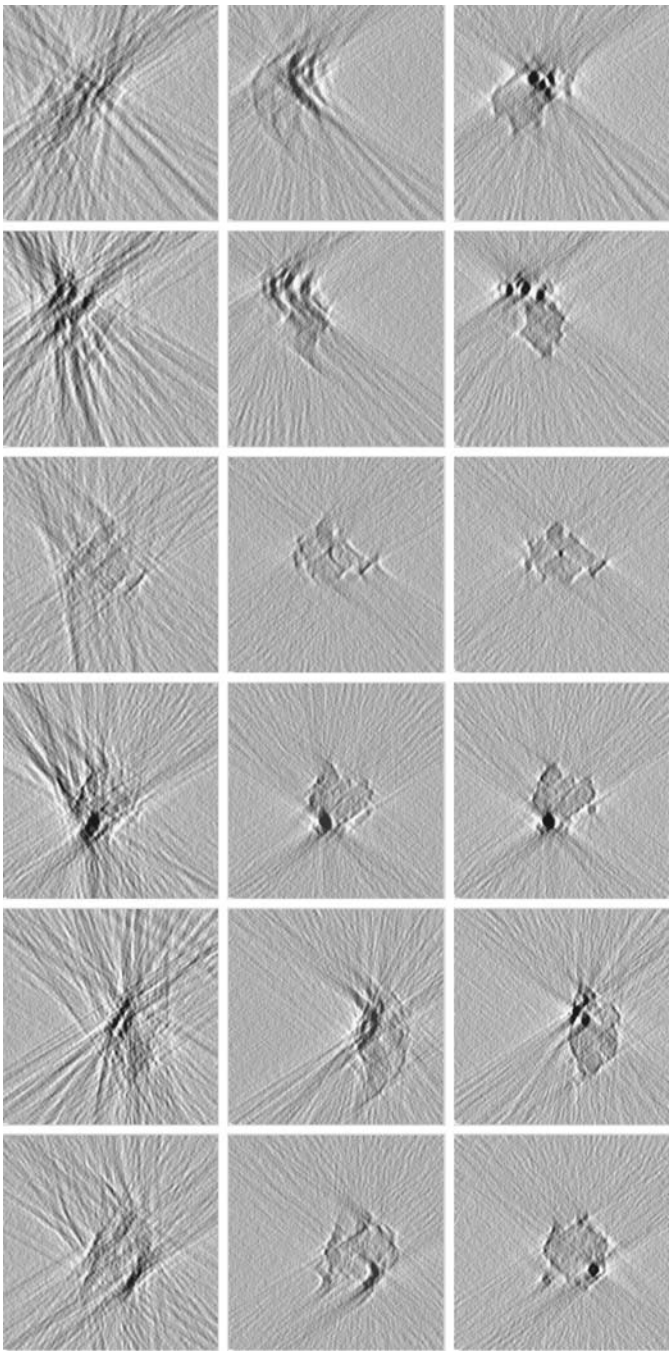


FIGURE 8. Cross-sections of the filtered back-projection reconstructions of the microvillus corresponding to the (left) cross-correlation alignment with hand-tuned orientation of the tilt axis, (middle) the epipolar alignment and (right) the trifocal alignment.

accuracy it provides is usually comparable with the level of accuracy that can be achieved by using fiducial markers. For certain samples, even hundreds of thousands of feature point measurements can be used (Brandt and Ziese, 2006) which should imply, if the assumptions made are valid, a very high level of accuracy because the estimation error decreases due to the averaging of the measurement noise.

7. CONCLUSIONS

In this chapter, we have discussed automatic image alignment in electron tomography. We started by reviewing the principles of conventional cross-correlation-based image alignment methods and the common line method as it has been proposed for the alignment of TEM tilt series. Then we continued by describing the second category in our taxonomy, i.e. the feature-based alignment approaches, after which we reviewed the third category for image alignment that we refer to as 3D model-based methods. For all the three categories, we discussed the strengths and weaknesses of the approaches. Finally, we showed some alignment examples. Let us now summarize the main points of this chapter.

The cross-correlation alignment method is one of the earliest image alignment methods in electron tomography. At present, it is computationally fast and straightforward to implement, at least when only translational alignment is needed. However, the method is usually found to be less accurate than fiducial-based alignment methods because of, among other things, sequential alignment and error accumulation. The method is also less appropriate for thick samples, because the 2D projections of an object are similar generally, after the cosine stretching, only if the samples are relatively flat. The common line method (Liu *et al.*, 1995) is an alternative for alignment of tilt series, where it can be used to estimate the plane rotation and the shift parameter in the direction of the tilt axis. The cross-correlation and common line methods have been additionally used together as a pre-alignment method (Owen and Landis, 1996).

Feature-based methods form a modern approach for automatic image alignment in electron tomography. The alignment problem can be seen as one form of a more general computer vision problem known as the structure-from-motion problem or, alternatively, as an extension of the fiducial-based techniques where automatically extracted features are used instead of fiducials. The feature-based alignment methods are currently based on the construction of reliable feature tracks over the image sequence using the available geometric constraints of a small number of views (Brandt *et al.* 2001a; Brandt and Ziese, submitted). When the feature tracks have been established, the total projection geometry of the image set can be solved, and images aligned.

At present, our experience is that the latest feature-based alignment (Brandt and Ziese, 2006) is usually able to give a competitive level of accuracy when compared with the fiducial alignment in single-axis electron

tomography of critical-point-dried specimens of whole mounts and sections. Cases where the feature-based alignment is not expected to work include at least the cases where feature detection fails, e.g. due to noise, or where the image intensity neighborhood does not give enough information to obtain a sufficient number of reliable point correspondences. Another possible case of failure may occur due to the inaccuracy of the geometry estimation or due to the final parameter optimization getting trapped in a local minimum.

As a third category of alignment methods for tilt series, we reviewed the methods that solve the imaging geometry by iteratively constructing a 3D model of the object that is reprojected onto the image plane and the model projection is matched to the measured projection. A common feature for the early approaches of this category has been that the model construction and the projection matching have been treated as independent problems. However, an approach involving simultaneous solving of projection geometry and the 3D model have also been proposed recently (Brandt and Kolehmainen, 2004; Brandt *et al*., in preparation; Yang *et al*. 2005). The model-based approaches are relatively complex computationally but we nevertheless deem the approach promising for obtaining accurate alignment, especially if a statistical point of view is considered.

As far as the relationship between the feature-based and 3D-model-based methods is concerned, it seems to us at present that both ways are promising. Both approaches have their pros and cons, but the methods in the first category also seem to have their place. For instance, cross-correlation pre-alignment can be performed on a routine basis and it may significantly simplify the computations in the more sophisticated methods. The future challenge will be to find a reliable and accurate method, especially for such cryosamples that are beyond the capabilities of the alignment methods currently available. Perhaps the 3D model-based approaches are going to be stronger in the alignment of such low signal-to-noise projections because to us it seems a better idea to rely on all the intensity measurements of the cryoprojection images than on a smaller number of individual feature points whose matching is currently relatively vulnerable to noise.

ACKNOWLEDGEMENTS

We would like to thank J. Frank, C. Renken, and P. Engelhardt for their comments. This work has been supported by the Academy of Finland, Project no. 129286.

REFERENCES

Brandt, S. (2002). Theorems and Algorithms for Multiple View Geometry with Applications to Electron Tomography. Doctoral thesis for the degree of Doctor of Science in Technology, Helsinki University of Technology.

Brandt, S., Heikkonen, J. and Engelhardt, P. (2001a). Automatic alignment of transmission electron microscope tilt-series without fiducial markers. *J. Struct. Biol.* **136**:201–213.

Brandt, S., Heikkonen, J. and Engelhardt, P. (2001b). Multiphase method for automatic alignment of transmission electron microscope images using markers. *J. Struct. Biol.* **133**:10–22.

Brandt, S. S. and Kolehmainen, V. (2004). Motion without correspondence from tomographic projections by Bayesian inversion theory. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR 2004) Vol. I. Washington, DC, pp. 582–587.

Brandt, S. S. and Palander, K. (2005). A Bayesian approach for affine auto-calibration. In *Proceedings of the 14th Scandinavian Conference on Image Analysis*. Joensuu, Finland, pp. 577–578.

Brandt, S. S. and Ziese, U. (2006). Automatic TEM image alignment by trifocal geometry. *J. Microsc.* **222**:1–14.

Coleman, T. F. and Li, Y. (1996). An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim*. **6**:418–445.

Cong, Y., Kovacs, J. A. and Wiggers, W. (2003). 2D fast rotational matching for image processing of biophysical data. *J. Struct. Biol.* **144**:51–60.

Crowther, R. A. (1971). Procedures for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs. *Philos. Trans. R. Soc. B* **261**:221.

Crowther, R. A., Amos, L. A., Finch, J. T., De Rosier, D. J. and Klug, A. (1970). Three dimensional reconstructions of spherical viruses by Fourier synthesis from electron micrographs. *Nature* **226**:421–425.

Dengler, J. (1989). A multi-resolution approach to the 3D reconstruction from an electron microscope tilt series solving the alignment problem without gold particles. *Ultramicroscopy* **30**:337–348.

Engelhardt, P. (2000). Electron tomography of chromosome structure. In *Encyclopaedia of Analytical Chemistry* (R. A. Meyers, ed.), Vol. 6. John Wiley & Sons Ltd, pp. 4948–4984.

Faugeras, O. and Luong, Q.-T. (2001). *Geometry of Multiple Images*. MIT Press, Cambridge, Massachusetts.

Fishler, M. and Bolles, L. (1981). Random sample consensus. A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**:381–385.

Frank, J. (1980). The role of correlation techniques in computer image processing. In *Computer Processing of Electron Microscope Images* (P. W. Hawkes, ed.). Springer-Verlag, Berlin, pp. 187–222.

Frank, J. and McEwen, B. F. (1992). Alignment by cross-correlation. In *Electron Tomography: Three-Dimensional Imaging with the Transmission Electron Microscope* (J. Frank, ed.). Plenum Press, New York. pp. 205–213.

Frank, J., McEwen, B. F., Radermacher, M., Turner, J. N. and Rieder C. L. (1987). Three-dimensional tomographic reconstruction in high voltage electron microscopy. *J. Electron Microsc. Tech.* **6**:193–205.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M. and Leith, A. (1996). SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol*. **116**:190–199.

Frank, J., Shimkin, B. and Dowse, H. (1981a). SPIDER—a modular software system for electron image processing. *Ultramicroscopy* **6**: 343–357.

Frank, J., Verschoor, A. and Boublik, M. (1981b). Computer averaging of electron micrographs of 40S ribosomal subunits. *Science* **214**:1353–1355.

Gonzalez, R. C. and Woods, R. E. (1993). *Digital Image Processing*. Addison Wesley.

Guckenberger, R. (1982). Determination of a common origin in the micrographs of tilt series in three-dimensional electron microscopy. *Ultramicroscopy* **9**:167–174.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151

Hartley, R. and Zisserman, A. (2000). *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Huber, P. J. (1981). *Robust Statistics*. Wiley.
Irani, M. and Anadan, P. (2000). Factorization with uncertainty. In *Proceedings of the 6th European Conference on Computer Vision*, Dublin, Ireland, pp. 539–553.
Joyeux, L. and Penczek, P. A. (2002). Efficiency of 2D alignment methods. *Ultramicroscopy* **92**:33–46.
Kak, A. C. and Slaney, M. (1988). *Principles of Computerized Tomographic Imaging*. IEEE Press.
Kenney, J., Karsenti, E., Gowen, B. and Fuller, S. D. (1997). Three-dimensional reconstruction of the mammalian centriole from cryoelectron micrographs: the use of common lines for orientation and alignment. *J. Struct. Biol.* **120**:320–328.
Kremer, J. R., Mastronarde, D. N. and McIntosh, J. R. (1996). Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**:71–76.
Kuipers, J. B. (2002). *Quaternions and Rotation Sequences: A Primer with Applications to Orbits, Aerospace and Virtual Reality*. Princeton University Press.
Lauren, P. D. and Nandhakumar, N. (1997). Estimating the viewing parameters of random, noisy projections of asymmetric objects for tomographic reconstruction. *IEEE Trans. Pattern Anal. Machine Intell.* **19**:417–430.
Lawrence, M. C. (1983). Alignment of images for three-dimensional reconstruction of non-periodic objects. In *Proceedings of the Electron Microscopy Society of Southern Africa*, Vol. 13, pp. 19–20.
Lawrence, M. C. (1992). Least-squares method of alignment using markers. In *Electron Tomography: Three-Dimensional Imaging with the Transmission Electron Microscope* (J. Frank, ed.). Plenum Press, New York, pp. 197–204.
Lim, J. S. (1990). *Two-dimensional Signal and Image Processing*. Prentice Hall, Englewood Cliffs, New Jersey.
Liu, Y., Penczek, P. A., McEwen, B. and Frank, J. (1995). A marker-free alignment method for electron tomography. *Ultramicroscopy* **58**:393–402.
Lindahl, M. (2001). Strul—A method for 3D alignment of single-particle projection based on common line correlation in Fourier space. *Ultramicroscopy* **87**:165–175.
Mastronarde, D. N. (1997). Dual-axis tomography: an approach with alignment methods that preserve resolution. *Journal of Structural Biology* **120**:343–352.
Mühlich, M. and Mester, R. (2001). Subspace methods and equilibration in computer vision. In *Proceedings of the 12th Scandinavian Conference on Image Analysis*. Bergen, Norway, pp. 415–422.
Owen, C. H. and Landis, W. J. (1996). Alignment of electron tomographic series by correlation without the use of gold particles. *Ultramicroscopy* **63**:27–38.
Penczek, P., Grassucci, R. A. and Frank, J. (1994). The ribosome at improved resolution: new techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles. *Ultramicroscopy* **53**:251–270.
Penczek, P., Marko, M., Buttle, K. and Frank, J. (1995). Double-tilt electron tomography. *Ultramicroscopy* **60**: 393–410.
Penczek, P., Radermacher, M. and Frank, J. (1992) Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy* **40**:33–53.
Quan, L. (1996) Self-calibration of an affine camera from multiple views. *Int. J. Comput. Vis*. **19**:93–105.
Saxton, W. O. (1994). Accurate alignment of sets of images. *J. Microsc.* **174**:61–68.
Saxton, W. O., Baumeister, W. and Hahn, M. (1984) Three-dimensional reconstruction of imperfect two-dimesional crystals. *Ultramicroscopy* **13**:57–70.
Saxton, W. O. and Frank, J. (1977). Motif detection in quantum noise-limited electron micrographs by cross-correlation. *Ultramicroscopy* **2**:219–227.
Schmid, C., Mohr, R. and Bauckhage, C. (2000). Evaluation of the interest point detectors. *Int. J. Comput. Vis.* **37**:151–172.
Schmid, C. and Zisserman, A. (2000). The geometry and matching of curves over multiple views. *Int. J. Comput. Vis.* **40**:199–233.

Shan, Y. and Zhang, Z. (2002). New measurements and corner-guidance for curve matching with probabilistic relaxation. *Int. J. Comput. Vis.* **46**:199–233.

Taylor, K. A., Tang, J., Cheng, Y. and Winkler, H. (1997). The use of electron tomography for structural analysis of disordered protein arrays. *J. Struct. Biol.* **120**:372–386.

Tomasi, C. and Kanade, T. (1992). Shape and motion from image streams under orthography: a factorisation approach. *Int. J. Comput. Vis.* **9**:137–154.

Triggs, B., McLauchlan, P., Hartley, R. and Fitzgibbon, A. (2000). Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice* (B. Triggs, A. Zisserman and R. Szeliski, eds), Vol. 1883 of *LNCS*. Springer, pp. 298–372.

van Heel, M. (1987). Angular reconstitution: *a posteriori* assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **21**:111–124.

van Heel, M., Schatz, M. and Orlova, E. (1992). Correlation functions revisited. *Ultramicroscopy* **46**:307–316.

Winkler, H. and Taylor, K. A. (2003). Focus gradient correction appied to tilt series image data used in electron tomography. *J. Struct. Biol.* **143**:24–32 .

Xu, G. and Zhang, Z. (1996). *Epipolar Geometry in Stereo, Motion and Object Recognition*. Kluwer.

Yang, C,, Ng, E. G. and Penczek, P. A. (2005). Unified 3-D structure and projection orientation refinement using quasi-Newton algorithm. *J. Struct. Biol.* **149**:53–64.

Zhang, Z., Deriche, R., Faugeras, O. and Luong, Q. (1994). A robust technique for matching two uncalibrated through the recovery of the unknown epipolar geometry. *Artif. Intell.* **78**:87–119.

7

Algorithms for Three-dimensional Reconstruction From the Imperfect Projection Data Provided by Electron Microscopy

Jose-Maria Carazo, Gabor T. Herman, Carlos O. S. Sorzano and Roberto Marabini

Jose-Maria Carazo • Centro Nacional de Biotecnología (CSIC), Universidad Autónoma, 28049 Cantoblanco, Madrid, Spain

Gabor T. Herman • Department of Computer Science, The Graduate Center, City University of New York, New York, NY 10016, USA

Carlos O. S. Sorzano • Escuela Politécnica Superior, Universidad San Pablo-CEU, 28668 Boadilla del Monte, Madrid, Spain

Roberto Marabini • Escuela Politécnica Superior, Universidad Autónoma, 28049 Cantoblanco, Madrid, Spain

1. INTRODUCTION

Since the 1970s, it has become increasingly evident that transmission electron microscopy (TEM) images of typical thin biological specimens carry a large amount of information on 3D macromolecular structure. It has been shown many times how the information contained in a set of TEM images (2D signals) can determine a useful estimate of the 3D structure under study.

In its most general form, the 3D reconstruction problem in TEM can be defined by the statement: given a collection of projection images (2D data) g, determine the 3D structure f that produced the images g. This problem has to be solved under conditions in which the image data, as well as the information about the geometry of data collection that relates g to f, are imperfect; in particular, both the gray level information in the images and the information regarding the direction of projections are corrupted by substantial noise. We are interested in knowing under which conditions g is adequate for producing an $\hat{f}$ that is close to f in some sense. For experimental reasons, we always have three basic limitations on the collection of the image data set g: the image gray level noise, the imperfect information about the data collection geometry and, finally, the restriction to a finite number of images. Can we do something to ameliorate the situation? Can we qualify our confidence in the reproducibility of the reconstructed structures (in other words, do small changes in g produce radical changes in the estimate $\hat{f}$?). Formulated in this way, it becomes clear that the topics covered under 3D reconstruction in TEM belong to the broad class of signal-recovery/inverse problems. It is within this framework that we discuss how the physical limits directly affect the fidelity of the estimated 3D structures produced by various reconstruction algorithms.

2. OVERVIEW OF OUR APPROACH

In this section, we present the concepts that we consider fundamental to a careful discussion of our topic. We illustrate these concepts as we go along on the relatively simple problem of recovering a 2D structure from its 1D images.

In the area of inverse problems, it is assumed that we have some understanding of the data collection process. Usually we describe it by some mathematical idealization, combined with a description of how the actual data collection process differs from the ideal one. For example, in the field of (2D) image reconstruction from (1D) projections, the ideal data consist of the line integrals, for all lines, of the image f to be reconstructed. In practice, we can have data corresponding to only finitely many lines, and the data are likely to be contaminated with noise (see Fig. 1). To approximate f from such data, there are two basic approaches that we call transform methods and series expansion methods, respectively.

FIGURE 1. A 341 × 341 pixel representation of a 2D image (left) and its 1D projections (right). Each projection consists of noisy versions of line integrals for 495 parallel lines (the distance between them is the same as the length of an edge of a pixel in the image); they correspond to individual columns in the picture on the right. The angle between the lines corresponding to one column and the next one is 2°; there are 70 columns corresponding to direction angles ranging from –69° to 69° with the vertical. The noise is additive zero-mean Gaussian noise, with a standard deviation that is the integral for a vertical line through the largest of the 12 central rectangular shapes in the image. (The image is approximately a cross-section through a 3D phantom, designed by Fernández *et al.* (2002), of mitochondria; the just discussed rectangular shapes are sections through the simulated cristae.)

The essence of transform methods is to find a mathematical procedure that describes the recovery of f from its ideal data, and then implement this procedure (as best as one can) making use of the actual data. In the case of 2D image reconstruction from 1D projections, the ideal data consist of all $p(l,\theta)$ that are the line integrals of f along lines at a distance l from the origin, making an angle θ in radians with the vertical. In that case, we have the formula (Herman (1980), p. 95)

$$f(r,\phi)=\frac{1}{2\pi^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\int_{-E}^{E}\frac{1}{r\cos(\theta-\phi)-l}p_1(l,\theta)\mathrm{d}l\mathrm{d}\theta \tag{1}$$

for f in polar coordinates with ϕ in radians, where $p_1(l, \theta)$ denotes the partial derivative of $p(l, \theta)$ with respect to l, and E is a number large enough so that $p(l, \theta) = 0$, for all $l > E$ and all θ. This formula is derived, based on the pioneering work of Radon (1917), in Section 16.3 of Herman (1980). The inner integral in equation (1) is referred to as a convolution, while the outer integral is referred to as a back-projection.

For the data collection method described in Fig. 1, we have only (noisy versions of) the data $p(nd, \Delta_m)$, where d is the length of an edge of a pixel

in the image on the left of Fig. 1, the n are integers in the range from –247 to 247, the m are integers in the range from –34 to 35, and Δ_m is $(2m - 1)^\circ$ in radians. We now discuss one way of finding an approximation $\hat{f}$ to the f in equation (1) based on such data. What we describe is referred to as the convolution method by Herman (1980); it is more popularly known as the filtered back-projection method (FBP), and it is very similar in nature to the weighted back-projection method (WBP) for the reconstruction of 3D structures from 2D projection images that is discussed below in Section 4.1.

To evaluate the convolution, we do the following. We define, for $-247 \le \bar{n} \le 247$ and $-34 \le m \le 35$,

$$p_c(\bar{n}d, \Delta_m) = d \sum_{n=-247}^{247} p(nd, \Delta_m) q((\bar{n} - n)d), \tag{2}$$

(see p. 128 of Herman, 1980), where q is the so called filter function. There are a number of ways to define q; they will lead to different reconstructions $\hat{f}$. In the illustration given here, we use

$$q(u) = \frac{2A}{\pi} \int_0^{\frac{A}{2}} \sin\frac{\pi U}{A} \cos(2\pi U u) dU, \tag{3}$$

see Herman (1980), p. 126, with the *Sinc* window from Table 8.1 chosen for F_A. Note that q depends on the parameter A. Choosing A small has a smoothing effect on the reconstruction; the ideal choice of A depends on the relative nature of the structure we are attempting to reconstruct (the signal) and of the data collection (including the noise).

The back-projection, for the data collection geometry described in Fig. 1, is approximated by

$$\begin{aligned}\hat{f}(r,\phi) = 11\Delta[p_c(r\cos(\Delta_{-34} - \phi), \Delta_{-34}) + p_c(r\cos(\Delta_{35} - \phi), \Delta_{35})] \\ + \Delta \sum_{m=-33}^{34} p_c(r\cos(\Delta_m - \phi), \Delta_m)\end{aligned} \tag{4}$$

where Δ is 2° in radians, and values of $p_c(r \cos(\Delta_m - \phi), \Delta_m)$ are evaluated from values provided by equation (2) using interpolation (typically linear). Bigger weights have been given to the first and the last angles in this approximation to the outer integral of equation (1), due to the fact that there is a big gap between these two angles for which no measurements are taken. (An alternative way of handling such missing data problems is discussed below in Section 4.1.)

The reconstruction based on equations (2–4) is shown on the left of Fig. 2. Pixel values were calculated by evaluating $\hat{f}(r,\phi)$ at the centers of the pixels. The value of A was selected to minimize the norm of the difference between the reconstruction and the original image in a central region; its optimal value is 0.1.

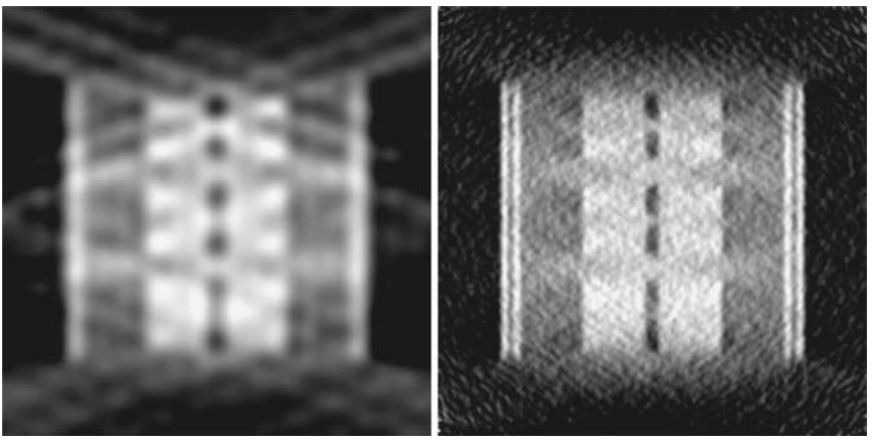


FIGURE 2. Reconstructions from the data shown in Fig. 1; by FBP on the left and by ART with blobs on the right.

Now we give an example, in the same context, of a series expansion method. Such a method is basically different from a transform method, in as much as no attempt is made to find a mathematical expression for the solution of (an idealized version) of the original problem, but rather the problem is converted into a finite-dimensional one by restricting our search for an approximation $\hat{f}$ to only those that can be expressed as a linear combination of some fixed basis functions, i.e. it is assumed *a priori* that

$$\hat{f}(r,\phi) = \sum_{j=1}^{J} x_j b_j(r,\phi) \tag{5}$$

(Herman, 1980, p. 98), and the task becomes that of estimating an appropriate J-dimensional vector $\mathbf{x}$ (whose jth component is x_j). If the data collection method is linear (as is indeed the case with line integrals), then we can say that the ith measurement ($1 \le i \le I$) is

$$y_i \simeq \sum_{j=1}^{J} r_{i,j} x_j \tag{6}$$

(Herman, 1980, p. 100), where $r_{i,j}$ is what the ith measurement would be if the structure consisted of only the jth basis function. Our understanding of the data collection procedure usually allows us to calculate (or, at least, to estimate) the $r_{i,j}$. (For the data collection scheme described in Fig. 1, $I = 70 \times 495 = 34{,}650$.) Our task is now as follows: given a data vector $\mathbf{y}$ (whose ith component is y_i) and knowing the system matrix R (whose (i,j)th entry is $r_{i,j}$), find an $\mathbf{x}$ that 'satisfies' equation (6).

The algebraic reconstruction techniques (ART) are a family of algorithms designed to solve such problems. We now describe and illustrate a specific ART algorithm, called the relaxation method for systems of equalities by Herman (1980), p. 187. (Variants of this algorithm will be discussed and illustrated later in Section 4 for the reconstruction of 3D structures from 2D projection images.)

The algorithm produces a sequence of J-dimensional vectors $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}, \ldots$. Typically we choose $\mathbf{x}^{(0)}$ to be the vector of all zeros, and we stop the process after we have cycled through all the data some integer number of times. In the step going from $\mathbf{x}^{(k)}$ to $\mathbf{x}^{(k+1)}$, we pick the next equality from equation (6) to be considered; we denote the index i associated with that equality by i_k. (The order in which the equalities are picked can be of great practical importance; see Herman and Meyer (1993). In the illustrations below, we follow the data access ordering recommended in that paper; its implication for the reconstruction of 3D structures from 2D projection images is discussed in Section 4.) Then

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda \frac{y_{i_k} - \mathbf{r}_{i_k} \cdot \mathbf{x}^{(k)}}{\mathbf{r}_{i_k} \cdot \mathbf{r}_{i_k}} \mathbf{r}_{i_k}, \tag{7}$$

where $\mathbf{r}_i$ (for $1 \le i \le I$) denotes the J-dimensional vector whose jth component is $r_{i,j}$ and $\cdot$ denotes the usual inner product. A property of this algorithm (follows from Section 16.8 of Herman (1980)) is that if there is an $\mathbf{x}$ satisfying equation (6) exactly, then the algorithm will converge to such an $\mathbf{x}$, provided that $0 < \lambda < 2$. However, the choice of λ is important for efficient early behavior, especially in the case of noisy data.

Note that the algorithm, as described above, does not depend on the choice of the basis functions. However, this choice is important in practical applications. For the illustration in Fig. 2, we have chosen the so-called blob basis functions of Lewitt (1990, 1992), following the recommendations of Matej and Lewitt (1995, 1996). A comment to be made is that using blobs as basis functions is efficacious in noisy situations (Marabini *et al.*, 1998; Matej *et al.*, 1994). The blobs that we used for the reconstruction of 3D structures from 2D projection images are defined in Section 4.2.2 below.

We applied this algebraic reconstruction technique to the data described in Fig. 1, cycling through the data 10 times. Using the same criterion as in the case of the FBP for optimizing A, we found that the optimal value of λ is 0.05. On the right of Fig. 2, we show the resulting reconstruction.

The figures of merit (FOMs), measuring the quality of the reconstructions (defined in this case as one minus a normalized squared distance between the reconstruction and the original image in a central region), are 0.18 and 0.23 for the FBP and the ART reconstructions, respectively. While this numerical difference is small, the appearances of the reconstructions

are strikingly different, indicating that this particular FOM is not good at capturing the usefulness of a reconstruction for a specific task.

3. DATA COLLECTION, MISSING INFORMATION AND IMPERFECT DATA

3.1. General Considerations

To perform a 3D reconstruction of a macromolecular structure from TEM images, it is necessary to collect a set of views of the structure from different directions. The way in which these different views are obtained varies with the type of aggregate into which the specimens are arranged.

The simplest case is presented by specimens that are aggregated into a helical superstructure (for the purpose of this presentation, it does not matter whether these aggregates are natural or artificially induced). In this case, it is clear that a single image of the helix already contains views from different directions of the individual specimens; this is so because the relative orientation between such a specimen and the electron beam changes along the helix in a well-defined way. This property makes it possible to perform a complete 3D reconstruction of the helix from one image of the helical aggregate (at least up to some aliasing-limited resolution, see DeRosier and Moore, 1970).

There are other types of aggregations that are also highly symmetrical; this is the situation, for instance, for icosahedral viruses, where a general view of the specimen already provides 59 other symmetry-related views around the specimen (Crowther *et al*., 1970). For the general case, however, we cannot count on any type of symmetries, since we may well have only non-aggregated specimens of an asymmetric structure. In this latter situation, the set of experimentally obtained projection images determining the reconstruction must be collected explicitly for a number of projection directions around the structure. The same situation occurs for specimens arranged in a 2D crystal, since most of the symmetry elements are not orientation-dependent in the tilt direction perpendicular to the grid plane (except for a possible screw axis).

In the rest of this chapter, we focus on the case of reconstructing a structures with no translational symmetries (such as 2D crystals) or point symmetries (such as icosahedral particles). A differentiation has to be made between the 'single-particle reconstruction problem' and the 'tomographic reconstruction problem'. The key conceptual difference is that in the latter case, we aim at reconstructing 'unique' specimens, while in the former case we assume that we have multiple copies of structurally identical specimens, and the task is the reconstruction of the common structure.

The differentiation introduced in the previous paragraph has very strong implications for the approaches that are used to collect data for the 3D reconstruction, as well as for the degree of completeness of the data so

obtained. For the tomographic case, we obtain different views of the specimen from different projection directions by tilting the specimen in the microscope with the help of a goniometer. The projection directions are known, at least in an approximate way, by the angles that are set at the goniometer. For the single-particle case, however, where the different views from identical copies of the macromolecular structure are encountered, the differences between the images are due to the differences in projection directions relative to the common structure. The experimental views from different copies are combined in the reconstruction as if they were coming from the same specimen at different projection directions. Clearly, if a sufficiently large number of views is obtained, then there is no need to tilt the specimens in the microscope. On the other hand, the directions of the projections are not known *a priori* and have to be determined before (or possibly as part of) the reconstruction process.

There are technical limits to the maximum tilt angle that commercial goniometers can achieve, usually around ±60°. Even if specially designed goniometers are constructed, providing, in principle, unlimited tilting (Chalcroft and Davey, 1984), there are physical limits to the maximum tilt angle that can be reached while still obtaining useful images. These limits arise from the increase in the effective specimen thickness that is proportional to 1/(*cos*(*tilt angle*), which makes multiple scattering events more probable at high tilt. For the single-particle case, these limitations do not apply in principle, since the different views can be obtained by combining images from different particles, but the problem then becomes that of 'discovering' the projection direction of each view.

The general problem of reconstructing an asymmetric object is thus posed as the 3D reconstruction of a structure from a finite set of noisy 2D projection images over a possibly restricted angular range and with an approximate knowledge of the data collection geometry. It is due to the combined presence of noise in the image pixels' values, limited angular coverage and lack of information of the projection directions that it becomes necessary to consider reconstruction algorithms capable of dealing with imperfect data.

3.2. Fourier Space Formulation

Much insight into how imperfect data may affect a reconstruction is gained by introducing the Fourier transform (FT) operator. The FT F of a function f is defined by

$$F(\mathbf{R}) = \int_{\mathbb{R}^n} f(\mathbf{r}) exp(-2\pi i \mathbf{r}\cdot\mathbf{R}) d\mathbf{r}. \tag{8}$$

In Cartesian coordinates for the case $n = 3$, this definition becomes (where now f represents a 3D structure)

$$F(X,Y,Z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y,z) exp(-2\pi i(xX + yY + zZ)) dxdydz. \tag{9}$$

The inverse FT f of F is defined by

$$f(\mathbf{r}) = \int_{\mathbb{R}^n} F(\mathbf{R}) exp(2\pi i \mathbf{r}\cdot\mathbf{R}) d\mathbf{R}. \quad (10)$$

We thus obtain pairs of functions (f, F), where F is the FT of f, and f is the inverse FT of F. The convention of denoting such pairs by lower case and upper case letters (lower case associated with real space and upper case with Fourier space) will be used throughout this chapter, in relation to both functions and variables. After applying the FT operator to the original function f, the coordinate system associated with the variable $\mathbf{R}$ of the new function F is different from the coordinate system of the variable r of the original function f. In the following, we will use the term *Fourier space* to refer to the coordinate system of F, and the term *real space* to refer to the original coordinate system of f.

One of the most important results relating f and F to each other is the so-called central section theorem, which states that the FT of an ideal projection image (containing noiseless line integrals) of a 3D structure is equal to a central section of the 3D FT of the structure (e.g. Crowther *et al.*, 1970; others refer to the same result as the projection theorem, e.g. Herman, 1980). Figure 3 shows this relationship. Let us consider the z-axis pointing in the direction of the electron beam. Without loss of generality, let us place the tilt axis along the x-axis. If the ideal projection image g is obtained by tilting the object by θ, then its FT G will be a central section tilted around the X-axis by θ of the 3D FT of the structure. It is common practice to measure tilt angles between +90° and –90° with reference to the z- (or Z-) axis; so, if θ is, say, +60° (with the usual convention that clockwise rotations are negative), then G is a central section that includes the X-axis and makes an angle of $\theta' = -30°$ with the Z-axis.

The central section theorem is fundamental to the following discussion, helping us to visualize how the different practical data collection procedures necessarily result in the partial absence of information regarding the macromolecular structure.

3.3. Single-axis Tilt, Conical Tilt and Generalized Geometries

Single-axis tilt is a common data collection procedure in electron tomography. The specimen is tilted by small increments in the microscope around a fixed axis, and a micrograph of it is obtained in each orientation (Fig. 4a). By calculating the FTs of each of these projections and using the central section theorem, we obtain a set of tilted planes in Fourier space that have the tilt axis in common (Fig. 4b and c). Note that for tilting around a single axis, the 3D problem can be divided into a set of 2D problems, each being the reconstruction in a plane perpendicular to the tilt axis, which can be done by the methods discussed in Section 2. The maximum tilt angle achievable in a microscope is limited, and this limit defines the maximum

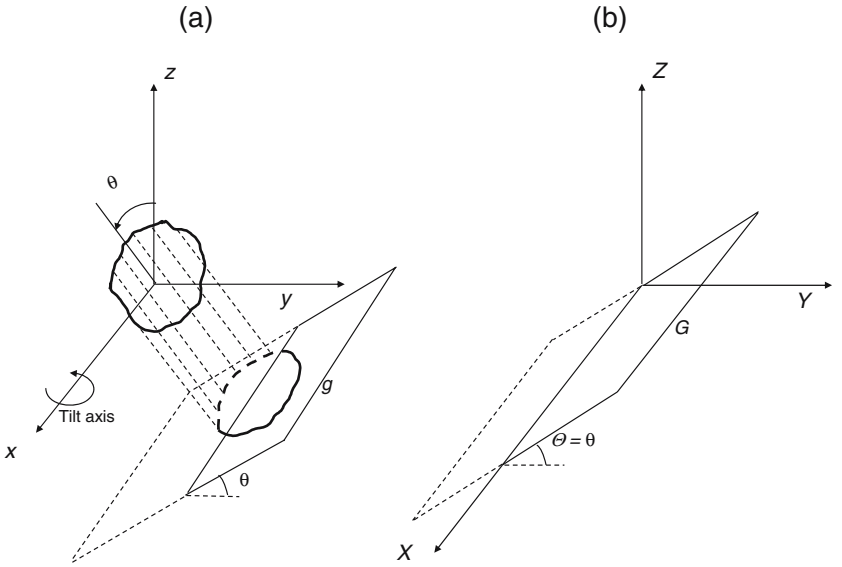


FIGURE 3. The central section theorem. Real space coordinate axes are labeled (*x*, *y*, *z*), while Fourier space coordinate axes are labeled (*X*, *Y*, *Z*). (a) An object is shown in real space placed near the origin, together with a representation of its projection image *g*, formed by integrals along parallel lines that are all orthogonal to the *x*-axis and are tilted with respect to the *z*-axis by θ. (b) The FT of the object extends over the entire Fourier space. The FT *G* of the projection *g* coincides with the FT of the object restricted to a plane through the origin of the Fourier space (a central section) that contains the *X*-axis and is tilted by θ with respect to the *XY* plane.

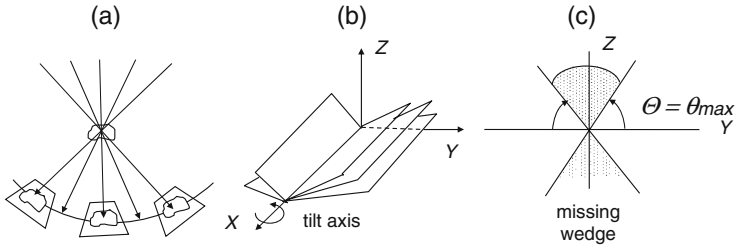


FIGURE 4. Single-axis tilt data collection geometry. (a) Tilting in real space around an axis perpendicular to the plane of this page (which is assumed to be the *x*-axis). (b) Fourier space representation of the information presented in (a); the different projections provide values in the central sections through the 3D FT of the object with the *X*-axis in common. (c) Close-up view of (b) from the the direction of the positive *X*-axis, showing the missing wedge. ((a) is from Radermacher (1980), reproduced with permission.)

tilt angle of the corresponding sections in Fourier space (Fig. 4c). It is then clear that there exists a wedge-shaped region in Fourier space for which no data can be measured (Fig. 4c), and it is consequently termed the missing wedge region. Correspondingly, the measurable area is also wedge shaped, and it is usually termed the *data wedge*.

The missing wedge is centered on the *Z*-axis in Fourier space and, assuming a maximum tilt angle of 60°, has a width of 60° (from –30° to +30° off the *Z*-axis), while the data wedge is centered around the tilt axis in the *XY* plane and has a width of 120°. Being more precise, there are two missing wedges and two data wedges (Fig. 4c).

A generalization of single-axis tilt results in the so-called conical tilt schema for data collection. Conceptually, one can get to this mode of data collection by combining single-axis tilts for tilt axes in many directions. In practice, the same geometry of data collection can be achieved by first tilting the specimen in the microscope by the maximum attainable tilt angle, θ_{max}, and then rotating it in the tilted plane by small angular increments (Fig. 5a); see Radermacher (1988) and Radermacher *et al.* (1987).

Following the same line of reasoning that was used for the single-axis tilt data collection, it is clear that there exists a portion of the FT of the structure for which no data can be measured. The shape of this missing region in Fourier space is a cone centered at the *Z*-axis with a half-width of $90° - |\theta_{max}|$ (Fig. 5b); consequently, it is referred to as the missing cone region.

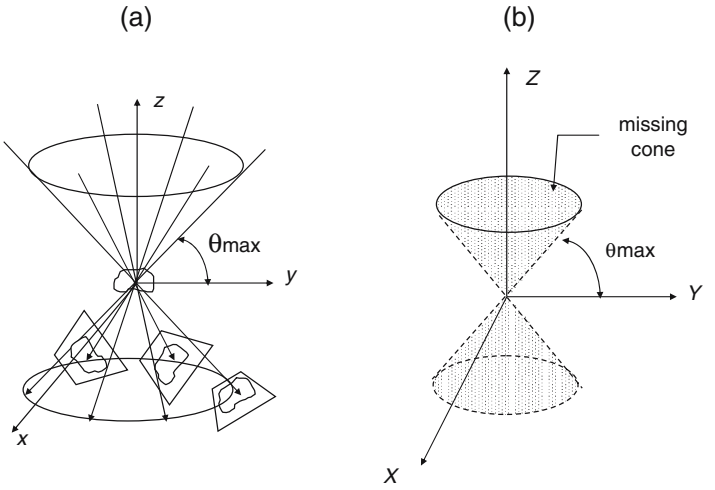


FIGURE 5. Conical tilt data collection geometry. (a) Tilting in real space by θ_{max}, around multiple tilt axes in the *xy* plane. (b) Fourier space representation of the information collected in (a); no experimental data are available in a 'missing cone' around *Z*. (From Radermacher (1980), reproduced with permission.)

Having considered single-axis tilt and conical tilt, the concept of 'generalized geometries' can now be introduced. It refers to those cases in which the pattern of projection directions is not fixed *a priori*. The most common experimental situation illustrating this schema happens in the field of single-particle reconstruction, where we have typically thousands of images that are considered to be views of the same structure from a random set of projection directions whose orientations have to be found before (or possibly during) the reconstruction process.

4. SAMPLE ALGORITHMS

4.1 Weighted Back-projection

There are inversion formulae for the reconstruction of 3D structures from 2D projection images that are very similar in nature to equation (1) for the reconstruction of 2D images from 1D projections; see, for example, Theorem 2.16 of Natterer and Wübbeling (2001). Such formulae consist of an inner integral (corresponding to the convolution in equation (1), but this time the integrations are over the 2D projection planes rather than the 1D projection lines) and a back-projection (which is this time a back-projection into 3D space of the processed 2D projection images). As in all transform methods, a computational procedure has to be designed to implement such ideal inversion formulae for the reconstruction of a 3D structure from the noisy samples of its projection images that is provided by TEM. WBP is one such computational procedure.

Before going into the details of the WBP method, we demonstrate the need for weighting (or, in terms of the approach of Section 2, convolution) with the back-projection. If we do a simple back-projection without weighting (or convolution), then all we do is just 'smear' each of the projections back along the direction from which it was taken. The result is a blurry approximation of the real structure (see Fig. 6a). The purpose of the weighting in WBP (and of the convolution in FBP) is to process the measured projection images so that when the images processed thus are back-projected, the blurring produced by a simple back-projection is avoided, or at least considerably reduced (see Fig. 6b).

In FBP, we convolve with the same filter function q in all projections (see equations (2) and (3)), irrespective of the arrangement of the projection directions. This arrangement is taken into consideration only in the back-projection, and that is why different weights are assigned to different convolved projections in equation (4). The WBP is based on an alternative philosophy: it attempts to design filters (different ones for different projections, each dependent on the geometrical arrangement of all the projections) so that when the convolved projections are simply back-projected (each with the same weight), then the the blurring seen in Fig. 6a (bottom right) will not be observed.

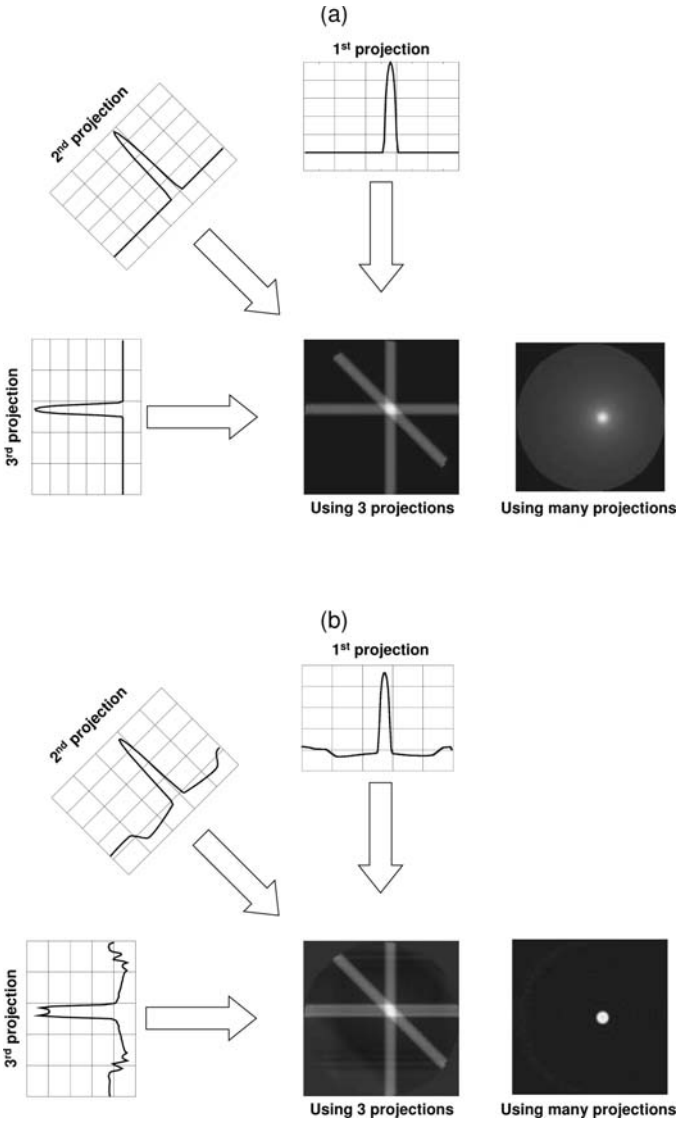


FIGURE 6. Schematic representation of simple back-projection (a) and weighted back-projection (b) for data collected using single-axis tilt geometry. In all cases, a small sphere has been reconstructed and a slice of the reconstruction perpendicular to the tilt axis and intersecting the sphere center is shown. Three plots showing one line (perpendicular to the tilt axis) of three different projections are shown with (b) and without (a) weighting based on the H_m described in the text.

A heuristic way of arriving at the WBP method is the following. Consider a point object at the origin (this is mathematically described by the Dirac delta distribution $\delta(\mathbf{r})$ whose FT has value 1 everywhere) enclosed in a sphere of radius a. Projections of such a 3D point object will be 2D point objects, and when those are back-projected into the enclosing sphere they give rise to line objects of length $2a$ going through the origin. There is one such line for each projection; see Fig. 6a. We can work out mathematically the FT H of this back-projected object. Note that H depends on a, as well as on the geometry of data collection. Restricting H to the central section in Fourier space that is parallel to the plane of the mth projection in real space (see Fig. 3) results in a function $H_m(X_m, Y_m)$ where X_m and Y_m are the coordinates within the central section in question. $H_m(X_m, Y_m)$ can also be worked out mathematically (more details on how this may be done can be found in Chapter 8 of this volume). For the point object of our study, the value of the FT in this central section is the constant 1: this implies that we should be dividing the values of the FT of the back-projection by $H_m(X_m, Y_m)$. This is impossible to do if $H_m(X_m, Y_m) = 0$ and is dangerous to do if $H_m(X_m, Y_m)$ is very small or negative, and so we introduce a (small) positive number *threshold* and replace $H_m(X_m, Y_m)$ by *threshold* whenever $H_m(X_m, Y_m) < threshold$.

Based on these ideas, we can describe the WBP algorithm as follows:

- *Weighting*: for $1 \le m \le M$ (M is the number of 2D projections), calculate p_m^c as the inverse FT of $P_m(X^m, Y^m)/H_m(X^m, Y^m)$, where P_m is the FT of the mth projection and H_m is defined as in the paragraph above.
- *Back-projection*: sum the 'smeared' versions (into the 3D reconstruction volume) of the p_m^c to form $\hat{f}$.

Note that the reconstruction depends on the choices of a and threshold.

In comparing WBP with FBP, we note the following. In the derivation of the weighting in WBP, we have taken into consideration the geometrical arrangement of the projections; in contrast, the corresponding convolving functions in FBP ignore this geometry. However, the weighting in WBP is calculated based on the assumption that we wish to reconstruct a point object at the origin and it is, strictly speaking, not valid for reconstructing point objects away from the origin. In fact, the appearance of a reconstructed point object will be dependent on its location in space. This variation can be reduced by increasing the size of the radius a. The convolving function in FBP is derived without any assumption on the location of the object to be reconstructed, hence the appearance of the reconstruction of a point object is independent of its location. (This last statement is not strictly true: there is also a lesser variation in the appearance of a reconstructed point object due to the fact that the original projections are sampled and the convolved/weighted projections are also calculated at the sample points, resulting in having to use interpolation during the back-

projection. This property is shared by WBP and FBP; its effects were nicely illustrated by Rowland (1979).)

4.2. Block ART

In this section, we discuss a particular variant of ART (introduced in Section 2) that we have used in the experiments reported below. The algorithm depends on the choice of the basis functions b_j in the 3D series expansion (which corresponds to equation (5) that was used for the 2D case). In our work, we have been using two kinds of basis functions (voxels and blobs, to be defined below), but the mathematical nature of the algorithm is the same in both cases, as will be now elaborated.

The basis functions are defined by a grid G that is a finite set $\{\mathbf{g}_1, \cdots, \mathbf{g}_J\}$ of points in 3D space and a basic basis function b that is just a function of three variables. For $1 \le j \le J$, the individual basis functions b_j are defined by

$$b_j(\mathbf{r}) = b(\mathbf{r} - \mathbf{g}_j), \tag{11}$$

i.e. the jth individual basis function is obtained from the basic basis function by shifting its center to the jth grid point. The set of grid points consist of that subset of a regular lattice (e.g. a simple cubic lattice or a body-centered cubic lattice) that lie within a sphere large enough to include the structure to be reconstructed. As in the case of WBP discussed in the previous section, we assume that there are M 2D projections. We further assume that in each 2D projection there are N measurements. Thus the total number of measurements is $I = MN$. Just as in Section 2, we use $r_{i,j}$ to denote what the ith measurement would be if the structure consisted of only the jth basis function and $\mathbf{r}_i$ (for $1 \le i \le I$) to denote the J-dimensional vector whose jth component is $r_{i,j}$. The basic difference between the variant of ART that is discussed in Section 2 and the block ART that is presented here is that in the latter the algorithm proceeds by taking into consideration not only one measurement at a time, but rather a block of measurements that come from a particular 2D projection. This results in the replacement of equations (7) by

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \lambda \sum_{i=m_k N+1}^{(m_k+1)N} \frac{y_i - \mathbf{r}_i \cdot \mathbf{x}^{(k)}}{\mathbf{r}_i \cdot \mathbf{r}_i} \mathbf{r}_i, \tag{12}$$

where m_k is the index of the 2D projection used in the kth iterative step. Just as in the 2D case, we choose $\mathbf{x}^{(0)}$ to be the vector of all zeros. The theory of such block ART algorithms was introduced by Eggermont *et al.* (1981). Their limiting convergence properties have been carefully studied in that and consequent publications (see Censor and Zenios, 1997); however, such limiting convergence results are not necessarily relevant in practical applications in which (for reasons of computational time and cost) we wish to

stop the iterative process as early as possible. One of the reasons for this early termination is that many times the solution in the limit is something like a least squares solution, which are known to fit the noise as well as the signal, and so the corresponding reconstructed volumes are too noisy (since they tend to minimize the squared error between the noisy experimental projections and the projections taken from the reconstructed volume).

The order in which the projections are picked is important. Equation blocks are determined by all those pixels belonging to the same projection image. We order the projections in such a way that each one is as nearly orthogonal as possible to the previous two projections (a projection is orthogonal to another projection if their corresponding projection directions are orthogonal).

In order to turn equation (12) into a reconstruction algorithm, we need to decide how to choose the parameter λ and at which value of k to stop the iterative process.

These choices depend on many things (see, for example, Marabini *et al*., 1997, 1998). Generally speaking, for the same number of iterations, the value of λ should be smaller if the data are noisier. Typically, in the single-particle reconstruction problem (in which the number M of 2D projections tends to be quite large), it is sufficient to cycle through the measurement data only once (i.e. to stop the iterative process at $k = M$). In the tomographic reconstruction problem (in which M tends to be much smaller), one needs to cycle through the data several times.

4.2.1. Block ART with Voxels

A basic voxel basis function depends on a variable that we will call here VEL (for voxel edge length): it is defined to have the value 1 at points strictly inside the cube that is centered at the origin and that has edges of length VEL parallel to the coordinate axes, and to have the value 0 at points strictly outside this cube. (While this does not make any difference in any possible application, for completeness we can define the basic voxel basis function to have the value 1/2 at the faces, the value 1/4 at the edges, and the value 1/8 at the corners of the cube.) The associated grid G is defined as the set of all points of the form VEL$\mathbf{R}$, where the coordinates of the vector $\mathbf{R}$ are integers, that lie within a sphere that is large enough to include the structure to be reconstructed. A consequence of this definition is that any reconstruction produced by block ART with voxels will have a constant value within cubes of edge length VEL.

4.2.2 Block ART with Blobs

A basic blob basis function is a generalization of a well-known class of functions in digital signal processing called Kaiser–Bessel (see Lewitt, 1990); it is spherically symmetric (i.e. its value at a point depends only on

the distance r of that point from the origin), has non-zero values only in a sphere (of radius a) around the origin, and smoothly decreases from a positive value at the origin to zero at the surface of that sphere. The general form of a blob is:

$$b(m,\alpha,a;r)=\begin{cases}\dfrac{I_m\left(\alpha\sqrt{1-\left(\dfrac{r}{a}\right)^2}\right)}{I_m(\alpha)}\left(\sqrt{1-\left(\dfrac{r}{a}\right)^2}\right)^m, & \text{if} \quad 0\le r\le a,\\ 0, & \text{if} \quad r>a,\end{cases} \tag{13}$$

where I_m denotes the modified Bessel function of order m and α is a parameter controlling the blob shape. The three parameters m (a non-negative integer), a and α (non-negative real numbers) control the smoothness and shape of a blob and influence the results yielded by reconstruction and visualization algorithms; therefore, the appropriate selection of them is highly important. In our work, we set m equal to 2, which makes the blobs have continuous first derivatives everywhere.

The choice of the grid G is also important. It was shown by Petersen and Middleton (1962) that the body-centered cubic (*bcc*) grids provide the most efficient sampling of $\mathbb{R}^3$. The *bcc* grids are defined by

$$B_\Delta=\left\{\Delta\mathbf{k}\,\middle|\,\mathbf{k}\in\mathbb{Z}^3 \text{ and } k_1\equiv k_2\equiv k_3 (mod\ 2)\right\}, \tag{14}$$

where $\mathbf{k}$ is a 3D vector, whose components (k_1, k_2 and k_3) belong to the set of integers denoted by $\mathbb{Z}$, and Δ is a positive real number (the sampling distance). In order to visualize this grid, we can use a small portion of it and take advantage of its periodic repetition (see Fig. 7). For reconstruction purposes, Matej and Lewitt (1995) demonstrated that whenever a linear combination of blobs is employed to obtain a reconstruction, the *bcc* grids provide desirable sets of locations for the center of the blobs.

Having decided that we use $m = 2$ and the *bcc* grid, there are three parameters to be chosen: Δ, a and α. Clearly, to be able to approximate

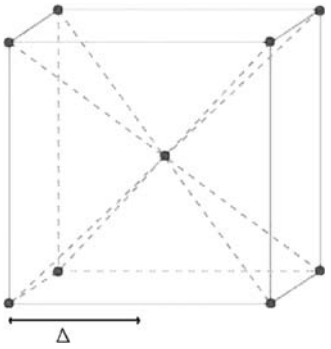


FIGURE 7. Points in the body-centered cubic grid in a $2 \times 2 \times 2$ portion of space (assuming $\Delta = 1$). The rest of the points can be obtained by filling in space by the most natural repetition of the indicated $2 \times 2 \times 2$ portion.

arbitrary structures, the value of Δ should be small. However, in a fixed volume of space, the number of grid points (and consequently the computational cost of a reconstruction algorithm) is proportional to $1/\Delta^3$ and so practical considerations do not allow us to choose Δ to be very small. The cost of reconstruction is also proportional to a^2. The computational cost does not depend on α, and so this parameter may be chosen purely based on the quality of the resulting reconstructions. Matej and Lewitt (1995) proposed a method for the selection of the parameters with the aim of ensuring that ART will produce 'good' reconstructions. In particular, they postulate that a linear combination of blobs with $x_j = 1$, for $1 \leq j \leq J$, should approximate a constant-valued function. They show that for this we should select Δ, a and α such that

$$\alpha = \sqrt{2\pi^2\left(\frac{a}{\Delta}\right)^2 - 6.987932^2} \tag{15}$$

is satisfied.

Garduño and Herman (2004) reported that the parameters yielded by the methodology suggested by Matej and Lewitt (1995) produced in some cases non-convex reconstructions from data obtained from convex structures, causing a significant inaccuracy in the visualization of the resulting surfaces. To correct this problem, Garduño and Herman (2004) proposed an additional criterion for the selection of the parameters Δ, a and α: a should be chosen as small as possible, consistent with both satisfying equation (15) and achieving the result that if two blobs at nearest grid points in the grid B_Δ (those separated by $\sqrt{3}\Delta$) are given coefficients 1 with all other blobs given coefficients 0, then the surface obtained by thresholding at $t = 0.5$ should enclose a convex set. The selection of the latter criterion is not arbitrary but is based on the fact that the location of the two nearest blobs will determine the best resolution that can be achieved by an approximation using blobs. Furthermore, given a choice of Δ, such a criterion provides us with a unique pair a and α that satisfies equation (15). This is the methodology used in this chapter for determining the blob parameters.

5. PRACTICAL ILLUSTRATION

5.1. General Considerations

In this section, we present practical cases of how different reconstruction methods treat the 'imperfect data' problem that is typical in 3D electron tomography. As a general remark, it is important to stress that 'the best reconstruction method' does not exist. Each technique makes its own assumptions about the possible set of solutions, the noise or the type and extent of *a priori* knowledge. Depending on the specific problem at hand, these assumptions may or may not be the most appropriate ones.

There are two important points directly related to 3D electron microscopy applications that need to be addressed here. One is the characterization of the image degradations introduced by the electron microscope as an image-forming system, the other is the noise model to be used in the reconstruction process.

Not surprisingly, the electron microscope cannot be considered a perfect imaging system. It is well known that it introduces some degradations into the images it renders. Erickson and Klug (1971) presented an analysis on how these instrumental degradations can be modeled, within certain limitations, in terms of a single-system optical transfer function (OTF), providing a parametric model for it (in the field of electron microscopy, the OTF is usually called the contrast transfer function or CTF). The corresponding point-spread function is referred to as the 2D TEM-PSF. In general, the 2D TEM-PSF, being the inverse FT of the instrument CTF, does not have a finite support, although in practice it can be approximated by a finite-support blurring function.

With respect to the contents of this chapter, knowledge of the CTF is important because it tells us that the relationship between the 3D object and the 2D image obtained from it is not that of a simple projection, but rather a projection convolved by the 2D TEM-PSF (defined in the previous paragraph). This fact should be kept in mind when modeling the image formation.

In a first approximation to the reconstruction problem, one might think of neglecting this instrumental degradation. However, this is not an acceptable choice in the quest for high-resolution structural information. Additionally, the fact that the geometry of data collection is known only in an approximate way introduces imperfections in our initial data set as well.

Regarding the noise model, it is clear that part of the noise is due to electron counting, and thus follows a Poisson process. It is also clear that there are other contributions coming from the specimen embedding and supporting material, from the photographic emulsion and processing, the photographic grain, and inaccuracies in the scanning devices. Other sources of 'noise' are more difficult to consider, such as the inaccuracies in the determination of the geometry of data collection or inaccuracies in the determination and restoration for the CTF. It follows from all these considerations that realistic noise models are not trivial, and therefore should always be used with some caution, being aware of their limitations.

5.2. Test Data Set Description and Testing Approach

In this section, we present a practical study of two widely used reconstruction methods: WBP and block ART. Only block ART with blobs has been studied, based on its well-established superiority to block ART with voxels (Matej and Lewitt, 1996; Matej *et al.*, 1994). The two methods have been tested on three simulated data sets, two typical of high-resolution

studies of single particles (one typical of the conical tilt geometry as depicted in Fig. 5, and the other typical of a geometry with even angular distribution) and the third typical of electron tomography. Each data set consists of a number of independently generated collections of projection images (each collection in the data set corresponds to one instance of what is denoted by g in Section 1). The performance of the methods is measured and compared using a well-defined FOM in a statistical manner. This work has been done using the X-Window-based Microscopy Image Processing Package (XMIPP) that is freely available from the authors' web site (http://xmipp.cnb.uam.es/).

The first test data set consisted of 30 collections of 10,000 projection images which were obtained using random projection directions of the thermosome from *Thermoplasma acidophilum* (Ditzel *et al*., 1998). The sampling interval, in both the images and the volume, was 0.266 nm. The dimensions of the images and of the volumes were 96×96 and $96 \times 96 \times 96$, respectively. The simulation was intended to represent a practical case of high-resolution (below 1 nm) single-particle reconstruction.

The second test data set consisted of 10 collections of 10,000 projection images each that were obtained randomly within a conical tilt data collection geometry with a tilt angle of 60° of the thermosome from *T. acidophilum* (Ditzel *et al*., 1998).

The third test data set consisted of 10 collections of 70 regularly distributed projection images following the single-axis tilt data collection geometry with a maximum tilt angle of ±60° of the mitochondria-like phantom described by Fernández *et al*. (2002) and Bilbao-Castro *et al*. (2004). Each of the images was of dimensions 256×256 pixels, and the reconstructed volume was $256 \times 256 \times 256$ voxels. The rationale here was to simulate the case of a tomographic study of a cell organelle.

Note that these data sets represent very different mathematical situations when they are translated into a system of conditions defining the reconstruction in equation (6). In the first two cases, we have a very large collection of images and the task is the reconstruction of a relatively small volume, while it is exactly the opposite that happens in the third case, since here we have relatively few large images and the task is to reconstruct a large volume. To be precise, in the first two cases, the total number of pixels in a collection g of projection images is $10{,}000 \times 96 \times 96$, while the number of voxels in the volume is $96 \times 96 \times 96$; in the third case, the total number of pixels in g is $70 \times 256 \times 256$ and the number of voxels is $256 \times 256 \times 256$. As a consequence, the system (equation 6) is overdetermined in the first two cases by a factor of ~100, and is underdetermined in the third case. Additionally, as the size of the images become larger and larger in tomographic applications with a reduced number of views, the tomographic reconstruction problem will become more and more underdetermined, while in the single particle cases, as the number of images coming from different specimens become larger, the reconstruction problem becomes more and more overdetermined.

Another difference is the coverage of the Fourier space that can be achieved using g. This is reasonably uniform in the first case, has a missing cone in the second case (see Fig. 5) and has a missing wedge in the third case (see Fig. 4). WBP and ART each have their ways of handling such data: WBP takes the geometry of projection directions into consideration in designing the weighting function, while the system (equation 6) that ART attempts to solve is explicitly based on the projection directions.

Regarding the generation of the projection images of the thermosome, each of them was calculated from the set of atomic coordinates contained in Protein Data Base (PDB) entry 1A6D by replacing each non-hydrogen atom in the structure with a blob (equation 13) with appropriately chosen parameters, which allows a perfect analytical description of the projections. Projections were then sampled at intervals of 0.266 nm. The generation of the mitochondrion-like phantom was performed as in Fernández *et al.* (2002). This phantom consists of hollow cylinders representing the membranes and a set of solid cylinders simulating the cristae. The cristae are embedded in a region of intermediate density, resembling the mitochondrial inner matter.

The electron microscope-induced projection blurring was taken into account using a family of CTFs whose average is plotted in Fig. 8. The defocus value was randomly changed between views within the interval [–0.315, –0.285] μm for the single particle case and the interval [–1.575, –1.475] μm for the tomographic reconstruction. In the experiment with random conical tilt, the tilt angle was fixed to 60°. Restoration of the CTF was performed such that it tended to resemble a practical experimental case: only a phase flipping was performed based on a CTF estimate provided by all the projection images.

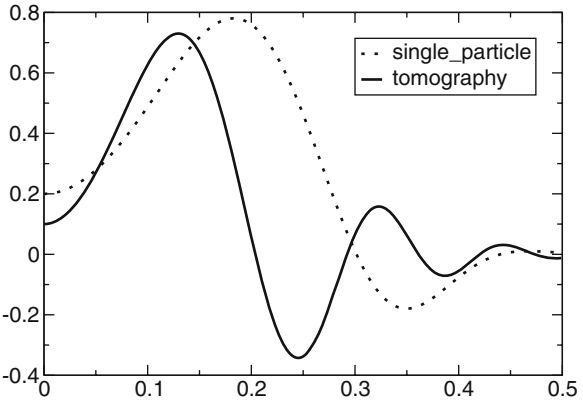


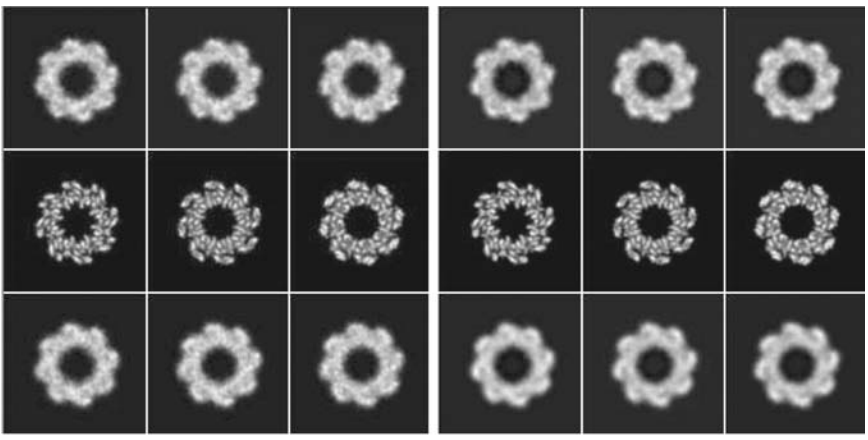
FIGURE 8. Plot of the average CTFs used in the single particle (dotted line) and the tomographic (solid line) experiments, respectively. The horizontal axis represents normalized frequencies so that the maximum frequency given by the Nyquist theorem is mapped to 0.5.

As for inaccuracies related to the determination of the actual geometry of projection directions, Gaussian noise was added to all the parameters determining it, so that views were not perfectly centered, but they had a noise with 2 pixel edge lengths standard deviation in both x and y, and 2° in each of the three Euler angles. Gaussian noise was added to each pixel value so that the signal-to-noise ratio was 1/3 in the case of single particles and 2/3 in the case of tomography.

The testing of the different algorithms should be based on a statistical comparison of their relative ability to perform well-defined and relatively simple tasks coded in the form of FOMs. However, since all algorithms have free parameters that have to be adjusted, prior to their comparisons it is necessary to perform a parameter optimization approach in which, for all reconstruction algorithms, their free parameters are adjusted to optimal performance. The parameter optimization and the algorithm evaluation should be performed on the sames FOMs. This methodology has been presented in a number of previous works, such as Furuie *et al*. (1994), Matej *et al*. (1994) and Sorzano *et al*. (2001). For this chapter, we have selected, as an illustrative example, a FOM based on the Euclidean distance between the reconstructed volume and the original volume: $scL2 = 1 - \frac{1}{V}\sum_{v=1}^{V}\left(\frac{p_v - r_v}{2}\right)^2$ (where p_v and r_v are the values assigned to the vth voxel of the phantom and the reconstruction, respectively, and V is the total number of voxels within a sphere of radius half the reconstructed volume edge). Clearly, this FOM measures only a global characteristic of reconstruction accuracy.

The parameters to be optimized for WBP, as described in Section 4.1, are a and *threshold*. In fact, based on past experience, we have in all cases fixed a so that it is half the length of the edge of the cubic volume within which we perform the reconstruction. For block ART, we need to select the relaxation parameter, λ, which controls the magnitude of the update of the reconstructed volume during each iteration; see equation (12). Note that the number of iterations for block ART is normally also a free parameter. However, we set it to M (the number of projection images) for the overdetermined case following the results and discussions presented in Marabini *et al*. (1998), and to $10M$ for the underdetermined case (see Marabini *et al*., 1997). With these choices for the number of iterations, the selected values of λ were 0.3124, 0.0625 and 0.5, respectively, for the three test data sets.

In Fig. 9a, we present the results corresponding to the first data set. In this overdetermined case, the performances of the two reconstruction algorithms according to this FOM called *scL2* were similar, rendering FOM values of 0.62 ± 0.03 for block ART and 0.60 ± 0.03 for WBP. The first value is the mean of the FOM over the 30 collections of 10,000 projection images each, and the second value is the standard deviation. Bearing in mind that the standard deviation of the mean of 30 samples is $(1/\sqrt{30})$ times the standard deviation of the original distribution, we see that the difference



(a) Radomly distributed images (b) Conically distributed images

FIGURE 9. Three slices perpendicular to the *Z*-axis corresponding to, from top to bottom, the block ART reconstruction, the phantom volume and the WBP reconstruction, respectively.

between the means is over three times the standard deviations of the means, implying that the observed difference has statistical significance, albeit not a large one. (The phrase 'statistical significance' is used here in the sense that is standard in hypothesis testing. Its meaning is not that the difference between the means of the outcomes of the experiments using the two methods is large, or even that it is necessarily important from the point of view of the application. Rather it is a measure of our confidence by which we can reject the null hypothesis that the methods are equally good, from the point of view of the FOM, in favour of the alternative hypothesis that the one for which the mean is larger is the better one. This confidence depends not only on the difference between the means, but also on the variability of the experiments, as indicated by the standard deviations. Thus statistical significance is indicated even by a very small difference in means, as long as the standard deviations are correspondingly small.)

Regarding the case in which there was a missing cone of data, the results are presented in Fig. 9b. This time, the global FOMs (0.53 ± 0.01 for block ART and 0.49 ± 0.02 for WBP) are less similar: bearing in mind again that the means are averages of several (this time 10) samples, the difference between the mean FOMs for the two reconstruction methods is even more significant than in the previous case, with block ART presenting more details than back-projection. This type of result was expected, since we found in our previous work (Marabini *et al.*, 1997) that WBP is more sensitive to a missing angular range than ART with blobs.

As for the third case of a tomographic reconstruction problem, it is an underdetermined situation with a missing angular range. The results

FIGURE 10. One slice perpendicular to the *Z*-axis corresponding, from left to right, to the ART reconstruction, the phantom volume and the WBP reconstruction, respectively.

obtained are presented in Fig. 10. The FOM specified above provides values of 0.9958 ± 0.0001 for the block ART reconstruction and of 0.9954 ± 0.0002 for WBP that are again significantly different. Two observable disadvantages of WBP when compared with block ART are: a blurring along the z-axis (notice the two extra cylinders that appear on the left in the WBP reconstruction that have been 'smeared' from another plane) and the loss of the resolution in the 'bilayer' that surrounds the specimen as it goes farther away from the center (a behavior that is in agreement with the theory, since as we move away from the center we increasingly violate the assumed nature of the point-spread function of simple back-projection, and so the correction becomes less reliable).

6. DISCUSSION

In this work, we have presented two different families of reconstruction algorithms (transform methods and series expansion methods) used in the field of 3D electron microscopy, including a more detailed analysis of two common algorithms: WBP and block ART. There are several lessons to be learned from this exercise.

The first one is the realization that no algorithm can be treated as a black box, and that special attention has to be given to the actual value of the parameters that characterize each algorithm in the context of a particular application. In this study, we have presented an approach in which the value of these parameters is selected according to the maximization of a well-defined FOM. Clearly, there is ample opportunity to define other FOMs and, consequently, to 'tune' the algorithms for a better recovery of certain types of information, such as 'feature separability', 'contrast' etc., which is a common practice in the field of computed tomography (CT) in medicine. This review does not pretend to explore this issue in depth, as it

deserves a work of its own, but simply presents this situation explicitly since many times it is hidden in the common use of the methods.

The second lesson is the importance of limits, other than the ones imposed by the reconstruction algorithms, on our capacity to achieve better reconstructions and, in this way, higher resolution biological information from our 3D reconstructions. Of utmost importance are the errors due to our imperfect knowledge of the geometry of data collection and, also, due to the inaccuracies introduced by the instrument's CTF. This consideration indicates that a very interesting way to overcome these limitations could be a combination, into a single algorithm, of the functionalities of reconstruction techniques and geometry estimation methods (see Scheres *et al*. (2005) for a way to advance in this direction).

Finally, the third reflection is that under the general title of '3D EM Reconstruction Methods' there are quite different types of situations. A clear example is the realization that typically single-particle reconstruction problems are greatly overdetermined in terms of the basic characteristics of the systems of equations that arise from them, while tomographic reconstruction problems fall at the opposite extreme. Furthermore, the two types of problems are evolving in such a way that the former one becomes even more overdetermined, while the latter one becomes more and more underdetermined. Being so different in their basic mathematical characteristics, it is clear that the relative performance of the methods in these two areas could also be very different: what performs 'best' in the one case may be inferior in the other.

ACKNOWLEDGEMENTS

The authors are thankful to Dr J. J. Fernandez for kindly providing the mitochondria phantom. We acknowledge partial support from the 'Comunidad Autónoma de Madrid' through grants CAM-07B-0032-2002 and GR/SAL/0342/2004, the 'Comisión Interministerial de Ciencia y Tecnología' through grants BIO2001-1237, BIO2001-4253-E, BIO2001-4339-E, BIO2002-10855- E, BFU2004-00217/BMC and TIC990361, the European Union through grants QLK2-2000-00634, QLRI-2000-31237, QLRT-2000-0136 and QLRI-2001-00015, the NIH through grant HL70472, and the 'Fundación de Investigación Sanitaria' through network FIS G03/185 and grant PI040683.

REFERENCES

Bilbao-Castro, J. R., Sorzano, C. O. S., García, I. and Fernández, J. J. (2004). Phan3D: design of biological phantoms in 3D electron microscopy. *Bioinformatics* **20**:3286–3288.

Censor, Y. and Zenios, S. A. (1997). *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, New York.

Chalcroft, J. P. and Davey, C. L. (1984). A simply constructed extreme-tilt holder for the Philips eucentric stage. *J. Microsc*. **134**:41–48.

Crowther, R. A., DeRosier, D. J. and Klug, F. R. S. (1970). The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. *Proc. R. Soc. A* **317**:310–340.

DeRosier, D. J. and Moore, P. B. (1970). Reconstruction of three-dimensional images from electron micrographs of structures with helical symmetry. *J. Mol. Biol.* **52**:355–369.

Ditzel, L., Lowe, J., Stock, D., Stetter, K. O., Huber, H., Huber, R. and Steinbacher, S. (1998). Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT. *Cell* **93**:125–138.

Eggermont, P. P. B., Herman, G. T. and Lent, A. (1981). Iterative algorithms for large partitioned linear systems with applications to image reconstruction. *Linear Algebra Appl*. **40**:37–67.

Erickson, H. P. and Klug, A. (1971). Measurement and compensation of defocusing and aberrations by Fourier processing of electron micrographs. *Philos. Trans. R. Soc. B* **261**:105–118.

Fernández, J., Lawrence, A., Roca, J., García, I., Ellisman, M. and Carazo, J. (2002). High performance electron tomography of complex biological specimens. *J. Struct. Biol*. **138**:6–20.

Furuie, S. S., Herman, G. T., Narayan, T. K., Kinahan, P. E., Karp, J. S., Lewitt, R. M. and Matej, S. (1994). A methodology for testing for statistically significant differences between fully 3D PET reconstruction algorithms. *Phys. Med. Biol.* **39**:341–354.

Garduño, E. and Herman, G. T. (2004). Optimization of basis functions for both reconstruction and visualization. *Discrete Appl. Math.* **139**:95–112.

Herman, G. T. (1980). *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*. Academic Press, New York.

Herman, G. T. and Meyer, L. B. (1993). Algebraic reconstruction techniques can be made computationally efficient. *IEEE Trans. Medical Imaging* **12**:600–609.

Lewitt, R. M. (1990). Multidimensional digital image representations using generalized Kaiser–Bessel window functions. *J. Optic. Soc. Am. A* **7**:1834–1846.

Lewitt, R. M. (1992). Alternatives to voxels for image representation in iterative reconstruction algorithms. *Phys. Med. Biol*. **37**:705–716.

Marabini, R., Herman, G. T. and Carazo, J. M. (1998). 3D reconstruction in electron microscopy using ART with smooth spherically symmetric volume elements (blobs). *Ultramicroscopy* **72**:53–65.

Marabini, R., Rietzel, E., Schröder, R., Herman, G. T. and Carazo, J. M. (1997). Three dimensional reconstruction from reduced sets of very noisy images acquired following a single-axis tilt schema: application of a new three-dimensional reconstruction algorithm and objective comparison with weighted backprojection. *J. Struct. Biol.* **120**:363–371.

Matej, S., Herman, G. T., Narayan, T. K., Furuie, S. S., Lewitt, R. M. and Kinahan, P. E. (1994). Evaluation of task-oriented performance of several fully 3D PET reconstruction algorithms. *Phys. Med. Biol*. **39**:355–367.

Matej, S. and Lewitt, R. M. (1995). Efficient 3D grids for image reconstruction using spherically-symmetric volume elements. *IEEE Trans. Nucl. Sci*. **42**:1361–1370.

Matej, S. and Lewitt, R. M. (1996). Practical considerations for 3-D image reconstruction using spherically symmetric volume elements. *IEEE Trans. Med. Imaging* **15**:68–78.

Natterer, F. and Wübbeling, F. (2001). *Mathematical Methods in Image Reconstruction*. SIAM, Philadelphia.

Petersen, D. P. and Middleton, D. (1962). Sampling and reconstruction of wavenumber limited functions in *n*-dimensional Euclidean spaces. *Inform. Control* **5**:279–323.

Radermacher, M. (1980). Dreidimensionale Rekonstruktion bei kegelförmiger Kippung im Electronenmikroskop. PhD Thesis, Technische Universität München.

Radermacher, M. (1988). Three-dimensional reconstruction of single particles from random and nonrandom tilt series. *J. Electron Microsc. Tech*. **9**:359–394.

Radermacher, M., Wagenknecht, T., Verschoor, A. and Frank, J. (1987). Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal unit. *J. Microsc.* **146**:113–136.

Radon, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig. *Math. Phys. Klasse* **69**:262–277.

Rowland, S. (1979). Computer implementation of image reconstruction formulas. In *Image Reconstruction from Projections* (G.Herman, ed.). Springer-Verlag, Berlin, pp. 29–79.

Scheres, S. H. W., Valle, M., Núñez, R., Sorzano, C. O. S., Marabini, R., Herman, G. T. and Carazo, J. M. (2005). Maximum-likelihood multi-reference refinement for electron microscopy images. *J. Mol. Biol.* **348**:139–149.

Sorzano, C. O. S., Marabini, R., Boisset, N., Rietzel, E., Schröder, R., Herman, G. T. and Carazo, J. M. (2001). The effect of overabundant projection directions on 3D reconstruction algorithms. *J. Struct. Biol.* **133**:108–118.

8

Weighted Back-projection Methods

Michael Radermacher

Michael Radermacher • University of Vermont Collage of Medicine, Department of Molecular Physiology and Biophysics, HSRF 120, Burlington, VT 05405, USA

1. INTRODUCTION

Traditionally, 3D reconstruction methods have been classified into two major groups, *Fourier reconstruction methods* and *direct methods* (e.g. Crowther *et al*., 1970; Gilbert, 1972). Fourier methods are defined as algorithms that restore the Fourier transform of the object from the Fourier transforms of their projections and then obtain the real-space distribution of the object by inverse Fourier transformation. Included in this group are also equivalent reconstruction schemes that use expansions of object and projections into orthogonal function systems (e.g. Cormack, 1963, 1964; Smith *et al*., 1973; Chapter 9 of this volume). In contrast, direct methods are defined as those that carry out all calculations in real space. These include the convolution back-projection algorithms (Bracewell and Riddle, 1967; Gilbert, 1972; Ramachandran and Lakshminarayanan, 1971) and iterative algorithms (Gordon *et al*., 1970; Colsher, 1977). Weighted back-projection methods are difficult to classify in this scheme, since they are equivalent to convolution back-projection algorithms, but work on the real-space data as well as the Fourier transform data of either the object or the projections. Both convolution back-projection and weighted back-projection algorithms are based on the same theory as Fourier reconstruction methods, whereas iterative methods normally do not take into account the Fourier relationships between object transform and projection transforms. Thus it seems justified to classify the reconstruction algorithms into three groups: *Fourier reconstruction methods*, *modified back-projection methods* and *iterative direct space methods*, where the second group includes convolution back-projection as well as weighted back-projection methods.

Each reconstruction algorithm requires a set of projections of the object, recorded under different projecting directions. While the choice of the data collection geometry is determined by the properties of the specimen, e.g. by its radiation sensitivity, by the degree of variation among particles in the preparation and by orientational preferences of the particles, this geometry in turn determines which algorithm is most efficient in calculating the 3D structure.

As explained in more detail in the introductory chapter by Frank, four kinds of data collection schemes are used in electron microscopy: single-axis tilting, conical tilting, random-conical tilt and general random tilt. For the collection of a single-axis tilt series, the specimen is tilted in the microscope in a range of typically –60° to +60° in small increments, e.g. 1–5°, and an image of the same particle is recorded for each specimen position. For a conical tilt series, the specimen is tilted by one fixed angle in the range of 45–60° and then rotated within this plane by small angular increments. Again an image of the same particle is recorded for each specimen position. Both data collection schemes are mainly used for preparations that are radiation resistant or contain particles that individually have different shapes, i.e. averaging over different particles is either not required or not

possible. Conical tilt reconstructions have seen a recent revival in the application to thin sections and replicas (Bellon *et al*., 2005, Zampighi *et al*., 2005) Reconstruction schemes based on a random or random-conical tilt geometry intrinsically average over a large number of identical particles, and the electron exposure of each single particle can be kept very low. The random-conical data collection scheme (Frank and Goldfarb, 1980; Radermacher *et al*., 1986, 1987) makes use of the fact that many particles tend to exhibit a preferred orientation with respect to the specimen support plane, leaving only one rotational degree of freedom, which is a rotation around an axis perpendicular to this plane. The set of images from many such particles in a micrograph of a tilted specimen form a conical tilt series with random azimuthal angles. To achieve the maximum amount of 3D information, a large tilt angle is required, typically between 45 and 65°. If a preferred orientation is not present, then any micrograph of the specimen showing particles in random orientation provides a random tilt series (e.g. Vogel and Provencher, 1988). Intermediate to the latter situation is a specimen of particles with more than one preferred orientation. Each such orientation can be used as a random-conical tilt series, or the sets of particles with different preferred orientations can be combined into a single random tilt series (Carazo *et al*., 1989; Frank *et al.*, 1988).

Among all types of reconstruction algorithms, most work has been devoted to the reconstruction of objects from a single-axis tilt series with equal angular increments. The Fourier algorithms have been extensively optimized, and most of them perform very efficiently. Many variations of the convolution and weighted back-projection algorithms can be found that are designated for single-axis tilt geometry (e.g. Gilbert, 1972; Kwok *et al*., 1977; Ramachandran and Lakshminarayanan, 1971).

For the regular conical tilt geometry, the choice of reconstruction algorithms is much smaller. Although many iterative direct space algorithms can be used that do not rely on a specific geometry, their performance can be very slow. In many Fourier reconstruction methods, the problem of Fourier interpolation and the associated matrix inversion becomes prohibitively large even for most of currently available computers. Recently developed methods have overcome many of these limitations by using a moving window Shannon interpolation technique (Bellon and Lanzavecchia, 1997, Bellon *et al*., 1999; Lanzavecchia and Bellon, 1998). Methods that use orthogonal series expansions still need more development to be efficient in applications to regular conical tilt geometries. Weighted back-projection methods, on the other hand, have been extended to the reconstruction from conical projection series (Radermacher, 1980; Radermacher and Hoppe, 1978) and offer a very efficient solution.

A similar situation can be found for the reconstruction from arbitrary non-regular geometries with randomly distributed tilt angles. As in the case of regular conical tilt geometries, the iterative methods are directly applicable. However, as the number of input images in reconstructions from

randomly distributed projections is necessarily much larger than in reconstructions from regular tilt geometries, the problem of computing time requirements increases even further. Again the weighted back-projection method in the form developed to deal with arbitrary geometries (Harauz and van Heel, 1986; Radermacher *et al.*, 1986, 1987) has become a very efficient reconstruction algorithm. The two-step Radon inversion algorithm, described later in this chapter, has at its core an *R*-weighted back-projection and is the fastest of all the current reconstruction algorithms for arbitrary geometry. In this context, we must also mention the series expansion method developed for arbitrary geometries (Provencher and Vogel, 1988; Vogel and Provencher, 1988). However, this very elegant method has not found any widespread use beyond its original publication.

For the reasons stated, weighted back-projection methods and iterative methods are currently the most widely used algorithms for the reconstruction of single, asymmetrical particles in electron microscopy. The version developed for single-axis tilt geometry is extensively used for the reconstruction of tomographic data sets. This includes tilt series of structures embedded in thick sections (e.g. Frank *et al.*, 1987; McEwen *et al.*, 1986) and, increasingly, tomographic data sets of cryospecimens (e.g. Hsieh *et al.*, 2002; Stöffler *et al.*, 2003). The weighted back-projection algorithm for arbitrary geometry is used for the reconstruction of macromolecular structures from single-exposure random-conical tilt series (Radermacher *et al.*, 1986), for which this algorithm was specifically developed, and as part of angular reconstitution techniques. In many situations, this algorithm can be replaced by the two-step Radon inversion algorithm which is based on *R*-weighted back-projection.

Besides performing faster on large data sets, weighted back-projection algorithms have the important advantage over most iterative algorithms that all operations involved are linear and the outcome of the reconstruction is entirely determined by the experimental input data. This linearity facilitates a description of weighted back-projection methods using the concept of *point-spread function*s and *transfer functions*. Constraints based on a priori knowledge can be applied using POCS-style methods in a separate step, independent of the reconstruction algorithm itself.

This chapter starts with a short introduction to the concept of linear systems. Following this, the transfer function of a simple back-projection is determined, which yields the weighting functions for the weighted back-projection. First, the most general weighting function, for a reconstruction from projection with arbitrary angular distribution, is derived. Afterward, specific analytical forms of the weighting functions are shown that can be derived for data sets that consist of projections obtained with equal angular increments.

Many other weighting functions have been proposed which differ from the functions shown here mainly by the incorporation of low-pass filters. The functions derived here do not contain implicit low-pass filters; instead,

the formulae will be given that describe the dependence of the resolution on the number of projections available. With these formulae available, the reconstructions can be low-pass-filtered independently of the reconstruction algorithm used. This approach allows for flexibility in the choice of the low-pass filter function and ensures that maximum use is made of the available data. At the end of this section, some examples of computer implementations of the reconstruction algorithms will be given.

2. THE CONCEPT OF POINT-SPREAD FUNCTIONS AND TRANSFER FUNCTIONS

The point-spread function of an imaging system describes the image of a single point as it results after using a perfect point as input to the system. Such a system can be any linear optical system or an algorithm that 'images' an input distribution onto an output distribution, or even the combination of optical and digital systems. The complete system involved in 3D electron microscopy is the microscope that images the object onto a photographic plate, image plate or CCD camera, the digitizer that transfers the image to the computer, the algorithms used to process the images, and the media and equipment to record the final result.

For a more detailed analysis of point-spread and transfer functions, we refer the reader to Goodman (1968), whose terminology will be closely followed, and to Papoulis (1968).

A system is defined as a 'black box' that maps a set of input functions onto a set of output functions. The system will be described by the operator S. If $g_1(x', y')$ represents a 2D input and $g_2(x, y)$ is the output of the system, then we use the notation

$$g_2(x, y) = S[g_1(x', y')] \tag{1}$$

Because the extension to 3D and higher dimensional systems is straightforward, only 2D systems will be regarded here. If the operator S (or the system it is associated with) is linear, then

$$S[a \cdot f(x', y') + b \cdot g(x', y')] = a \cdot S[f(x', y')] + b \cdot S[g(x', y')] \tag{2}$$

The response of a linear system can be described as the superposition of responses to elementary functions into which the input can be decomposed. One possible decomposition of the function $g_1(y', y')$ would be

$$g_1(x', y') = \iint g_1(\xi, \eta)\delta(x' - \xi, y' - \eta)d\xi d\eta \tag{3}$$

where δ is the 2D δ-function. The function g_1 passing the system results in

$$g_2(x, y) = S\left\{\iint g_1(\xi, \eta)\delta(x' - \xi, y' - \eta)d\xi d\eta\right\} \tag{4}$$

and from the linearity of S, it follows that

$$g_2(x, y) = \iint g_1(\xi, \eta) S\{\delta(x' - \xi, y' - \eta)\} d\xi d\eta \tag{5}$$

Here the input function $g_1(\xi, \eta)$ simply appears as a coefficient of the elementary function $\delta(x' - \xi, y' - \eta)$. To describe the mapping of g_1, only the system's response h to the delta-function, called the *impulse response* or (in optics) the *point-spread function*, needs to be known for every location (x, y):

$$h(x, y; \xi, \eta) = S\{\delta(x' - \xi, y' - \eta)\} \tag{6}$$

Note, however, that the system's response still can depend on the location of the point.

If the system is *shift-invariant*, a property also called *isoplanatic* in optics, then the system response is independent of the absolute coordinates and depends only on the difference vector $(x - \xi, y - \eta)$. The point-spread function then can be written as

$$h(x, y; \xi, \eta) = h(x - \xi, y - \eta) \tag{7}$$

Combining equations (5–7), we see that the mapping by a linear shift-invariant system can be described as the convolution of the input function with the point-spread function:

$$g_2(x, y) = \iint g_1(\xi, \eta) \cdot h(x - \xi, y - \eta) d\xi d\eta \tag{8}$$

In connection with linear shift-invariant systems, the convolution theorem of Fourier transforms is of importance:

$$g * f = \boldsymbol{F}^{-1}\{\boldsymbol{F}\{g\} \cdot \boldsymbol{F}\{f\}\} \tag{9}$$

The asterisk indicates convolution, and $\boldsymbol{F}$ indicates the Fourier transform. The convolution of a function g with a function f is the inverse Fourier transform of the product between the Fourier transform of g and the Fourier transform of f. This relationship will be extensively used. Let the function $f(x, y)$ be mapped onto the function $g\ (x, y)$:

$$g(x, y) = f(x, y) * h(x, y) \tag{10}$$

where $h(x, y)$ is the point-spread function of the system. The Fourier transform of $g(x, y)$ is:

$$G(X, Y) = F\ (X, Y) \cdot H\ (X, Y) \tag{11}$$

where $H(X, Y)$, the Fourier transform of $h(x, y)$, is the transfer function of the system, and $F(X, Y)$ is the Fourier transform of $f(x, y)$. Throughout this

chapter, upper case letters are used for coordinates in Fourier space. The original function $f(x, y)$ can be recovered from $g(x, y)$ by dividing the Fourier transform of $g(x, y)$ by the Fourier transform of $h(x, y)$ followed by an inverse Fourier transform:

$$f(x, y) = \boldsymbol{F}^{-1}\left\{\frac{G(X, Y)}{H(X, Y)}\right\} \text{ for } H \neq 0, \tag{12}$$

The division of the Fourier transform $G(X, Y)$ by $H(X, Y)$ corresponds to a deconvolution of $g(x, y)$ by the function $h(x, y)$.

3. 3D RECONSTRUCTION

The input data to any 3D reconstruction are projections of a 3D distribution. Bright-field images in electron microscopy are in approximation projections of the potential distribution of the object. Use will be made of the projection theorem, which states that the Fourier transform of a projection is a central section through the Fourier transform of the 3D distribution.

For understanding weighted back-projection methods, it is necessary to understand the simple back-projection or summation technique, which will be explained first. A weighted back-projection is a simple back-projection followed by a deconvolution with the point spread function of the simple back-projection algorithm. This deconvolution is done by dividing the Fourier transform of the back-projection by its transfer function. One divided by the transfer function is called the *weighting function*, a term from which the term *weighted back-projection* originates. Convolution back-projections are algorithms that perform the equivalent convolution in real space. The convolution kernel for the real-space convolution is the inverse Fourier transform of the weighting function.

The weighting functions derived in the following sections only corrects for the transfer function of the simple back-projection algorithm. Specifically they do not take into account the limits to the resolution caused by the limited number of projections available. Thus, to obtain a faithful reconstruction of the object, the 3D reconstruction must be low-pass-filtered to the resolution that is determined, by virtue of Shannon's sampling theorem, by the number and spacing of the available projections. The resolution formulae will be given in Section 4.

In the derivation of the analytical form of the weighting function for single-axis tilting, it will be assumed that data are available over the full tilt range from $-\pi/2$ to $+\pi/2$. Distortions of the image point that are caused by the missing wedge in single-axis tilt geometries or the missing cone in conical tilt geometries cannot be corrected by a weighting function or deconvolution, because the Fourier transform of the point-spread function

$H(X, Y, Z)$ becomes 0 in these regions and thus violates the condition stated in equation (12).

Distortions caused by missing data can be reduced, however, using *a priori* knowledge, such as the possible density range or the overall size of the reconstructed object. The technique of projection onto convex sets allows this information to be combined with the reconstruction (e.g. Barth *et al.* 1989; Carazo and Carrascosa 1987; Levi and Stark 1983). Essentially, data consistent with the *a priori* information are inserted into the missing range of the Fourier transform of the reconstructed object, while data that are determined by the experimental projections are left unchanged. *A priori* information can also be used to extend the resolution of the reconstruction by filling in data where the angular sampling is insufficient to fulfill Shannon's sampling theorem. Great caution has to be used, however, with any application of *a priori* information. If the *a priori* information is incorrect, a structure will still emerge, consistent with the data and the *a priori* information; however, it may not reflect the the true structure of the object under investigation. *A priori* information also may bias resolution measurements. If, as is often done, the data set is split in half and two separate reconstructions are calculated that are compared using a resolution criterion that determines the information common to both reconstructions, and both reconstructions are carried out using the same *a priori* information, the resolution measurement may reflect the consistency between the *a priori* information rather than the true resolution of the structure.

A powerful *a priori* information is, for example, a volume constraint, which postulates that outside of a specified region in three dimensions, no structure is present. The imposition of this constraint is comparable with the solvent-flattening approach used in X-ray crystallography. In medical tomography, the object dimensions can be measured. In 3D electron microscopy, the precise outline of the molecule is, in general, not available. For example, when particles are reconstructed that are embedded in negative stain, the surrounding stain distribution is still part of the structure and must be included in any volume mask. Otherwise strong artifacts can appear.

Many of the iterative reconstruction algorithms implicitly use *a priori* information about the object, and the same precautions must be taken. The *a priori* information has to be reliable, and resolution measurements may require an independent reference, for example the X-ray structure of a subregion of the macromolecule.

The weighted back-projection algorithms described here do not make use of *a priori* information. Thus, the resolution and quality of the reconstruction are ruled by Shannon's sampling theorem. The angular sampling must be consistent with the resolution achievable and, after the application of the algorithms described below, the reconstructions must be low-pass-filtered to the resolution following from Shannon's sampling theorem. In all calculations, the pixel sampling of the projections and volumes should be kept twice as fine as required by Shannon's sampling theorem, to avoid

any loss of resolution through limitations in the interpolations used. For example, if the expected resolution in three dimensions is 2 nm, the projections should be scanned with a pixel size not larger than 0.5 nm. In all applications of Fourier filters, the projections should be padded to twice their size to avoid wraparound artifacts.

3.1. The Simple Back-projection Algorithm

An intuitively simple reconstruction method is the back-projection or summation technique. The technique can be explained most easily using the example of a simple 2D binary object consisting of three disc-shaped regions with value 1 inside and value 0 outside. This object is to be reconstructed from its 1D projections (Fig. 1).

The object O (Fig. 1) appears in the projections P at angles θ_1, θ_2, θ_3, θ_4 and θ_5. As a first step in the reconstruction process, the projections are smeared out to form so-called back-projection bodies (B in Fig. 1) (Hoppe *et al.*, 1976). To reconstruct the object, one has to sum all the back-projection bodies.

The simple back-projection calculates only an approximation of the object. In the example in Fig. 1, the original three disks are identifiable in the back-projection because of their higher intensity relative to the surrounding. If the object were not binary, the ambiguity in the back-projection would be even greater. The quality, however, improves when more projections are used. It can be shown that a simple back-projection reconstructs the object with a point-spread function that enhances the low spatial frequency components.

Let $f(x, y, z)$ be a 3D distribution, which is projected under the angles θ_j, ϕ_j to form a series of projections $p_j(x^j, y^j)$. Let $\boldsymbol{r}^j = (x^j, y^j, z^j)$ be the coordinates in the coordinate system of the projection p_j which forms the (x^j, y^j) plane. The geometrical relationship between the object coordinates $\boldsymbol{r} = (x, y, z)$ and $\boldsymbol{r}^j = (x^j, y^j, z^j)$ can be described using the rotation matrices D_{θ_j}, D_{ϕ_j}:

$$\boldsymbol{r}^j = D_{\theta_j} \cdot D_{\phi_j} \cdot \boldsymbol{r} \tag{13}$$

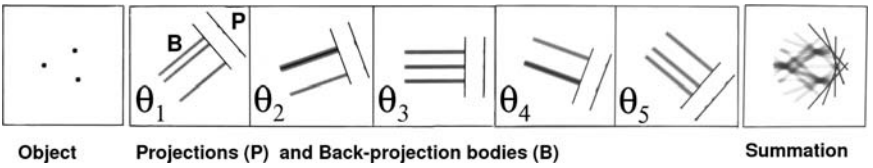


FIGURE 1. Principle of a simple back-projection. A binary object consisting of three points is projected onto five projections P at angles θ_1, θ_2, θ_3, θ_4, θ_5. From these projections, back-projection bodies B are created and the object is reconstructed by summation of these back-projection bodies. The figure shows an actual calculation, where three points have been projected, 'smeared back' (equation 19) and summed (equation 20).

with

$$D_{0_j} = \begin{pmatrix} \cos\theta_j & 0 & -\sin\theta_j \\ 0 & 1 & 0 \\ \sin\theta_j & 0 & \cos\theta_j \end{pmatrix} \tag{14}$$

and

$$D_{\phi_j} = \begin{pmatrix} \cos\phi_j & \sin\phi_j & 0 \\ -\sin\phi_j & \cos\phi_j & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{15}$$

The object is first rotated by ϕ_j around its z-axis, then tilted by $-\theta_j$ around the new y-axis, and then projected onto the (x^j, y^j) plane. Only two angles are needed to describe all projecting directions. A third angle would just describe a rotation in the plane of the projection, which will be disregarded here.

A projection along the direction z^j with angles θ_j, ϕ_j can be written as

$$p_j = \int f(x^j, y^j, z^j) dz^j \tag{16}$$

A back-projection body is formed by convolution of p_j in the (x^j, y^j) plane with the 3D spread function

$$l_j = \delta(x^j, y^j) c(z^j) \tag{17}$$

with

$$c(z^j) = \begin{cases} 1 & \text{for } -a \leqslant z^j \leqslant +a \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

The convolution conditions for an isoplanatic system are fulfilled within a sphere of diameter a if $2 \cdot a$ is at least twice the object diameter and the projections and the reconstruction volume are sufficiently large to include all back-projection rays. Alternatively, for theoretical purposes, projections and back-projection bodies can be considered as being of infinite extent. In practice, the parameter $2 \cdot a$ is chosen to be the object diameter for the calculation of the weighting function and infinite in the back-projection step. All projections and volumes are kept at twice the size of the object during all calculations.

The back-projection body now can be written as

$$\begin{aligned} p_j^b(x^j, y^j, z^j) &= \iint p_j(x^j, y^j) \cdot l(x^j - x'^j, y^j - y'^j, z^j) dx'^j dy'^j \\ &= \iiint f(x'^j, y'^j, z'^j) dz'^j \cdot l(x' - x'^j, y' - y'^j, z^j)\, dx'^j dy'^j \end{aligned} \tag{19}$$

and the back-projection algorithm becomes

$$b(x, y, z) = \sum_j p_j^b(x^j, y^j, z^j) \tag{20}$$

Under the conditions specified above, simple back-projection is a linear and shift-invariant algorithm. Thus, its performance can be described as the convolution of the original object with the point-spread function associated with the algorithm. To obtain a faithful reconstruction, the back-projected density has to be corrected. This correction can be achieved by the convolution with a function that acts as the inverse of the point-spread function. In weighted back-projection methods, this deconvolution is achieved by dividing the Fourier transform of the back-projected density by the transfer function that represents the reconstruction algorithm.

3.2. Weighting Function for Arbitrary Tilt Geometry

To find the weighting function for a weighted back-projection for arbitrary geometry, we must first analyze the point-spread function of a simple back-projection in more detail. The 3D object can be rewritten as a superposition of δ functions, using the 3D equivalent to equation (3):

$$f(x, y, z) = \iiint f(\xi, \eta, \zeta)\delta(x - \xi, y - \eta, z - \zeta)d\xi d\eta d\zeta \tag{21}$$

The point-spread function then can be found by analyzing how the back-projection algorithm affects a single point in three dimensions represented by the function

$$q = \delta(x, y, z) \tag{22}$$

The projection of q at angles θ_j, ϕ_j is

$$p_j(x^j, y^j) = \delta(x^j, y^j) \tag{23}$$

The back-projection body becomes, according to equations (17) and (19),

$$p_j^b(x^j, y^j, z^j) = \delta(x^j, y^j)c(z^j) \tag{24}$$

and the point back-projected in three dimensions is found by summation over θ_j, ϕ_j as

$$b(x, y, z) = \sum_j \delta(x^j, y^j)c(z^j) \tag{25}$$

Thus, $b(x, y, z)$ is the point-spread function of a back-projection calculated from a set of projections taken with arbitrary angles θ_j, ϕ_j. The transfer function is the Fourier transform of (25):

$$\begin{aligned} H(X,Y,Z) &= \boldsymbol{F}[b(x,y,z)] \\ &= \boldsymbol{F}\left\{\sum_j \delta(x^j,y^j)c(z^j)\right\} \\ &= \sum_j \boldsymbol{F}[\delta(x^j,y^j)c(z^j)] \end{aligned} \tag{26}$$

and

$$\begin{aligned} \boldsymbol{F}[\delta(x^j,y^j)c(z^j)] &= \int_{-\infty}^{+\infty}\iint \delta(x^j,y^j)c(z^j)e^{-2\pi i\left(x^jX^j+y^jY^j+z^jZ^j\right)}dx_j dy_j dz_j \\ &= 1\int_{-a}^{+a} e^{-2\pi i z^j Z^j}dz^j = \frac{\sin(2\pi a Z^j)}{\pi Z^j} \\ &= 2a\,\mathrm{sinc}(2a\pi Z^j) \end{aligned} \tag{27}$$

where $\mathrm{sinc}(x) = \sin(x)/x$.

With the rotation matrices given in equations (14) and (15), Z^j can be expressed in the coordinate system (X, Y, Z) of the object:

$$Z^j = X\sin\theta_j\cos\phi_j + Y\sin\theta_j\sin\phi_j + Z\cos\theta_j \tag{28}$$

and the transfer function becomes (Radermacher *et al.*, 1986):

$$H(X,Y,Z) = \sum_j 2a\,\mathrm{sinc}[2a\pi(X\sin\theta_j\cos\phi_j + Y\sin\theta_j\sin\phi_j + Z\cos\theta_j)] \tag{29}$$

Hence, the corresponding weighting function for arbitrary geometry is

$$\begin{aligned} W_a(X,Y,Z) &= \frac{1}{H(X,Y,Z)} \\ &= \left\{\sum_j 2a\,\mathrm{sinc}[2a\pi(X\sin\theta_j\cos\phi_j + Y\sin\theta_j\sin\phi_j + Z\cos\theta_j)]\right\}^{-1} \end{aligned} \tag{30}$$

Equation (30) is valid only for $H \neq 0$. In the implementation of the algorithm, a lower threshold for H is set to avoid division by 0 and to limit the enhancement of noise. While the threshold value 1 would avoid any enhancement, in practice a value of 0.6 has proven as a good guideline, allowing for some enhancement in undersampled areas without creating an undue amplification of noise. The original 3D distribution $o(x,y,z)$ can be recovered from the back-projection $b(x,y,z)$ by multiplication of its Fourier transform $B(X, Y, Z)$ by $W_a(X, Y, Z)$, followed by an inverse Fourier transform:

$$o(x,y,z) = \boldsymbol{F}^{-1}[O(X,Y,Z)] = \boldsymbol{F}^{-1}[B(X,Y,Z)W_a(X,Y,Z)] \tag{31}$$

The weighting function (30), which is three-dimensional, is applied to the 3D Fourier transform of an object reconstructed by simple back-

projection. Since the Fourier transforms of projections are central sections through the 3D Fourier transform of the object and since simple back-projection is essentially a summation which commutes with the Fourier transform operation, the weighting function can also be applied to the Fourier transform of the projections before the back-projection step. However, in an arbitrary tilt geometry, the location of a point in the central section relative to all other central sections is not independent of an in-plane rotation of the projection, and therefore a third Euler angle has to be included in the transformation (13). Only in a random-conical tilt series and simpler geometries can this third angle be kept 0 by aligning the tilt axis parallel to the same axis in each projection. Thus, for the application to the Fourier transforms of the projections, transformation (13) becomes

$$\boldsymbol{r}^j = D_{\psi_j} D_{\theta_j} D_{\phi_j} \boldsymbol{r} \tag{32}$$

Z^j in equations. (28) and (29) becomes the Z component of the vector

$$\boldsymbol{R}^j = D_{\psi_j} D_{\theta_j} D_{\phi_j} D_{\phi_k}^{-1} D_{\theta_k}^{-1} D_{\psi_k}^{-1} \boldsymbol{R}^k \tag{33}$$

where $\boldsymbol{R}^k$ is the vector $(X^k, Y^k, 0)$ in the section in Fourier space at angles ψ_k, θ_k, ϕ_k to which the weighting function is being applied.

The advantage of applying the weighting function to the projection is that in most cases H is non-zero along the corresponding sections through the 3D Fourier transform, because each section is also the origin of one of the superimposed sinc-functions in H. The disadvantage of this mode of application is that the number of calculations needed is usually larger. Assuming that the dimensions of the Fourier transforms of projections and object are equal to their real-space dimensions, a common situation is that the xy dimensions of the projections are larger than the xy dimensions of the object and, secondly, that the number of projections is larger than the number of z slices in the reconstructed object. The most appropriate implementation of the algorithm is one that allows a choice of the mode of application.

In practical applications the weighting function (equation 30) performs well if the number of projections per angular interval varies smoothly. If, however, large gaps exist in the angular sampling and the number of projections per angular interval varies rapidly, then fringes may occur in the reconstruction. These artefacts can be eliminated by the use of a weighting function that is the Fourier transform of an apodized point-spread function. The discontinuous step function $c(z^j)$ in equation (18), which describes the length of the rays in the back-projection step, is replaced by a continuous step function with Gaussian fall-off. This function can be created by a convolution of the discontinuous step function with a Gaussian function. Because of the convolution theorem, its Fourier transform is the product of the sinc function (equation 18) and the Fourier transform of the

Gaussian, which is also a Gaussian function. For the calculation of the weighting function based on the apodized point-spread function, the elements in the sum in equation (29) are replaced by products of sinc functions and Gaussians.

3.3. Weighting Function for Single-axis Tilt Geometry with Equal Angular Increments

The weighting function derived in the previous section is applicable to any tilt geometry. However, for a single-axis tilt geometry with equal angular increments, an analytical form of the weighting function can be found. Although the weighting function for single-axis tilting can be derived from the inverse Radon transform or a comparison between a simple back-projection and a Fourier inversion method (e.g. Cormack, 1963, 1964; Gilbert, 1972; Ramachandran and Lakshminarayanan, 1971; Vainshtein and Orlov 1972; and Section 5 of this chapter), it will be derived here as a special case of the weighting function for arbitrary geometry.

In a single-axis tilt geometry, the object is tilted around a single axis in equal angular increments, and a projection is recorded for each orientation. Without loss of generality, the tilt axis is assumed to be the y-axis (Fig. 2); thus θ in equation (14) becomes the tilt angle, and ϕ in equation (15) is kept at 0. Now we make the following approximations: (i) the variable a in

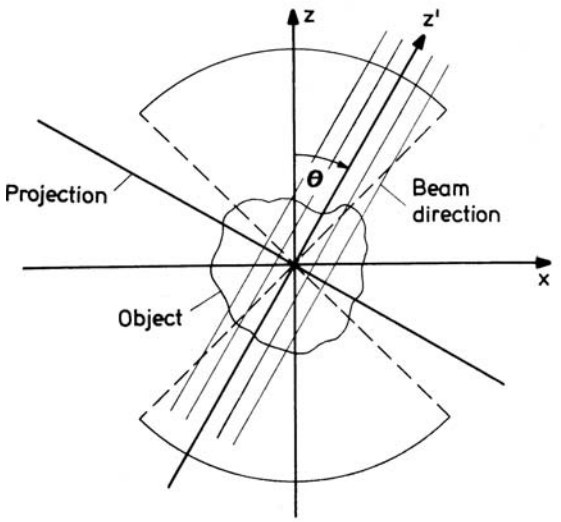


FIGURE 2. Single-axis tilt geometry. The tilt axis y is perpendicular to the plane of the page, θ is the tilt angle, x, y, z the coordinate system fixed to the object, x', y', z' is the coordinate system, with x', y' being the plane of the projection. (From Radermacher, 1988; reproduced with permission of Alan R. Liss, Inc.)

equation (18) is assumed to be infinite, which corresponds to back-projection bodies that are infinitely extended in the z^j direction; and (ii) the series of projections is available over a continuous range of θ from –90° to +90°. With these approximations, and with

$$\lim_{a\to\infty} a\,\mathrm{sinc}(a\cdot x) = \delta(x) \tag{34}$$

equation (29) becomes

$$H(X,Y,Z) = \sum_j \delta(X\sin\theta_j + Z\cos\theta_j) \tag{35}$$

From condition (ii) the sum in equation (35) can be replaced by an integral:

$$H(X,Y,Z) = \int_{-\pi/2}^{+\pi/2} \delta(X\sin\theta + Z\cos\theta)\,d\theta \tag{36}$$

Since the tilt axis is assumed to be the y-axis, $H(X, Y, Z)$ is independent of Y. The transfer function is constant in the direction of the tilt axis and varies only in the planes perpendicular to Y. If the coordinates X and Z are replaced by the cylindrical coordinates R and Γ within the XZ planes.

$$X = R\cos\Gamma, \quad Z = R\sin\Gamma, \quad Y = Y \tag{37}$$

equation (36) becomes

$$H(R,Y,\Gamma) = \int_{-\pi/2}^{+\pi/2} \delta(R\cos(\Gamma-\Theta))\,d\Theta \tag{38}$$

The solution to this integral can be found using the identity

$$\delta(f(x)) = \sum_n \frac{1}{|f'(x_n)|}\delta(x-x_n), \quad \text{with} \quad f'(x_n) = \left.\frac{df(x)}{dx}\right|_{x=x_n} \tag{39}$$

for $f'(x_n) \neq 0$. The x_n are the zeros of $f(x)$ within the integration range. There is only one zero of f within the integration interval $\Theta\varepsilon(-\pi/2, \pi/2]$; $f(\Theta) = 0$ for $\Theta = \Gamma - \pi/2$, and the solution of (38) becomes

$$H(R,Y,\Gamma) = \frac{1}{R} \tag{40}$$

We thus obtain the result that, for the deconvolution of a simple back-projection from a single-axis tilt series with equal angular increments, the Fourier transform of the back-projections has to be multiplied by the weighting function

$$W_s(R, Y, \Gamma) = R \tag{41}$$

where R is the radius in Fourier space perpendicular to the tilt axis Y. Again the multiplication by R can be applied either to the projections or directly to the 3D Fourier transform of the back-projection. If applied to the projections, the application of the weighting function corresponds to a multiplication of their Fourier transform by X.

3.4. Weighting Function for Conical Tilt Geometry with Equal Angular Increments

For collection of a conical tilt series in the microscope the specimen is tilted by a fixed angle θ_0 and then rotated in this inclined plane by equal angular increments $\Delta\phi$. In each position, a projection is recorded. The geometry is equivalent to moving the beam direction along the surface of a cone (Fig. 3).

To calculate the geometrical relationship between the coordinate system in the projections and the coordinate system fixed to the object, again the rotation matrices (14) and (15) are employed with $\theta_j = \theta_0$ constant, and $\phi_j = j\Delta\phi$. To find the analytical form of the weighting function, the same assumptions are made as in the calculation of the analytical weighting for single-axis tilting, i.e. the back-projection bodies are infinitely extended in the z_j direction, and the azimuthal angle ϕ is assumed to be continuous. As before, this results in a replacement of the sinc function in equation (29) by a δ function and a replacement of the sum by an integral. The transfer function of a simple back-projection from a regular conical tilt series thereby becomes

$$\begin{aligned} H(X,Y,Z) &= \int_{-\pi}^{\pi} \delta(Z_\phi)\, d\phi \\ &= \int_{-\pi}^{\pi} \delta[X \sin\theta_0 \cos\phi + Y \sin\theta_0 \sin\phi + Z \cos\theta_0]\, d\phi \end{aligned} \qquad (42)$$

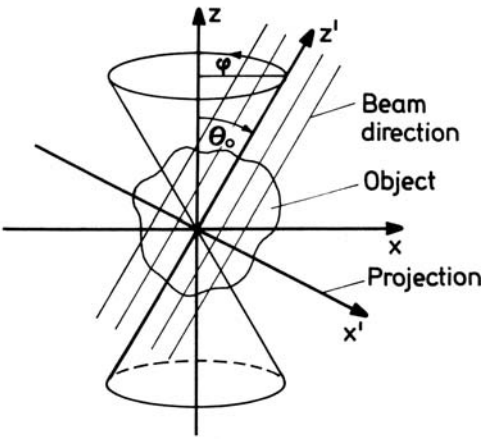


FIGURE 3. Conical tilt geometry. As in Fig. 2, the y-axis and y'-axis are perpendicular to the plane (viewing direction at $\phi = 0$). x, y, z, are object-fixed coordinates, θ_0 the fixed tilt angle around the y-axis and ϕ the azimuthal angle around z. (From Radermacher, 1988; reproduced with permission of Alan R. Liss, Inc.)

By expressing X, Y, Z in cylindrical coordinates R, Γ, Z and making use of equation (39), we finally obtain (Radermacher, 1980; Radermacher and Hoppe, 1978):

$$H(R,\Gamma,Z)=\frac{2}{|R\sin\theta_0\sin\arccos[-Z\cot(\theta_0)/R])|} \tag{43}$$

Equation (43) is valid in the region $R > Z\cot\theta_0$, i.e. only in the region of the Fourier transform where measured data are available. It is not valid within the missing cone. Because $0 \leqslant \theta_0 \leqslant \pi/2$,

$$\arccos\left(\frac{-Z\cot\theta_0}{R}\right)=\arcsin\left(\sqrt{1-\frac{Z^2\cot(\theta_0)}{R^2}}\right) \tag{44}$$

and equation (43) becomes

$$H(R,\Gamma,Z)=\frac{2}{\sqrt{R^2\sin^2\theta_0-Z^2\cos^2\theta_0}} \tag{45}$$

The weighting function that needs to be applied to the 3D Fourier transform of the back-projection from a regular conical tilt series then becomes

$$W_c(R,\Gamma,Z)=\frac{1}{2}\sqrt{R^2\sin^2\theta_0-Z^2\cos^2\theta_0} \tag{46}$$

Here, as in the previous cases, the weighting function may also be applied to the projections before the back-projection is carried out. In the central section in Fourier space that corresponds to the projection at the angles θ and ϕ, equation (47) becomes

$$W_c(X^j, Y^j) = Y^j\sin\theta_0 \tag{47}$$

Equation (47) was derived by reverting to Cartesian coordinates in equation (46) and transforming to the coordinate system of the projection at angles θ_0, ϕ_j using

$$\boldsymbol{r}^j = D_{\theta_0}D_{\phi_j}\boldsymbol{r} \tag{48}$$

with D_ϕ and D_θ as previously defined in equations (14) and (15).

For a reconstruction from a conical tilt series with equal angular increments, the weighting step corresponds to a multiplication of the Fourier transforms of the projections by $Y\sin\theta_0$, where y is the direction of the tilt axis belonging to the angle θ_0. In contrast to the weighting function for

single-axis tilt geometry, which only affects the projection perpendicular to the tilt axis, for conical tilting the weighting function increases in the direction of the tilt axis and is constant perpendicular to it.

3.5. Other Forms of Weighting Functions

For reconstruction from projections with arbitrary projecting angles, a weighting function different from the one derived in equation (29) has been proposed (Harauz and van Heel, 1986). The filter, which the authors term 'exact filter', is an approximation to the function shown here. The essential difference is that the sinc function is replaced by the simpler triangular function. The use of the triangular function in the summation of the single projection's contribution to the weighting function poses smaller computational demands and thus results in a faster performance of the algorithm. However, if the sine function that appears in equation (29) as part of the sinc function calculation is tabulated, the difference in computational effort between the two approaches should be minimal.

For single-axis tilting, a large variety of weighting functions can be found that combine the R weighting with a window of the Fourier transform that corresponds to a low-pass filter (e.g. Kwok *et al.*, 1977; Ramachandran and Lakshminarayanan, 1971; Suzuki, 1983). The inclusion of a window function $w(R)$, however, becomes necessary if the deconvolution is to be carried out in real space. The real-space convolution kernel is calculated as the inverse Fourier transform of $R \cdot w(R)$. As R is not square-integrable and does not have a compact support, its inverse Fourier transform does not exist. The window function is needed to limit the weighting function to a compact support, make it square-integrable and, possibly, to carry out this inverse Fourier transform.

3.6. A Variant of Weighted Back-projection, the Two-step Radon Inversion

A variation of the weighted back-projection algorithms described above is the *Two-step Radon Inversion Algorithm* in which two consecutive R-weighted back-projections are employed to invert a 3D Radon transform. This algorithm can be used for the reconstruction from projections with any geometry. The problem of accounting for the different projection direction is solved in the step of creating a 3D Radon transform from the projection set. This 3D Radon transform is sampled in a regular spherical coordinate system.

As will be shown in Section 5, a series of 2D projections of a 3D object does not represent the 3D Radon transform of the object. To calculate the 3D Radon transform, first the 2D Radon transforms of all projections have to be calculated. Each single line in these 2D transforms represents a line through the origin of the 3D Radon transform. The orientation of each of these lines in three dimensions is known from the projection angles and

from their angular coordinate within the 2D Radon transform. In a large data set, many of these lines represent redundant measurements of the 3D transform. This 3D Radon transform can then be calculated by an averaging procedure. The result is a 3D Radon transform, averaged from the measurements and stored in a polar coordinate system (r, θ, ϕ). The inversion can now be carried out in two steps: first, all planes (r, ϕ) (θ = const) are sectioned out of the 3D Radon transform. An R-weighted back-projection is used to calculate the real-space image of these sections. Each reconstruction section now represents a 2D projection of the object at a specific angle θ. These reconstructed slices form a single-axis tilt projection series of the 3D object, which can be retrieved with a second R-weighted back-projection. Detailed descriptions of these algorithms can be found in Radermacher (1994, 1997) and in Bellon *et al*. 1999.

4. BAND LIMIT OF A RECONSTRUCTION FROM A LIMITED NUMBER OF PROJECTIONS

For all reconstructions, only a discrete, limited number of projections is available. If no band limit is imposed on the reconstruction, then there exists an infinite number of possible reconstructions that are all consistent with the available measurements, and the result obtained would be just one of these possible solutions. However, by using Shannon's sampling theorem (Shannon, 1949), a band limit can be determined up to which the reconstruction is unique, presenting a faithful image of the original object. The band limitation is not taken into account by the weighting functions derived above and thus needs to be imposed in a separate step. Like the weighting function, the low-pass filter can be applied either to the projections before the reconstruction or to the reconstructed object. The application to the projections has the advantage that it avoids multiple Fourier transformations, since it can be done simultaneously with the application of the weighting function. On the other hand, the application of the low-pass filter to the reconstruction allows for more flexibility. It further reduces any high resolution artifacts that may have been caused by interpolation during the back-projection step. The low-pass filter can be adjusted, for example, to the resolution predicted by the sampling or to a measured resolution value (see, for example, Cardone *et al*., 2005).

4.1. Resolution for Single-axis Tilt Geometry

The equation for the dependence of the resolution on the number of projections in a single-axis tilt geometry with equal angular increments has been derived by Bracewell and Riddle (1967) and Crowther *et al.* (1970). Let d be the resolution of he reconstruction and D the diameter of the reconstruction volume, which is assumed to be cylindrical, the axis of the cylinder coinciding with the tilt axis. The reconstruction volume is defined

as the volume large enough to include the complete object. Let $\Delta\theta$ be the angular increment, and let $N = \pi/\Delta\theta$. Then the best resolution achievable is

$$d = \pi \frac{D}{N} \quad \text{for a tilt range from } -90° \text{ to } +90° \tag{49}$$

or

$$d = D\Delta\theta \tag{50}$$

4.2. Resolution for Conical Tilt Geometry

Let θ_0 be the fixed tilt angle, d the resolution, D the object diameter and N the number of projections evenly spaced in ϕ. The resolution of the reconstruction then depends on the number of projections in the following way (Radermacher, 1980, 1988; Radermacher and Hoppe, 1980):

$$d = 2\pi \frac{D}{N} \sin\theta_0 \quad \text{for } N \text{ even} \tag{51}$$

and

$$d = \frac{2\pi D \sin\theta_0 \cos\theta_0}{N\sqrt{\sin^2\theta_0 (\pi/2N)^2 + \cos^2\theta_0}} \quad \text{for } N \text{ odd} \tag{52}$$

Inversion of (52) yields

$$N = \frac{\pi D}{2d} \tan\theta_0 \sqrt{16\cos^2\theta_0 - (d/D)^2} \tag{53}$$

The resolution formulae above do not take into account the missing regions in the Fourier transform of the object caused by the limited tilt range for single-axis tilting or by $\theta_0 < 90°$ for conical tilting. In a single-axis tilt series with a limited tilt range around the y-axis from $-\alpha_{max}$ to $+\alpha_{max}$ in increments of $\Delta\theta$, equation (49) gives the resolution in the x direction. The resolution in the z-direction depends on the maximum tilt angle, and the resolution parallel to the tilt axis y is equal to the resolution of the input projections. Similarly, for a conical tilt series, where the fixed tilt by the angle θ_0 is around y, equations (51) and (52) give the resolution in the directions parallel to the xy plane, whereas the resolution in the z direction depends on the size of the missing cone (Radermacher, 1980, 1988).

4.3. Resolution in a Flat Extended Reconstruction Volume

For the derivation of equations (49–52) a spherical shape of the reconstruction volume has been assumed, which is appropriate for most

reconstructions of macromolecular assemblies. However, especially in reconstructions of subcellular components in thick sections, the assumption of a flat extended slab for the shape of the reconstruction volume is more appropriate. Equations (49–52) could be applied by using for D the diameter of a sphere that circumscribes the complete slab. However, the resulting number of projections necessary to obtain a reconstruction with a reasonable resolution becomes prohibitive if calculated this way. Indeed, this approach is unnecessarily conservative. If the image point is analysed more closely, it can be seen that, for a given resolution d, the diameter D in any equation (49–52) is the radius of the volume surrounding a single image point where artifacts are minimal. If the object is considered as being built up by a superposition of image points, each of which carries a halo of rays in the directions of the contributing projections (Fig. 4), starting at a distance D from the centre of the image point, then a faithful reconstruction can be calculated up to a section thickness of

$$T = D\cos\gamma \tag{54}$$

where γ is the maximum tilt angle $\alpha_{\max}$ for single-axis tilting or the fixed angle θ_0 for conical tilting. Substitution of

$$D = \frac{T}{\cos\gamma} \tag{55}$$

in equations (49–52) will therefore give the resolution of a reconstruction of an extended slab with thickness T.

4.4. Resolution for Random and Random-conical Geometry

For reconstructions from randomly distributed projections, no formulae are available that predict the theoretical value of the final resolution. For reconstructions from a random-conical tilt series, however, the angular coverage can be analysed and a conservative estimate of the resolution can be made by using the largest angular step between two adjacent projections as the value for $\Delta\theta = 2\pi/N$ in equation (52).

4.5. Resolution in the z-direction

The resolution values that can be estimated with the equations above only apply to specific directions in the reconstruction. In a single-axis tilt series, the resolution in the direction of the tilt axis is essentially equal to the resolution of the micrographs, reduced by radiation damage that could introduce inconsistencies in the projection set, and by possible inaccuracies in the alignments of the projections before reconstruction. In the x direction, we find the resolution determined by equation (50). The resolution in z direction depends on the size of the missing wedge, the part of the angular

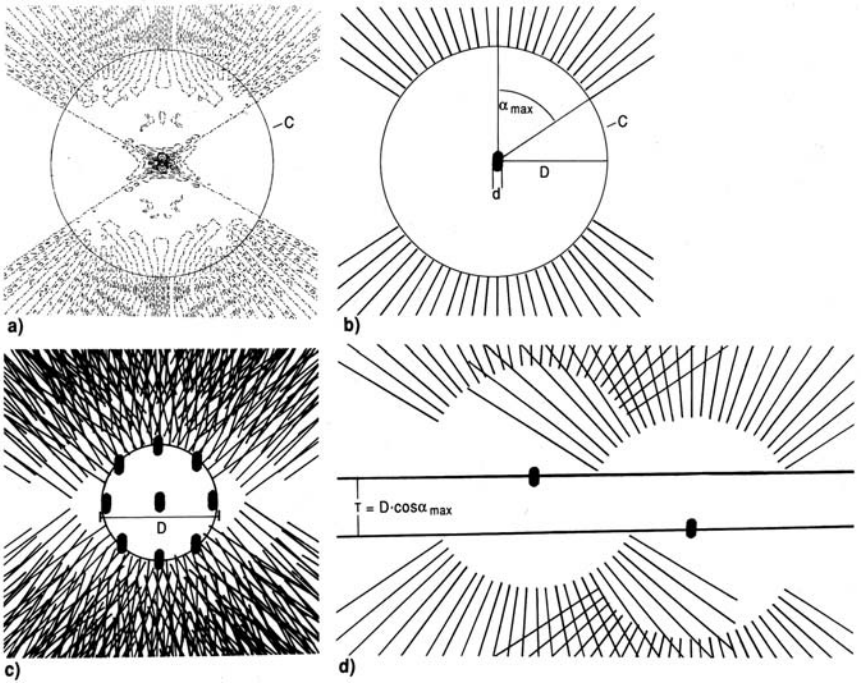


FIGURE 4. Reconstruction of an object within a spherical reconstruction volume and within an extended slab. (a) Contour plot of a cross-section through a 3D image point reconstructed from a single-axis tilt series containing 25 projections in the range of –60° to +60°, using *R*-weighted back-projection. Contour levels: dashed lines 2% increments <0 to 8%; dashed-dotted lines 0 contour; solid lines at 10, 20, 40 and 100%. Percentages are relative to the image maximum. (b) sketch of the image point in (a). C-circle with radius *D* (= diameter of the reconstruction volume as used in equation 54). *d*-resolution. (c) Superposition of reconstructed image points on the surface of a sphere with diameter *D*. The inside of the sphere is virtually artifact free. (d) Superposition of two image points at the surface of an extended slab. Within the thickness *T* of the slab, no overlap of the artifacts surrounding the image points occurs.

space that is not covered by projections. For conical and random-conical tilting, the equations above describe the resolution in the *x*–*y* plane and, again, the resolution in *z* direction is lower because of the limited tilt range.

In single axis tilting, the tilt range is typically ±60° and for conical and random-conical tilting, the maximum tilt angle typically is 55–60°. The missing wedge or missing cone in the 3D Fourier transform creates artifacts in the real-space reconstruction. Part of these artefacts are spurious features, defined as clutter (Radermacher, 1980), and the more obvious effect is a loss of resolution in the *z* direction (the direction perpendicular to the specimen plane). This resolution loss can be estimated as a ratio of the resolution in *z* direction to the resolution in the *x*–*y* plane. By approximation of the image point with a 3D ellipsoid, one can predict its axis ratio, depending on the two ellipsoid angles. Using this approach, the resolutions in the *z* direction are:

for single-axis tilting

$$e_{xz} = \sqrt{\frac{\alpha + \sin\alpha \cdot \cos\alpha}{\alpha - \sin\alpha \cdot \cos\alpha}}, \tag{56}$$

with α being the maximum tilt angle, and for conical tilting

$$e_{xz} = \sqrt{\frac{3 - \sin^2\theta_0}{2 \cdot \sin^2\theta_0}}, \tag{57}$$

with θ_0 being the fixed tilt angle in the conical tilt series (Radermacher, 1980, 1988).

5. RELATIONSHIP BETWEEN THE INVERSE RADON TRANSFORM, FOURIER INVERSION AND WEIGHTED BACK-PROJECTION METHODS

The inverse Radon transform (Cormack, 1963, 1964; Radon, 1917), Fourier inversion methods and weighted back-projection are equivalent formulations of the reconstruction problem in a single-axis tilt geometry. This is true for a single-axis tilt geometry only, where the problem can be reduced to the problem of reconstructing a 2D image from 1D projection.

The n-dimensional Radon transform is obtained by integration over (n – 1)-dimensional hyperplanes which for $n = 2$ results in line integrals. For $n = 3$, the integration has to be carried out over planes. Projections obtained in the electron microscope correspond to line integrals of the potential distribution of the object. In a single-axis tilt series with tilt axis y, the set of lines taken from each projection at the same coordinate y_0 form the 2D Radon transform of the xz section through the object at the location y_0. In a conical or arbitrary tilt geometry, the reconstruction problem can no longer be reduced to a set of planar reconstructions from a set of lines extracted from the projections. The set of 2D projections in a conical tilt geometry does not constitute the 3D Radon transform of the object. The latter would require a set of integrals over all planes and not the set of line integrals as is found in 2D projections. A method for reconstruction from a set of 1D projections that form a 3D Radon transform, however, can be found in earlier 3D imaging techniques applied in nuclear magnetic resonance imaging (Shepp, 1980).

To show the equivalence between Radon inversion, Fourier inversion and weighted back-projection methods, we follow the book of Deans (1983, p. 96ff) (see also, for example, Crowther *et al.*, 1970; Gilbert, 1972; Ramachandran and Lakshminarayanan, 1971; Smith *et al.*, 1973; Vainshtein and Orlov, 1972; Zwick and Zeitler, 1973.)

The 2D Radon transform of an object $f(\boldsymbol{r})$, $\boldsymbol{r} \in \boldsymbol{R}^2$ is the function

$$\tilde{f}(p, \boldsymbol{\xi}) = \int f(\boldsymbol{r}) \delta(p - \boldsymbol{\xi} \cdot \boldsymbol{r}) d\boldsymbol{r} \tag{58}$$

which is equivalent to a line integral along the line $p = \boldsymbol{\xi} \cdot \boldsymbol{r}$, with $\boldsymbol{\xi}$ being the unit vector in the direction of the line, $\boldsymbol{\xi} = (\cos\theta, \sin\theta)$, and the integration is carried out over the plane in $\boldsymbol{R}^2$. The Fourier transform of $f(\boldsymbol{r})$ is

$$\begin{aligned} F(\boldsymbol{R}) = \boldsymbol{F}_2\{f(\boldsymbol{r})\} &= \int f(\boldsymbol{r}) e^{-2\pi i \boldsymbol{r}\cdot\boldsymbol{R}} d\boldsymbol{r} \\ &= \int f(\boldsymbol{r}) e^{-2\pi i t} \delta(t - \boldsymbol{R}\cdot\boldsymbol{r}) dt d\boldsymbol{r} \end{aligned} \tag{59}$$

where $\boldsymbol{F}_2$ indicates the 2D Fourier transform. The Fourier transform along sections in the direction of the unit vector $\boldsymbol{\xi}$ can be found by replacing $\boldsymbol{R}$ with $s\boldsymbol{\xi}$. Substituting sp for t then yields

$$\begin{aligned} F(s\cdot\boldsymbol{\xi}) &= |s| \int_{-\infty}^{+\infty} \int f(\boldsymbol{r}) e^{2\pi i s p} \delta(sp - s\boldsymbol{\xi}\cdot\boldsymbol{r}) d\boldsymbol{r}\, dp \\ &= \int_{-\infty}^{+\infty} \int f(\boldsymbol{r}) \delta(p - \boldsymbol{\xi}\cdot\boldsymbol{r}) d\boldsymbol{r}\, e^{-2\pi i s p} dp \end{aligned} \tag{60}$$

Here, $g(p,\boldsymbol{\xi}) = \int f(\boldsymbol{r}) \delta(p - \boldsymbol{\xi}\cdot\boldsymbol{r}) d\boldsymbol{r}$ is the Radon transform of $f(r)$ or the 1D projection of f along $\boldsymbol{\xi}$. Equation (58) essentially is the projection theorem, which states that the 2D Fourier transform of the function $f(x, y)$ is the 1D Fourier transform along p of the 2D Radon transform of f. The function $f(y)$ can be recovered by application of an inverse 2D Fourier transform to the 1D Fourier transform along p of the Radon transform. This is the basis for all Fourier reconstruction methods.

The equivalence of the weighted back-projection method and Fourier reconstruction methods can be seen from the following. Let $p_\phi(q)$ be a 1D projection of a 2D object $o(x, y)$ at the angle ϕ. In the coordinate system of the object, we have $q = x\cos\phi + y\sin\phi$. Let $f(q, \phi)$ be the function describing the set of projections for a continuous angle ϕ. The back-projection in its continuous form then is

$$b(x, y) = \int_0^{\pi} f(x\cos\phi + y\sin\phi, \phi) d\phi = \int f(q(x, y, \phi), \phi) d\phi \tag{61}$$

Let $F(\boldsymbol{R})$ be the 1D Fourier transform of $f(q, \phi)$ along q.

$$F(R, \phi) = \int_{-\infty}^{+\infty} f(q, \phi) e^{-2\pi i q R} dq \tag{62}$$

Replacing f in equation (61) by the inverse Fourier transform of F yields

$$\begin{aligned} b(x, y) &= \int_0^{\pi} \int_{-\infty}^{+\infty} F(R, \phi) e^{2\pi i R q(x, y, \phi)} dR d\phi \\ &= \int_0^{\pi} \int_{-\infty}^{+\infty} F(R, \phi) e^{2\pi i R q(x\cos\phi + y\sin\phi)} dR d\phi \end{aligned} \tag{63}$$

and after introduction of polar coordinates r, γ with $x = r\cos\gamma$, $y = r\sin\gamma$, equation (63) becomes

$$\begin{aligned} b(r, \gamma) &= \int_0^{\pi} \int_{-\infty}^{+\infty} F(R, \phi) e^{2\pi i R r\cos(\gamma - \phi)} dR d\phi \\ &= \int_0^{\pi} \int_{-\infty}^{+\infty} \frac{1}{R} F(R, \phi) e^{2\pi i R r\cos(\gamma - \phi)} R dR d\phi \end{aligned} \tag{64}$$

which is the 2D inverse Fourier transform of $F(R, \phi)$ multiplied by $1/R$. The weighting function for single-axis tilting, R, compensates for this factor. Thus, a weighted back-projection is equivalent to a reconstruction by Fourier inversion methods.

ACKNOWLEDGEMENTS

I wish to thank Joachim Frank for many discussions and critically reading the original manuscript. I thank David Patlak, without whose help this second edition would not have been possible. This work was supported by NIH grants # 1R01 GM29169, 1R01 GM068650 and NSF grant #8313405.

A. APPENDIX: NOTES ON THE COMPUTER IMPLEMENTATION OF THE ALGORITHMS

Many schemes of computer implementation of algorithms can be found in the literature on numerical methods, and the following does not claim to be the most efficient way of programming the algorithms outlined in the foregoing chapter, but might be helpful in translating the equations into program code.

A.1. Implementation of the Simple Back-projection Algorithm

For the calculation of a simple back-projection, the projection bodies are summed into the 3D volume. Both the volume and the projections are available on a discrete sampling grid, and an interpolation is necessary to sum the rays of the back-projection body into the volume. One possible solution uses the following scheme. In a loop over the volume elements the coordinates of each point in the volume are calculated in the projection P_j at the angles θ_j, ϕ_j. The value of the projection at this point then is interpolated from the values of the surrounding points in the projection and added to the volume element. For a bilinear interpolation in the back-projection, the formula is

$$\begin{aligned} O(m,n,o) = \sum_i & P_i(p,q)(1-\Delta x)(1-\Delta y) + P_i(p+1,q)\Delta x(1-\Delta y) \\ & + P_i(p,q+1)(1-\Delta x)\Delta y + P_i(p+1,q+1)\Delta x\Delta y \end{aligned} \tag{65}$$

where $O(m, n, o)$ is the volume element at x, y, z indices m, n, o with $-M \leqslant m \leqslant +M$, $-N \leqslant n \leqslant +N$ and $-O \leqslant o \leqslant +O$. Here, p and q are the

x and y indices of pixels in the projections and are calculated using the matrices (14) and (15):

$$x = m\cos\theta_i\cos\phi_i + n\cos\theta_i\phi_i - o\sin\theta_i$$
$$p = \text{integer truncation of } x,\ \Delta x = x - p \tag{66}$$

and

$$y = -m\sin\phi_i + n\cos\phi_i$$
$$q = \text{integer truncation of } y,\ \Delta y = y - q \tag{67}$$

Of course, other interpolation schemes can be used, the fastest but crudest being a nearest-neighbour selection, and a much more elaborate one is a spline interpolation.

A.2. Implementation of the Weighting Scheme for Arbitrary Geometry

The following contains two sketches for possible layouts of a program that calculates and applies the weighting function for arbitrary geometry. One shows the application of the weighting function to the projections, and the second shows the application of the weighting function to the volume after simple back-projection (equations 28, 29, 31).

A.2.1. Application to the Fourier Transform of the Projections

Read the angles ψ_j, θ_j, ϕ_j for all projections
Loop 1 over all projections p_k, $1 \leqslant k \leqslant J$.
calculate the transfer function $H_k(X^k, Y^k)$ for projection p_k at angles θ_k, ϕ_k.
Loop 2 over all Fourier coordinates X^k, Y^k in the Fourier plane corresponding to the projection p_k.
Loop 3 over all projection angles ψ_j, θ_j, ϕ_j;
$1 \leqslant k \leqslant J$
calculate the Z_j coordinate of point X^k, Y^k, i.e. the Z coordinate of point $\boldsymbol{R}^k$ in the coordinate system of projection p_j, at angles ψ_j, θ_j, ϕ_j using
$\boldsymbol{R}^j = D_{\psi_j} D_{\theta_j} D_{\phi_j} D_{\phi_k}^{-1} D_{\theta_k}^{-1} D_{\psi_k}^{-1} \boldsymbol{R}^k$
calculate $H(X^j, Y^j) = H(X^j, Y^j) + 2a\,\mathrm{sinc}(2\pi a Z^j)$
end of loop 3 over projection angles
If $H(X^j, Y^j) <$ threshold, then $H(X^j, Y^j) =$ threshold
end of loop 2 over Fourier plane coordinates
Fourier transform projection $P_k(X^k, Y^k) = \boldsymbol{F}(p_k(x^k, y^k))$ compute $P_k(X^k, Y^k)/H(X^k, Y^k) := P_k^w(X^k, Y^k)$.
Inverse Fourier-transform: $\boldsymbol{F}^{-1}\{P_k^w(X^k, Y^k)\} := p_k^w(x, y)$.
end of loop 1 over projections.

A.2.2. Application to the 3D Fourier Transform

Read the angles ψ_j, θ_j, ϕ_j for all J projections

Loop 1 over all Fourier coordinates X, Y, Z in the 3D Fourier transform.

Loop 2 over all J projection angles ψ_j, θ_j, ϕ_j.

calculate the Z^j coordinate of point $\boldsymbol{R} = (X, Y, Z)$, i.e. the z coordinate of point $\boldsymbol{R}$ in the coordinate system of projection p_j at angles θ_j, ϕ_j (ψ can be neglected) using $\boldsymbol{R}^j = D_{\theta_j} D_{\phi_j} \boldsymbol{R}$

calculate $H(X, Y, Z) = H(X, Y, Z) + 2a\,\mathrm{sinc}(2\pi a Z^j)$

end of loop 2 over projection angles.

If $H(X, Y, Z) <$ threshold, then $H(X, Y, Z) =$ threshold

end of loop 1 over Fourier coordinates

Fourier transform back-projected volume: $\boldsymbol{F}\{o^b(\boldsymbol{r})\} := O^b(\boldsymbol{R})$

Divide: $O^b(\boldsymbol{R})/H(\boldsymbol{R}) := O(\boldsymbol{R})$

Inverse Fourier transform: $\boldsymbol{F}^{-1}\{O(\boldsymbol{R})\} := o(\boldsymbol{r})$

REFERENCES

Barth, M., Bryan, R. K. and Hegerl, R. (1989). Approximation of missing-cone data in electron microscopy. *Ultramicroscopy* **31**:365–378.

Bellon, P. L., Cantele, F., Kreman, M., Lanzavecchia, S., Wright, E., Zampighi, G. A. and Zampighi, L. (2005). Conical tomography of freeze-fracture replicas: a method for the study of integral membrane proteins inserted in phospholipid bilayers. *J. Struct. Biol.* **149**:87–98.

Bellon, P. L. and Lanzanvacchia, S. (1997). Fast direct Fourier methods, based on one- and two-pass coordinate transformations, yield accurate reconstructions of X-ray CT clinical images. *Phys. Med. Biol.* **42**:443–463.

Bellon, P. L., Lanzavecchia, S. and Radermacher, M. (1999). Fast and accurate three-dimensional reconstruction from projections with random orientations via Radon Transforms. *J. Struct. Biol.* **128**:152–164.

Bracewell, R. N. and Riddle, A. C. (1967). Inversion of fan-beam scans in radio astronomy. *Astrophys. J.* **150**:427–434.

Carazo, J. M. and Carrascosa, J. L. (1987). Information recovery in missing angular data cases: an approach by the convex projections method in three dimensions. *J. Microsc.* **45**:23–43.

Carazo, J.-M., Wagenknecht, T. and Frank, J. (1989). Variations of the three-dimensional structure of the *Escherichia coli* ribosome in the range of overlap views. *Biophys. J.* **55**:465–477.

Cardone, G., Grünewald, K. and Steven, A. C. (2005). A resolution criterion for electron tomography based on cross-validation. *J. Struct.* Biol. **151**:117–129.

Colsher, J. G. (1977). Iterative three-dimensional image reconstruction from tomographic projections. *Comput. Graph. Image Proc.* **6**:513–537.

Cormack, A. M. (1963). Representation of a function by its line integrals, with some radiological applications. *J. Appl. Phys.* **34**:2722–2727.

Cormack, A. M. (1964). Representation of a function by its line integrals, with some radiological applications. II. *J. Appl. Phys.* **35**:2908–2913.

Crowther, R. A., DeRosier, D. J. and Klug, A. (1970). The reconstruction of a three dimensional structure from projections and its application to electron microscopy. *Proc. R. Soc. A* **317**:319–340.

Deans, S. R. (1983). *The Radon Transform and Some of Its Applications.* Wiley, New York.
Frank, J., Carazo, J.-M. and Radermacher, M. (1988). Refinement of the random conical reconstruction technique using multivariate statistical analysis and classification. *Eur. J. Cell Biol. Suppl. 25* **48**:143–146.
Frank, J. and Goldfarb, W. (1980). Methods for averaging of single molecules and lattice fragments. in *Electron Microscopy at Molecular Dimensions* (W. Baumeister and W. Vogell, eds). Springer-Verlag, Berlin, pp. 261–269.
Frank, J., McEwen, B. F., Radermacher, M., Turner, J. N. and Rieder, C. L. (1987). Three-dimensional tomographic reconstruction in high-voltage electron microscopy. *J. Electron Microsc. Tech.* **6**:193–205.
Gilbert, P. F. C. (1972). The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. II: Direct methods. *Proc. R. Soc. B* **182**:89–102.
Goodman, J. W. (1968). *Introduction to Fourier Optics.* McGraw-Hill, New York.
Gordon, R., Bender, R. and Herman, G. T. (1970). Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography. *J. Theor. Biol.* **29**: 471–481.
Harauz, G. and van Heel, M. (1986). Exact filters for general three-dimensional reconstruction. *Optik* **73**:146–156.
Hoppe, W., Schramm, H. J., Sturm, M., Hunsmann, N. and GaBmann, J. (1976). Threedimensional electron microscopy of individual biological objects. I: Methods. *Z. Naturforsch.* **31a**:645–655.
Hsieh, C.-E., Marko, M., Frank, J. and Mannella, C. A. (2002). Electron tomographic analysis of frozen-hydrated tissue sections. *J. Struct. Biol.* **138**:63–73.
Kwok, Y. S., Reed, I. S. and Truong, T. K. (1977). A generalized $|\omega|$-filter for 3D reconstruction. *IEEE Trans. Nucl. Sci.* **NS24**:1990–2005.
Lanzavecchia, S. and Bellon P. L. (1998). Fast computation of 3D Radon transform via a direct Fourier method. *Bioinformatics* **14**:212–216.
Levi, A. and Stark, H. (1983). Signal restoration from phase by projections onto convex sets. *J. Opt. Soc. Am.* **73**:810–822.
McEwen, B. F., Radermacher, M., Rieder, C. L. and Frank, J. (1986). Tomographic three-dimensional reconstruction of cilia ultrastructure from thick sections. *Proc. Nat. Acad. Sci. USA* **83**:9040–9044.
Papoulis, A. (1968). *Systems and Transforms with Applications in Optics*. McGraw-Hill, New York; reprint, Robert E. Krieger, Florida, 1986.
Provencher, S. W. and Vogel, R. H. (1988). Three-dimensional reconstruction from electron micrographs of disordered specimes. I: Method. *Ultramicroscopy* **25**:209–222.
Radermacher, M. (1980). Dreidimensionale Rekonstruktion bei kegelförmiger Kippung im Elektronenmikroskop. PhD thesis, Technische Universität München, Germany.
Radermacher, M. (1988). Three-dimensional reconstruction of single particles from random and non-random tilt series. *J. Electron. Microsc. Tech.* **9**:359–394.
Radermacher, M. (1994). Three-dimensional reconstruction from random projections: orientational alignment via Radon transforms. *Ultramicroscopy* **53**:121–136.
Radermacher, M. (1997). Radon transform techniques for alignment and three-dimensional reconstruction from random projections. *Scanning Microsc.* **11**:171–177.
Radermacher, M. and Hoppe, W. (1978). 3-D reconstruction from conically tilted projections. In *Proc. 9th Int. Congr. Electron Microscopy* Vol. 1, pp. 218–219.
Radermacher, M. and Hoppe, W. (1980). Properties of 3-D reconstructions from projections by conical tilting compared to single axis tilting. In *Proceeding of the 7th European Congress on Electron Microscopy* Vol. 1, pp. 132–133.
Radermacher, M., Wagenknecht, T., Verschoor, A. and Frank, J. (1986). A new 3-D reconstruction scheme applied to the 50S ribosomal subunit of *E. coli. J. Microsc.* **141**: RP1–RP2.
Radermacher, M., Wagenknecht, T., Verschoor, A. and Frank, J. (1987). Threedimensional reconstruction from a single-exposure random conical tilt series applied to the 50S ribosomal of *Escherichia coli. J. Microsc*. **146**:113–136.

Radon, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Ber. Verh. König. Sächs. Ges. Wiss. Leipzig, Math. Phys. Kl.* **69**:262–277.

Ramachandran, G. N. and Lakshminarayanan, A. V. (1971). Three-dimensional reconstruction from radiographs and electron micrographs: Application of convolution instead of Fourier transforms. *Proc. Natl. Acad. Sci. USA* **68**:2236–2240.

Shannon, C. E. (1949). Communication in the presence of noise. *Proc. IRE* **37**:10–22.

Shepp, L. A. (1980). Computerized tomography and nuclear magnetic resonance zeugmatography. *J. Comput. Assist. Tomogr.* **4**:94–107.

Smith, P. R., Peter, T. M. and Bates, R. H. T. (1973). Image reconstruction from a finite number of projections. *J. Phys.* A **6**:361–382.

Stöffler, D., Feja, B., Fahrenkrog, B., Walz, J., Typke, D. and Aebi, U. (2003). Cryoelectron tomography provides novel insights into nuclear pore architecture: implications for nucleocytoplasmic transport. *J. Mol. Biol.* **328**:119–130.

Suzuki, S. (1983). A study on the resemblance between a computed tomographic image and the original object, and the relationship to the filterfunction used in image reconstruction. *Optik* **66**:61–71.

Vainshtein, B. K. and Orlov, S. S. (1972). Theory of the recovery of functions from their projections. *Sov. Phys. Crystallogr.* **17**:253–257.

Vogel, R. W. and Provencher, S. W. (1988). Three-dimensional reconstruction from electron micrographs of disordered specimes. II Implementation and results. *Ultramicroscopy* **25**:223–240.

Zampighi, G. A., Zampighi, L., Fain, N., Wright, E. M., Cantele, F. and Lanzavecchia, S. (2005). Conical tomography II: A method for the study of cellular organelles in thin sections. *J. Struct. Biol.* **151**:263–274.

Zwick, M. and Zeitler, E. (1973). Image reconstruction from projections. *Optik* **38**:550–565.

9

Reconstruction With Orthogonal Functions

Elmar Zeitler

1. INTRODUCTION

In 1917, Johann Radon posed the question of whether the integral over a function with two variables along an arbitrary line can uniquely define that function such that this functional transformation can be inverted. He also solved this problem as a purely mathematical one, although he mentioned some relationships to the physical potential theory in the plane. Forty-six years later, A. M. Cormack published a paper with a title very similar to that by Radon yet still not very informative to the general reader, namely 'Representation of a function by its line integrals'—but now comes the point: 'with some radiological applications'. Another point is that the paper appeared in a journal devoted to applied physics. Says Cormack, 'A method

E. Zeitler • Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4–6, D-14195 Berlin, Germany

is given of finding a real function in a finite region of a plane given by its line integrals along all lines intersecting the region. The solution found is applicable to three problems of interest for precise radiology and radiotherapy'. Today we know that the method is useful and applicable to the solution of many more problems, including that which won a Nobel prize in medicine, awarded to A. M. Cormack and G. N. Hounsfield in 1979. Radon's pioneering paper (1917) initiated an entire mathematical field of integral geometry. Yet it remained unknown to the physicists (also to Cormack, whose paper shared the very same fate for a long time). However, the problem of projection and reconstruction, the problem of tomography as we call it today, is so general and ubiquitous that scientists from all kinds of fields stumbled on it and looked for a solution—without, however, looking back or looking to other fields. Today there is a vast literature which cannot comprehensively be appreciated in this short contribution. It was Cormack (1963, 1964) who first made use of orthogonal functions for the solution of Radon's problem. Not only is their application elegant, but it also provides a good understanding about the inner relationships of a structure to its projections. The goal of this contribution is to demonstrate these relationships.

2. ORTHOGONAL POLYNOMIALS

As we do not presuppose any knowledge about orthogonal polynomials, we will compile a certain minimum which, however, is derived from knowledge certainly available to anybody engaged in advanced electron microscopy, diffraction theory and Fourier techniques. We are concerned with the properties of orthogonal polynomials, but abstain from general derivations. Rather, we demonstrate them for the special functions that are needed in tomography and point out that their properties are generally valid for all orthogonal polynomials. Without calculation, we all know that the integral over a periodic function vanishes when a number of plus and minus cycles proper for a complete cancelation fall within the range of integration

$$\int_0^{\pi} \cos n\phi \, d\phi = 0. \tag{1}$$

Only if we change n, the number of nodes (zeros) within the interval, to nil does the integration result in a finite value equal to π. However, actually no node means also that we violate the original agreement of periodic functions. One can retain the periodicity and yet avoid the cancellations by squaring the integrand; and, again without calculation and just by reasoning, one can arrive at

$$\int_0^{\pi} \cos^2 n\phi \, d\phi = \frac{\pi}{2}; \quad n \neq 0.$$

Multiplication of two such harmonic functions leads to

$$\cos n\phi \cos m\phi = \frac{1}{2}\cos(n+m)\phi + \frac{1}{2}\cos(n-m)\phi,$$

i.e. to an additive mixture of what we would call in acoustics a sum tone and a difference tone, whose integral amounts to zero except when the numbers of nodes m and n are the same. The product of two vectors can be zero although their lengths are not. This happens when their directions are orthogonal to each other. In analogy, the harmonic cosine functions are called orthogonal because of the properties we found.

Orthogonality:

$$\frac{2}{\pi}\int_0^\pi \cos n\phi \cos m\phi \, d\phi = \delta_{n,m}(1+\delta_{n,0}) \tag{2}$$

where

$$\delta_{n,m} = \begin{cases} 1 & n = m \\ 0 & \text{otherwise} \end{cases}$$

A typical feature of these orthogonal functions is the existence of a recurrence relationship which connects three members of the set.

Recurrence:

$$\cos n\phi = 2\cos\phi\cos(n-1)\phi - \cos(n-2)\phi.$$

This relationship, which follows from the simple rules of compound angles, enables one to set up the complete system

$$\begin{aligned}
&\cos 0\phi = 1 \\
&\cos 1\phi = \cos\phi \\
&\cos 2\phi = 2\cos\phi\cos\phi - 1 = 2\cos^2\phi - 1 \\
&\cos 3\phi = 2\cos\phi\cos 2\phi - \cos\phi = 4\cos^3\phi - 3\cos\phi \\
&\cos 4\phi = 2\cos\phi\cos 3\phi - \cos 2\phi = 8\cos^4\phi - 4\cos 2\phi - 3
\end{aligned} \tag{3}$$

Two remarkable facts are contained in these lines. First, solving the last line for $\cos^4\phi$, we obtain

$$\cos^4\phi = \frac{1}{8}\cos 4\phi + \frac{1}{2}\cos 2\phi + \frac{3}{8},$$

which amounts to the Fourier decomposition of $\cos^4\phi$ into its harmonics; in other words

$$\frac{1}{8} = \frac{2}{\pi}\int_0^\pi \cos^4 \cos 4\phi \, d\phi$$

is the corresponding Fourier coefficient usually found by means of orthogonality as shown in equation (2) above. The second fact is that the harmonic functions $\cos n\phi$ with n nodes can be expressed by polynomials in powers of $\cos\phi$ whose degree equals n. The right hand side of equation (3) shows polynomials, while the left hand side shows orthogonal harmonics; hence we are very close to orthogonal polynomials.

We introduce

$$\cos\phi = x \quad \text{or} \quad \phi = \cos^{-1} x \quad \text{then}$$
$$\cos 3\phi = \cos 3(\cos^{-1} x) = T_3(x).$$

This polynomial in x is called the Chebyshev polynomial of third *degree*, $T_3(x)$. With this change of variable from ϕ to x, the above results also lead to a new formulation of orthogonality.

Orthogonality of polynomials:

$$\frac{2}{\pi}\int_{-1}^{+1} T_n(x) T_m(x) \frac{dx}{\sqrt{1-x^2}} = \delta_{n,m}(1+\delta_{n,0}). \tag{4}$$

The integrand contains $\dfrac{1}{\sqrt{1-x^2}}$ as a so-called weight or weighting factor which, besides the range of integration, is typical for the particular set of polynomials.

Recurrence:

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x) \tag{5}$$

So far we have introduced only the $T_n(x)$, the first kind of Chebyshev's polynomials. However, after this lengthy preparation, we may apply more formal and quicker steps to arrive at Chebyshev's polynomials of a second kind. In the complex plane, $z_1 = e^{\imath\phi}$ represents a point on the unit circle, with the coordinates x and $y = \sqrt{1-x^2}$. Instead of dealing only with the real part we take the complex function

$$z_1^n = e^{\imath\phi} = (\cos\phi + \imath\sin\phi)^n = \left(x + \imath\sqrt{1-x^2}\right)^n$$
$$e^{\imath 0\phi} = 1 = T_0(x)$$
$$e^{\imath 1\phi} = x + \imath\sqrt{1-x^2} = T_1(x) + \imath\sqrt{1-x^2}U_0(x)$$
$$e^{\imath 2\phi} = x^2 - (1-x^2) + \imath\sqrt{1-x^2}\,2x = T_2(x) + \imath\sqrt{1-x^2}U_1(x)$$
$$e^{\imath 3\phi} = T_3(x) + \imath\sqrt{1-x^2}U_2(x)$$

In order to avoid the unappealing square root, the modern literature designates as the Chebyshev polynomials $U_n(x)$ of the second kind:

$$\Im\frac{z_1^{n+1}}{\sqrt{1-x^2}} = \Im\frac{e^{\imath(n+1)\phi}}{\sin\phi} = \frac{\sin(n+1)\phi}{\sin\phi} = U_n(x); \quad x = \cos\phi.$$

$$1 = \frac{\sin 1\phi}{\sin\phi} = U_0(x)$$

$$2x = \frac{\sin 2\phi}{\sin\phi} = U_1(x)$$

$$4x^2 - 1 = \frac{\sin 3\phi}{\sin\phi} = U_2(x).$$

From the imaginary part of

$$e^{\imath(n+1)\phi} = \left(e^{\imath\phi} + e^{-\imath\phi}\right)e^{\imath n\phi} - e^{\imath(n-1)\phi}$$

follows upon division by $\sin\phi$ the recurrence relationship

$$U_n(x) = 2xU_{n-1}(x) - U_{n-2}(x), \tag{6}$$

which is identical to that for the $T_n(x)$, the difference being that $T_0 = 1$, $T_1 = x$, whereas $U_0 = 1$; $U_1 = 2x$; through the changed weighting, the orthogonality expression also changes slightly to:

$$\frac{2}{\pi}\int_{-1}^{+1} U_n(x)U_m(x)\sqrt{1-x^2}\,dx = \delta_{n,m}. \tag{7}$$

Sometimes of help can be the so-called generating function $U(t, x)$, which contains, besides the variable of interest x, a dummy variable t that serves as a marker; when $U(t, x)$ is expanded as a power series in t, it has as coefficients the polynomials to be generated. It can readily be derived that

$$U(t, x) = \frac{1}{1-2xt+t^2} = \sum_0^\infty t^n U_n(x) \tag{8}$$

For the many connections of these polynomials to other classical orthogonal polynomials such as that of Gegenbauer or Jacobi, the literature must be consulted (e.g. Szego, 1975).

Here we shall try to involve only those types which occur in tomography, such as the Zernike polynomials.

We conclude this section with an example reminding us that functions can be approximated with the aid of orthogonal polynomials as long as the range of their independent variable is the same:

$$f(x) = \sum_0^n a_k T_k(x) = \frac{2}{\pi}\sum_0^n T_k(x)\int_{-1}^{+1} f(\xi)\frac{T_k(\xi)}{\sqrt{1-\xi^2}}\,d\xi. \tag{9}$$

The 'Fourier coefficients' a_k are readily calculated. The error of the approximation is always the minimum of the mean square deviation. In cases where a further improvement is required, more terms must be included. The advantage is that the coefficient from the previous approximations remain valid for the refined approximation; new ones are merely added. This is a very economical feature—coefficients calculated once keep their value. In the next section, we demonstrate the usefulness of these properties for solving the problem of interpolation.

3. INTERPOLATION AND QUADRATURE

Measurements of a physical property always render a set of discrete quantitative values for a selection of fixed parameters that are dictated by that experiment. On the other hand, a theoretical treatment of a measured sequence is often facilitated when an analytical expression that describes the data can be established, because then the required mathematical operations can be performed effectively. In other words, one needs values not only at discrete points but elsewhere as well. This problem is solved by interpolation. Therefore, this section is included to show why orthogonal polynomials are most suited for this purpose. With an interpolating polynomial at hand, the problem of determining the integral over an interpolated function is readily solved. As this problem of quadrature occurs in our reconstruction procedures as well, its solution is also included in this section. Although these matters go all the way back to Lagrange and Gauss, a brief review is still desirable. At every 'abscissa point' x_j, we have performed a measurement and found the outcome $M(x_j)$. We have obtained a table x_j, $M(x_j)$ or a curve with m points, $j = 1, 2, 3 \ldots m$. In accord with Lagrange, we seek an analytical expression

$$F(x) = \sum_{1}^{m} L(x, x_j) M(x_j) \tag{10}$$

$$F(x_i) = \sum_{1}^{m} L(x_i, x_j) M(x_j) = M(x_i) \tag{11}$$

Equation (10) resembles a discrete version of an integral representation of the function $F(x)$:

$$F(x) = \int L(x, y) M(y) dy; \tag{12}$$

instead of a continuous kernel we have m kernel-functions $L(x, x_j)$. (For convenience we suppress the indicator m; the notation should be either L_m instead of L, or $x_{m,j}$ instead of x_j.) The problem is solved if

$$L(x_i, x_j) = \delta_{i,j}. \tag{13}$$

The polynomial of mth degree,

$$p(x) = \Pi(x - x_j) \quad j = 1, 2, \ldots m \tag{14}$$

has m zeros at $x = x_j$; its first-order Taylor expansion about $x = x_i$ is

$$\begin{aligned} p(x - x_i) &= p(x_i) + (x - x_i)\, p'(x_i) \\ &= (x - x_i)\, p'(x_i) \end{aligned}$$

so that

$$\frac{p(x)}{(x - x_i)p'(x_i)} = L(x, x_i) \tag{15}$$

has the required property. Remember that the locations of the zeros x_j have not been committed. Instead of opting for the usually equidistant division, we select the locations such that the sampling of the outcome of our measurements, namely the sampling of the projections, becomes most profitable as input into the structure synthesis (reconstruction).

First, all the zeros $| x_j |$ fall into the range from zero to one; secondly, we like to sample the neighborhood of the rim of the unit circle more densely than that of the center in order to compensate for the decreasing projection length (see Fig. 1). As $x = \cos\phi$, the projection length = $\sin\phi$, hence we should avoid sampling points at $x_j = \pm 1$ or $\phi_j = 0, \pi$. All those requirements

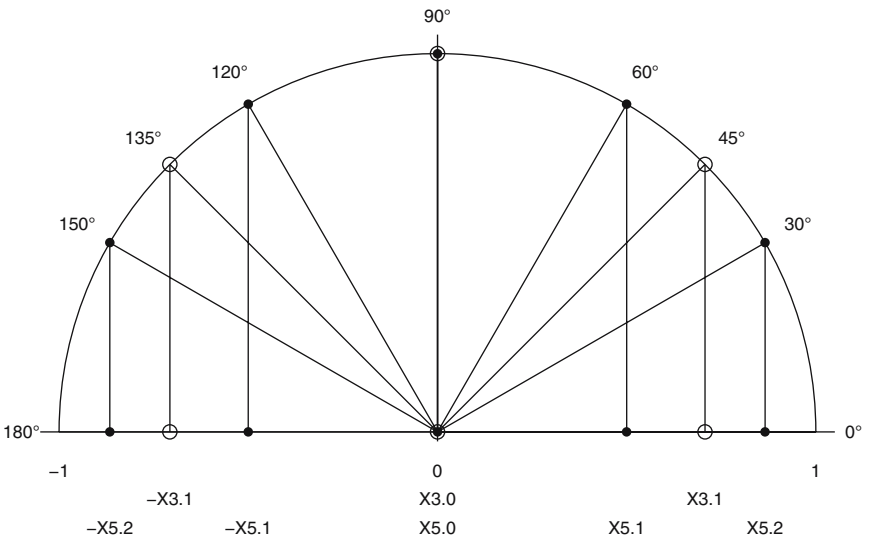


FIGURE 1. Two sets of zeros $x_{3,j}$ and $x_{5,j}$ of the Chebyshev polynomials $U_3(x)$ and $U_5(x)$. Note the interlacing of the sets and the higher density towards the limits plus or minus one.

are fulfilled when we sample in an equiangular division in ϕ rather than in an equidistant division in x, by means of the function $\sin(m+1)\phi/\sin\phi$, the denominator normalizing the projection length. This function has m zeros at $\phi_j = j\pi/(m+1)$, $j = 1, \ldots m$ with the corresponding zeros $x_j = \cos\phi_j$ (note that $\phi_0 = 0$ and $\phi_{m+1} = \pi$ are not zeros!). We recognize

$$p(x) = \frac{\sin(m+1)\phi}{\sin\phi} = U_m(\cos\phi) = U_m(x)$$

as the mth Chebyshev polynomial of the second kind, of which x_j is one of its m zeros (see Fig. 1). The Lagrange interpolation function is then

$$L(x, x_j) = \frac{U_m(x)}{(x - x_j)U'_m(x_j)}. \tag{16}$$

There are now two properties, one generally valid for orthogonal polynomials, the other especially for the derivatives of U_m(x), which permit simplification of $L(x, x_j)$.

1. The Christoffel–Darboux relationship

$$U_m(x)U_{m-1}(y) - U_{m-1}(x)U_m(y) = 2(x-y)\sum_0^{m-1} U_k(x)U_k(y)$$

becomes, upon identifying y with a zero x_j on account of

$$U_m(x_j) = 0, \quad U_{m-1}(x_j) = (-1)^{j+1},$$

identical to

$$\frac{U_m(x_j)}{x - x_j} = 2(-1)^{j-1}\sum_0^{m-1} U_k(x)U_k(x_j), \tag{17}$$

a formula that is reminiscent of the unity operator made up of eigenfunctions.

2. The expression for the derivative of the Chebyshev polynomial at the general coordinate x simplifies considerably at the location of a zero.

$$U'_m(x_j) = (-1)^{j+1}\frac{m+1}{1 - x_j^2}.$$

With these results, we obtain the Lagrange polynomials suited for our problem:

$$L(x, x_j) = \frac{2}{m+1}\sum_0^{m-1}(1 - x_j^2)U_k(x)U_k(x_j). \tag{18}$$

The representation of $F(x)$ found by interpolation, equation (2.1), can be seen by means of equation (18) as an expansion in m terms of orthogonal polynomials, namely

$$F(x) = \sum_{0}^{m-1} B_k U_k(x), \tag{19}$$

the Fourier coefficients being

$$B_k = \frac{2}{m+1} \sum_{1}^{m} (1 - x_j^2) U_k(x_j) F(x_j),$$

where $F(x_j) = M(x_j)$ has been utilized. As mentioned in Section 1, the usual way to obtain the Fourier coefficients makes use of the orthogonality, which here gives

$$B_k = \frac{2}{\pi} \int_{-1}^{1} \sqrt{1 - x^2}\, F(x) U_k(x) dx.$$

Equating the two last expressions is the essence of the well-known Gaussian quadrature: integrals of the type

$$I = \int_{-1}^{1} \sqrt{1 - x^2}\, G(x) dx$$

can be approximated by weighted sums

$$I = \sum_{1}^{m} w_j G(x_j),$$

whose error vanishes if $G(x)$ is a polynomial of power less than $2m$, since m sampling points x_j and m weights w_j are available (see, for example, Abramowitz and Stegun, 1965). Indeed, it can be shown that the error in equating both results is less than

$$R_m = \frac{2}{\pi} \frac{2^{-2m}}{2m!} \left(\frac{d}{d\xi} \right)^{2m} G(\xi), \quad |\xi| < 1.$$

Depending on the weight under the integral and on the range of the integration, different sets of orthogonal polynomials are invoked for the Gaussian quadrature. From the 'derivation' given, we can understand that for integrals of the type

$$I = \int_{-1}^{1} G(x) \frac{dx}{\sqrt{1 - x^2}} \quad \text{the approximation} \quad I = \sum_{0}^{m} w_j G(x_j) \quad \text{is valid,}$$

where the sampling points coincide with the zeros of the Chebyshev polynomials, $T_n(x)$ of the first kind:

$$\phi_j = \frac{2j-1}{m}\frac{\pi}{2}, \quad w_j = \frac{\pi}{2},$$

and the weights are conveniently constant; on the other hand, the remainder R_m is four times larger than the former one.

Incidentally, integrals over the area of the unit circle which are of the type, say, $\int_0^1 r^k G(r)\, r\, dr$ are approximated by sums over $G(r_j)$ at sampling points that are the zeros of certain Zernike polynomials, which are of course mutually orthogonal. In the general literature, they are known as *shifted Jacobi polynomials*, i.e. their argument is $2r^2 - 1$. Weights, zeros and remainder are rather complicated and hence are omitted here (see Abramowitz and Stegun, 1965).

Of equal importance to the radial integration over the unit circle is the azimuthal integration over 2π-periodic functions as, for example, over a 2D function at a fixed radius. The theory based on the orthogonal functions $f_k(\theta) = e^{\imath k\theta}$ gives a recipe for a Gaussian quadrature:

$$\int_0^{2\pi} f(\theta)d\theta \approx \frac{2\pi}{m}\sum_1^m f(\theta_j) + R_m(f). \tag{20}$$

The angular sectors are of equal size around the circumference

$$\theta_j = \frac{2\pi}{m}(j-1) + \alpha, \quad w_j = \frac{2\pi}{m}, \quad j = 1, 2, \ldots m,$$

with an arbitrary start angle α and constant weights.

The remainder $R_m(f_k)$ is zero as long as k is less than m.

4. A PRIORI *KNOWLEDGE AND ASSUMPTIONS ABOUT THE OBJECT*

The 3D object is composed of thin 2D slices. The mass density ρ of a slice varies from point to point (x,y) but does not vary along the third dimension. This requirement translates into a limitation of the slice's real thickness. The mass density is always positive and real; it differs from zero within a finite region.

1. We assume this region to be the unit disk

$$x^2 + y^2 = r^2 \le 1 \tag{21}$$

2. We assume that the mass density is periodic with respect to rotations about an axis normal to the disk (unit circle) through its center. Introducing cylindrical coordinates, r and ϕ, we have

$$\rho(r,\phi) = \rho(r,\phi + 2\pi) \tag{22}$$

The mathematical consequence of these realistic assumptions is that the mass density can be written as a Fourier series

$$\rho(r,\phi) = \sum_{-\infty}^{\infty} \rho_n(r) e^{in\phi}, \tag{23}$$

whose coefficients determine the radial dependence. Symmetry considerations, which are always quite powerful, restrict the mathematical form of the radial dependence.

3. We assume for the sake of argument that the highest occurring periodicity is of order m. We know from the previous section that

$$\begin{aligned} \Re e^{im\phi} &= \cos m\phi = T_m(\cos\phi) \\ &= a\cos^m \phi + b\cos^{m-2}\phi + \ldots \end{aligned}$$

Hence the highest exponent which $x = r\cos\phi$ as such can assume in the power expansion of the regular mass density function is also m. However, in the combination $r^2 = x^2 + y^2$, which is rotationally invariant, higher exponents may occur. It follows that the coefficient ρ_m must be of the form:

$$\rho_m(r) = r^m(\alpha_0 + \alpha_1 r^2 + \alpha_2 r^4 + \ldots \alpha_k r^{2k} + \ldots) \tag{24}$$

It is of interest to note that the Bessel function $J_m(r)$ of order m, which plays an important role in the theory of diffraction and of reconstruction, fulfills this requirement.

4. We assume that the total mass of the disk is unity (normalized); that means

$$\int_0^{2\pi}\int_0^1 \rho(r,\phi) r\, dr\, d\phi = 1 = 2\pi\int_0^1 \rho_0(r) r\, dr. \tag{25}$$

The total mass is contained in the 'dc' coefficient $\rho_0(r)$, whereas the 'ac' coefficients $\rho_j(r)$ represent 'decorative' yet massless features of the distribution.

5. Finally, we assume that the physical response of the electron microscope is proportional to the optical path length or to the mass thickness traversed by the electron, which we write as

$$\sigma[L] = \int_L \rho(x,y)ds \tag{26}$$

and call $\sigma[L]$, for the sake of brevity, the projection of the mass distribution ρ along the path L. For pertinent details and caveats, consult Chapter 3 of this volume. The problem at hand is to record many projections for various paths L and derive from the results the mass distribution inside the

disk. With the assumptions made, the problem can be formulated in a clear-cut way and be solved likewise. We prefer to present the attainment of this goal in an inductive manner rather than by means of rigorous mathematical deduction.

5. THE PROJECTION OPERATION

In this section, we specify the projection path L as the secant from rim to rim of the object disc parallel to its y-axis; the sum of the density values along this path we call the projection:

$$\sigma(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \rho(x, y)dy. \tag{27}$$

The various relationships of y to other variables permit interesting transformations of the projection operation and insights into the interrelationships between the original structure ρ and its projections σ, e.g.:

$$z = x + \imath y \Rightarrow dz = \imath dy \Rightarrow dy = -\imath dz$$
$$r^2 = x^2 + y^2 \Rightarrow r\,dr = y\,dy \Rightarrow dy = \frac{rdr}{\sqrt{r^2 - x^2}}$$

Since we need to perform these operations quite often, let us introduce symbols and operators for it. We define

$$PY(\ ) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} (\ )dy.$$

and try a few projections

$$\begin{aligned}
&PY(1) = 2\sqrt{1-x^2} \\
&PY(x) = 2\sqrt{1-x^2}\,x \qquad PY(y) = 0 \\
&PY(x^2) = 2\sqrt{1-x^2}\,x^2 \qquad PY(y^2) = 2\sqrt{1-x^2}\,\frac{1}{3}(1-x^2) \\
&PY(x^{2n}) = 2\sqrt{1-x^2}\,x^{2n} \qquad PY(y^{2n}) = 2\sqrt{1-x^2}\,\frac{1}{2n+1}(1-x^2)^n,
\end{aligned} \tag{28}$$

from which we can draw the following conclusions: (i) functions of x are merely multiplied by the path length through the disc; and (ii) odd functions of y yield zero projections; even functions project into even functions in x multiplied by the path length. We realize already here that the path length is the same square root that we encountered in the Chebyshev polynomials of the second kind. Normalizing the projections by this (trivial) length does remove the unappealing root. It also makes sense as it reflects

only the *a priori* knowledge of the object's spatial limits. Therefore, we include this normalization in the definition of the operators

$$PY(\ ) = \frac{1}{2\sqrt{1-x^2}} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} (\ )dy, \textit{i.e.},$$

PY stands for the averaged density $\sigma(x)$ encountered on a line parallel to y intersecting the x-axis at x.

We remember that density functions with arbitrary powers in r are not allowed; rather, their lowest term must be of the form $(re^{i\phi})^n$. The projection of this expression leads to a surprise:

$$PY\left(\left(re^{i\phi}\right)^n\right) = PY\left((x+\imath y)^n\right) = \frac{1}{n+1} \frac{\Im\left(x+\imath\sqrt{1-x^2}\right)^{n+1}}{\sqrt{1-x^2}} = \frac{1}{n+1} U_n(x).$$

This finding suggests the introduction of a second operator which performs the projection in the complex plane $z = (x, \imath y)$ on a density function $\rho = f(z)$ as a simple line integration between two conjugated points on the perimeter of the unit circle:

$$PZ(\ ) = \frac{-\imath}{z_1 - \overline{z}_1} \int_{\overline{z}_1}^{z_1} (\ )dz$$

$$z_1 = x + \imath\sqrt{1-x^2} = e^{i\phi}; \quad x = \cos\phi. \tag{29}$$

When $G(z)$ is the integral of $g(z)$ then its projection is

$$PZ(g(z)) = \frac{-\imath(G(z_1) - G(\overline{z}_1))}{z_1 - \overline{z}_1} = \frac{\Im G(z_1)}{\sqrt{1-x^2}}. \tag{30}$$

Also for the allowed combinations $z^n r^{2k}$ we can retain the convenience of this 'analytical' projection by writing for r^2,

$$r^2 = z\overline{z} = 2xz - z^2.$$

Let us compile a short table of permitted powers and their normalized projections PZ, as it can be done with ease and without computer:

$$\begin{aligned}
PZ(z^n) &= \frac{1}{n+1} U_n(x) \\
PZ(z^n r^2) &= \frac{2x}{n+2} U_{n+1}(x) - \frac{1}{n+3} U_{n+2}(x) \\
PZ(z^n r^4) &= \frac{4x^2}{n+3} U_{n+2}(x) - \frac{4x}{n+4} U_{n+3}(x) + \frac{1}{n+5} U_{n+4}(x).
\end{aligned} \tag{31}$$

Again, the special suitability of the Chebyshev polynomials is highlighted by the readily applicable recurrence formula, that enables us to remove the powers in x and thus project terms of $z^n r^{2k}$ into pure sums of Chebyshev polynomials; we find:

$$\binom{n+2k}{k} PZ(z^n r^{2k}) = \frac{1}{n+2k+1} \sum_0^k \binom{n+2k+1}{k-s} U_{n+2s}(x). \tag{32}$$

As a hint of the things to come, observe the possibility to create, by proper choice of the factors a_{nj}, linear combinations

$$e^{in\phi} R_{nk}(r) = z^n \sum_0^k a_{nj} r^{2j}$$

which project into a single Chebyshev polynomial $U_{n+2k}(x)$.

Making the proper choice is tantamount to finding the inverse of the matrix $A = [a_{nj}]$. This, however, is easy to do since A is already triangular. The polynomials $R_{nk}(r)$ will turn out to be orthogonal as well. They are the famous Zernike polynomials (see, for example, Born and Wolf, 1975). Before we conclude this section, one more word about a third projection operator PR which integrates over those values of the radius r that keep the product r cosϕ constant equal to x, a condition that makes ϕ a function of y. Therefore, we begin carefully with projecting a proper function of r and ϕ, namely:

$$\frac{1}{2\sqrt{1-x^2}} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} r^n e^{in\phi} g(r) dy.$$

After changing from the variable y to r and considering angular dependence and change of limits as well, we arrive at

$$PR(r^n e^{in\phi} g(r)) = \frac{1}{\sqrt{1-x^2}} \Re \int_{|x|}^1 r^n e^{in\phi} g(r) \frac{r\, dr}{\sqrt{r^2-x^2}} \tag{33}$$

$$= \frac{1}{\sqrt{1-x^2}} \int_{|x|}^1 r^n g(r) \frac{T_n(x/r)}{\sqrt{1-(x/r)^2}} dr. \tag{34}$$

It is certainly not accidental that the projection operation also invokes the Chebyshev polynomial of the first kind since the cosine function fixes the location whereas the sine function determines the length of the projection. The PR-operation reveals the relationship between the two kinds of polynomials, quite readily, $U_n(x)/(n+1) = PR(z^n)$, which follows from the formulae given above by setting $g(r) = 1$:

$$\frac{U_n(x)}{n+1} = \frac{1}{\sqrt{1-x^2}} \int_{|x|}^1 r^n \frac{T_n(x/r)}{\sqrt{1-(x/r)^2}} dr \tag{35}$$

The same effect of an r-projection could be achieved by a delta-function $\delta(x - r\cos\phi)$ and an integration over the entire domain r and ϕ, i.e. by a 2D integration over the unit disc. How does this projection kernel relate to the previous one? The delta-function is an even function periodic in ϕ and hence can be represented by a cosine Fourier series. Since expansions are the topic of this chapter, we ask for the expansion coefficients $G_n(x,r)$ of this series; they are directly found as

$$\begin{aligned} G_n(x,r) &= \frac{1}{\pi}\int_0^{\pi} \delta(x - r\cos\phi)\cos n\phi \, d\phi \\ &= \frac{1}{\pi}\int_{-r}^{r} \delta(x-u) T_n\left(\frac{u}{r}\right)\frac{du}{\sqrt{r^2-u^2}} \\ &= \frac{1}{\pi}\frac{T_n(x/r)}{\sqrt{r^2-x^2}}, \quad |x| \le r, \quad = 0, |x| > r \end{aligned} \tag{36}$$

In answering the above question, we see that both kernels render the same projection result. In the second case, the azimuthal integration will first filter out the n-fold symmetric structure while the consecutive r integration will add the seemingly missing r factor.

These findings describe succinctly specific properties of the projection operation.

6. FOURIER TRANSFORMATION AND PROJECTION

Confirmed by several contributions to this volume, Fourier transformations, decompositions and syntheses—continuous or discrete—play a major role in the processing of pictorial information. There is a simple computational aspect to it, i.e. very fast and economical implementations exist, and furthermore many other mathematical procedures such as convolution or correlation can be effected with great advantage by advanced Fourier algorithms. However, there is also a philosophical side. Questions as to the extension and to the contrast of an image detail are recast in Fourier language into questions of how many lines per unit length and of which intensity resemble best the image detail under investigation. The physical connection between these two reciprocal quantities, the one of a *length* and the other of a number (lines) *per length*, is established by the theory of diffraction or, experimentally, by the interaction of waves with matter. We like to capitalize on this knowledge. We would continue with our very simple 2D object, the homogeneous disk, this time of radius a. Its 2D Fourier transformation leads to the famous *Fraunhofer diffraction pattern* (Born and Wolf, 1975):

$$\oint\int_0^a e^{iwr\cos\phi} r \, dr \, d\phi = 2\pi a \frac{J_1(wa)}{w},$$

where $J_1(x)$ is the Bessel function of first order. If we normalize the integral by the area of the disk, πa^2 we obtain the well-known *Airy function*:

$$A_{00}(wa) = \frac{2J_1(wa)}{wa}.$$

The Airy function behaves very much like its 1D pendant,

$$\int_{-a}^{a} e^{\imath\eta y} dy = \frac{2\sin(\eta a)}{\eta} \tag{37}$$

which, after normalization (division by $2a$), turns into the sinc function:

$$\mathrm{sinc}(\eta a) = \frac{\sin(\eta a)}{\eta a}.$$

Both functions have very similar properties; they are even, they are unity for zero argument and they decay in an oscillatory fashion. Later on, when the angular modulation of the object also plays a role, we will encounter more general Airy functions, hence the special denotation.

$$\begin{aligned} \frac{2J_1(x)}{x} &= \frac{1}{2} A_{00}(x) \\ \frac{2J_{n+2k+1}(x)}{x} &= \frac{1}{\varepsilon_{nk}} A_{nk}(x) \\ \varepsilon_{nk} &= (1 + \delta_{n0}\delta_{k0}). \end{aligned} \tag{38}$$

When the spatial frequency η is chosen to be zero, the Fourier transform (37), becomes tantamount to an integration or projection, and the result is just the path length of the integral. A projection of the unit disk parallel to the y-axis at a distance x off-center gives the length $2\sqrt{1-x^2}$ and hence the Fourier transform

$$\frac{2\sin\left(\eta\sqrt{1-x^2}\right)}{\eta}.$$

We can readily remove the square root by expressing the 'impact parameter' x, the normal distance from the center, by $x = \cos\phi$. The sine function thus becomes periodic in ϕ and can be represented by a Fourier series. The result can be found in pertinent tables: the left hand side, representing an FM-sine wave, is decomposed on the right hand side into a multitude of AM-side bands whose amplitudes (Fourier coefficients) are Bessel functions of odd order. If we go back from ϕ to x and apply the functions already introduced, we obtain quite naturally

$$\frac{\sin\left(\eta\sqrt{1-x^2}\right)}{\eta\sqrt{1-x^2}} = 2\sum_0^\infty \frac{1}{\varepsilon_{0k}} A_{0k}(\eta) U_{2k}(x), \tag{39}$$

a generating function for either the Airy functions or the Chebyshev polynomials. Making use of their orthogonality, we arrive at their respective integral representations. In order to repeat in Cartesian coordinates the 2D Fourier transform of the disk, whose result in polar coordinates we already know, we must perform or look up the additionally required 1D x-transform

$$\frac{2}{\eta}\int_{-1}^{1} e^{i\xi x} \sin\left(\eta\sqrt{1-x^2}\right) dx = \pi A_{00}\left(\sqrt{\xi^2+\eta^2}\right);$$

indeed, we regain the previous Airy disc and recognize the rotational symmetry on account of the quadratic addition of the Cartesian frequency components

$$w^2 = \xi^2 + \eta^2.$$

Every central section through Fourier transform of the homogeneous disk looks the same. Hence, one projection suffices for reconstructing the disk. This situation will quickly change when more interesting structures arise. If we perform the same transformation on the single members of the right hand side in equation (39), we obtain again an Airy function, namely

$$\int_{-1}^{+1} e^{i\xi x} U_{2k}(x)\sqrt{1-x^2}\, dx = \pi(-1)^k (2k+1) A_{0k}(\xi)/\varepsilon_{0k}. \tag{40}$$

According to equation (39), the various transforms yield the important relationship:

$$A_{00}(w) = 2\sum_0^\infty (-1)^k \frac{(2k+1)}{\varepsilon_{0k}^2} A_{0k}(\xi) A_{0k}(\eta). \tag{41}$$

Mathematically, this relationship is a specialized form of the so-called *addition theorem* of Bessel functions (ξ and η must be in perpendicular directions, see, for example, Gradshteyn and Ryzhik, 1994), the physical message, however, is more interesting. The one side is, in polar coordinates, the 2D Fourier transform of the disk (i.e. its diffraction pattern), whereas the sum consists of products of 1D Fourier transforms, say, diffraction patterns of linear projections.

In order to clarify this important connection, we will elaborate on the 2D transformations of the allowed powers $z^n r^{2k}$, introduced earlier:

$$\frac{1}{\pi}\oint\int_0^\pi r^{n+2k} e^{in\phi} e^{i\omega\tau\cos\phi} r\, dr\, d\phi = \imath^n \int_0^1 r^{n+2k} J_n(wr) r\, dr. \tag{42}$$

With $J_n(\omega r)$ as kernel, the last expression appears as a special transformation in the radial domain of our 2D problem. This frequently occurring transformation is referred to as *general Hankel transformation HT* (general, because n is not just zero; see Bracewell, 1999), and for convenience we introduce the operator

$$HT(\ ) = \int_0^1 (\ ) J_n(wr) r\, dr.$$

Since $(x^{n+1}J_{n+1}(x))' = x^{n+1}J_n(x)$, the Hankel transforms of r^{n+2k} can be readily evaluated.

$$HT(r^n) = A_{n0}(w)\varepsilon_{n0}$$

$$HT(r^{n+2}) = \frac{n+1}{n+2} A_{n0}(\omega)/\varepsilon_{n0} - \frac{1}{n+2} A_{n1}(w)/\varepsilon_{n1} \tag{43}$$

$$HT(r^{n+4}) = \frac{n+1}{n+3} A_{n0}(w)/\varepsilon_{n0} - \frac{2}{n+4} A_{n1}(w)/\varepsilon_{n1} + \frac{2}{(n+3)(n+4)} A_{n2}(w)/\varepsilon_{n2} \tag{44}$$

Just as in the case of the operator PZ, we had to make use of the recurrence formula, this time for the Bessel functions, in order to reduce all the occurring powers $(1/\omega)^k$ to $1/\omega$, with the result:

$$\binom{n+2k}{k} HT(r^{n+2k}) = \sum_0^k (-1)^s \binom{n+2k+1}{k-s} \frac{n+2s+1}{n+2k+1} A_{ns}(w)/\varepsilon_{ns}. \tag{45}$$

Note the remarkable correspondence to the general $PZ(z^n r^{2k})$! Here we clearly recognize the above transforms as diffraction patterns of disk-like structures that have an n-fold azimuthal symmetry. The first index n refers to that symmetry, whereas the second index refers to the highest power of (the rotation invariant) r^2 which occurs in the series describing the radial dependence of the mass density in this particular n-fold symmetric structure. It is important to note that these generalized Airy functions have for a fixed n-fold symmetry, the welcome property of orthogonality, i.e.

$$\int_0^\infty A_{ns}(w) A_{nt}(w) w\, dw = \frac{\varepsilon_{ns}^2}{2(n+2s+1)} \delta_{ns},$$

so that in reciprocal space, spatial frequency functions with that symmetry can readily be expanded into sums over generalized Airy functions. Now we look for such linear combinations of allowed powers in r that Hankel-transform into a single 'clear-cut' Airy function. The above table shows that, for example,

$$HT((n+1)r^n - (n+2)r^{n+2}) = A_{n1}(w)$$

$$HT((n+1)(n+2)r^n - 2(n+2)(n+3)r^{n+2} + (n+3)(n+4)r^{n+4}) = 2A_{n2}(w) \tag{46}$$

We could go on finding more expressions by inverting the triangular equation system (matrix), but there an even simpler way of generating these interesting polynomials. Of course they do have a name; again they are the Zernike polynomials $R^n_{n+2k}(r)$. As their definition we require that their Hankel transform yields an Airy function with identical indices:

$$\int_0^1 R^n_{n+2k}(r) J_n(wr) r dr = (-1)^k A_{nk}(w) / \varepsilon_{nk}. \tag{47}$$

The straightforward yet slightly cumbersome stepwise inversion described above can be simplified by applying directly the inverse Hankel transform which, by the way, is self-reciprocal and thus yields

$$(-1)^k \varepsilon_{nk} R^n_{n+2k}(r) = \int_0^\infty A_{nk}(w) J_n(wr) w \, dw. \tag{48}$$

In physical terms we can describe these transformations by stating that a Zernike-mass density diffracts into an Airy pattern, and vice versa. When you look up these integrals in a table, you will find them written in the form of the more general Jacobi polynomials; hence more information is found by checking under those (e.g. in the well known handbook by Abramowitz and Stegun, 1965). We could have also synthesized the Zernike polynomials by *orthogonalizing* the powers of r^2 with a weighting function r^{2n+1} over an integration range from zero to one.

It seems indicated that we briefly summarize the results we have obtained thus far. We have dealt with three operations: (i) the projection PY along a line, which could be seen as a special case of (ii) the second transformation, i.e. the 1D Fourier transform (1D-FT) and finally (iii) the 2D-FT. In Cartesian coordinates, the 2D-FT is simply obtained by a duplication of the 1D-FT. In polar coordinates, however, the azimuthal transform leaves as the second 1D transform over the radius, a special transform known as Hankel transform (HT). Throughout our treatment of the reconstruction problem, all three transforms are intimately connected, as demonstrated by our finding of special functions and polynomials that transform into each other. As a special bonus, they are orthogonal and remain so after transformation.

7. ZERNIKE POLYNOMIALS AND THE SELECTION RULE

Up to now we have encountered the Zernike polynomials always in connection with some other orthogonal function. In this section, we introduce them more formally in their own right. The Zernike polynomials are chosen such that the functions

$$\psi_{ns}(r, \phi) = e^{in\phi} R^{|n|}_{|n|+2s}(r) \tag{49}$$

form a complete orthogonal system over the unit disk, i.e.

$$\oint\int_0^1 \psi_{ns}(r,\phi)\overline{\psi}_{mt}(r,\phi)r\,dr\,d\phi = \pi\frac{\delta_{nm}\delta_{st}}{\lambda_{ns}};$$
$$\lambda_{ns} = |n| + 2s + 1. \tag{50}$$

The polynomials $R^n_{n+2s}(r)$ are then orthogonal over the interval (0, 1), with a weighting function r, and normalized such that

$$\int_0^1 R^n_{n+2s}(r)R^n_{n+2t}(r)r\,dr = \frac{\delta_{st}}{2\lambda_{ns}}; \quad R^n_{n+2s}(1) = 1. \tag{51}$$

Most important for our purpose is the generating function (see, for example, Born and Wolf, 1975):

$$g_n(t,r) = \sum_{s=0}^{\infty} t^s R^n_{n+2s}(r)$$
$$= \frac{1}{R(t,r)}\left[\frac{1+t-R(t,r)}{2tr}\right]^n \tag{52}$$

whereby $R(t,r) = \sqrt{1-2t(2r^2-1)+t^2}$

If we identify t with $e^{\iota 2\theta}$, we obtain very simple expressions, namely

$$R(e^{\iota 2\theta}, r) = 2e^{\iota\theta}\cos\theta\sqrt{\cos^2\theta - r^2}$$

and, with $x = \cos\theta$,

$$g_n(x,r) = e^{-\iota(n+1)\theta}\left(\frac{x}{r} - \sqrt{\frac{x^2}{r} - 1}\right)^n \frac{1}{2\sqrt{x^2-r^2}} \tag{53}$$

Observing that the imaginary part

$$\Im\left(\frac{\left(x/r - \sqrt{(x/r)^2 - 1}\right)^n}{\sqrt{x^2-r^2}}\right) = \frac{T_n(x/r)}{\sqrt{r^2-x^2}}$$

is just the kernel for the projection operator PR, the generating function assumes a very simple meaning:

$$g_n(x,r)e^{\iota(n+1)\theta} = \sum_0^{\infty} R^n_{n+2s}(r)e^{\iota(n+2s+1)\theta}$$

whose imaginary parts can be rearranged into

$$\frac{2T_n(x/r)}{\sqrt{r^2-x^2}} = \sqrt{1-x^2}\sum_0^{\infty} R^n_{n+2s}(r)U_{n+2s}(x). \tag{54}$$

This formula encompasses all the relationships between the set of orthogonal functions in real space involved in projection and reconstruction. A Fourier transform of the formula translates the relationships into reciprocal space. Multiplying both sides of the formula by a Zernike polynomial $R^n_{n+2s}(r)$ and integrating over r leads on the left hand side to a projection, whereas on the right hand side one Chebyshev polynomial $U_{n+2k}(x)$ is singled out by orthogonality with the same indices as the projected Zernike polynomial. Playing this trick again, yet multiplying both sides now by a Chebyshev polynomial $U_{n+2k}(x)$, leads to a reconstruction of a Zernike polynomial. In other words, the projection operator has reconstructive properties as well; however, this time in the x or angular domain. Indeed, this is what tomography is all about: to obtain the structure from many projections of a tilted object. Therefore, it pays to rewrite this reconstruction operation in terms of angles

$$\frac{x}{r} = \cos\alpha, \quad T_m\left(\frac{x}{r}\right) = \cos m\alpha$$

$$R^m_{m+2k}(r) = \frac{2}{\pi}\int_0^{\pi/2} U_{m+2k}(r\cos\alpha)\cos m\alpha \, d\alpha \tag{55}$$

Forgetting the weight $\cos m\alpha$ for a moment, this integral expressing the structure R^m_{m+2k} is the summation of projections taken under various tilts. This operation is often referred to as *back-projection*. However, the analysis shows that back-projection can be correct only if the cosine weighting factor is included. It is remarkable that due to the symmetries, an angular range from zero to $\pi/2$ suffices. This goes beyond the fact that the projection $\sigma(x, 0)$ and the projection $\sigma(-x, \pi)$ of the flipped over structure are identical, and is a feature of the function system employed.

We conclude this section with some remarks about the selection rule. The Zernike polynomials have two distinct indices, n and s, while the Chebyshev polynomials and the Airy functions have only one, namely m, which is the sum $m = n + 2s$. For a given m, this *Diophantic equation* has many solutions (n, s), the parity of m and n being, however, always the same. How can we understand this ambiguity? The 'quantum number' n was already introduced as the number of zero lines running through the n-fold symmetric structure which we call spokes, for convenience. The quantum number $2s$ is the number of zeros (nodes) which occur in the radial extension of one particular Zernike polynomial. These are s concentric grooves running around the center of the 'eigen structure' $\psi_{n,s}(r, \phi)$. The spokes and the grooves together add up to $m = n + 2s$ zeros. In the projection, only the total number of zeros counts and not their type, azimuthal or radial, as in the eigen structure. This phenomenon, which is typical for a reduction of dimensions as it occurs in the process of projection, for example, is well known in other fields—in crystallography and signal processing the term *aliasing* has been coined for that phenomenon.

8. THE NUMERICAL SOLUTION OF THE RADON PROBLEM

In Fig. 2, we present the coordinate systems of the experimental set-up. The recording 'plane' is the x-axis. The projection is performed parallel to the y-axis, the optical axis of the microscope. The object with its own coordinate system (ξ, η) or (r,ϕ) centered on the tilt axis of the microscope's stage is tilted by an angle θ with respect to the x-axis. The measured results are a set of projections $\sigma(x)$ obtained at discrete tilts θ.

The problem is to find from these results the density distribution $\rho(r, \phi)$ within the object (we consider the angle still as a continuous variable). As pointed out already, it is sensible to normalize the projections by the length $2\sqrt{1-x^2}$ which contributes within the object to this projection. The Radon integral expresses the relationship between the measured results and the mass density to be deduced from them:

$$\tau(x,\theta) = \frac{\sigma(x,\theta)}{2\sqrt{1-x^2}} = \oint \int_0^1 G(x,\theta;r,\phi)\rho(r,\phi) r\, dr\, d\phi. \tag{56}$$

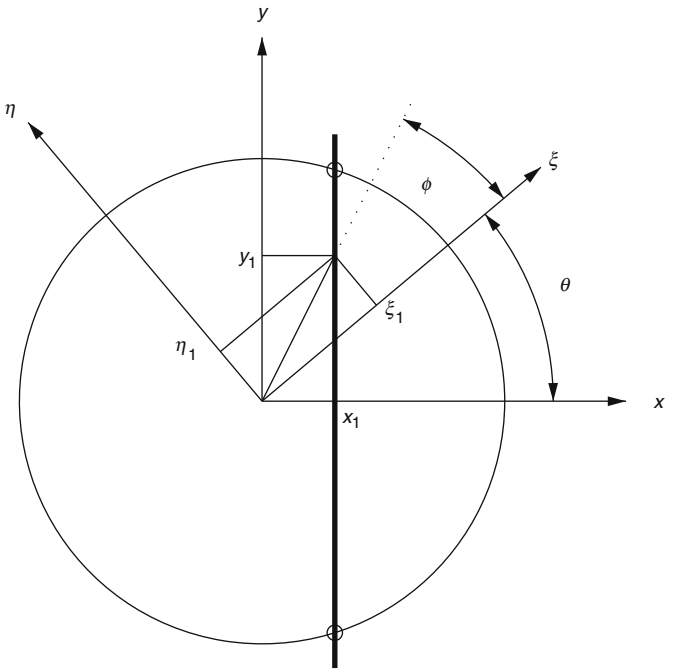


FIGURE 2. Coordinate systems of the projection experiment. System (ξ, η) fixed within the tilted object and θ is the tilt angle. System (x,y) fixed in the microscope (y is the optical axis).

The Radon kernel

$$G(x,\theta;r,\phi)=\frac{\delta(x-r\cos(\theta+\phi))}{2\sqrt{1-x^2}} \tag{57}$$

is a delta-function, whose Fourier components we could already identify as the kernel of the projection operator for the n-fold symmetric mode. The general solution must include all modes and all tilt angles. Therefore, we introduce, in analogy to the $\psi_{n,s}$ of the previous section, the two-dimensional projection polynomials $\pi_{n,s}$,

$$\pi_{n,s}(x,\theta)=U_{|n|+2s}(x)e^{-in\theta} \tag{58}$$

which also are orthogonal over the domain –1, 1 and $-\pi$, π:

$$\oint\int_{-1}^{1}\pi_{n,s}(x,\theta)\pi_{m,t}(x,\theta)dxd\theta=\pi^2\delta_{nm}\delta_{st}. \tag{59}$$

The Radon kernel then becomes,

$$G(x,\theta;r,\phi)=\frac{1}{\pi}\sum_{n-\infty}^{\infty}\sum_{s=0}^{\infty}\pi_{ns}(x,\theta)\overline{\psi}_{n,s}(r,\phi), \tag{60}$$

a product kernel whose factors each belong to one of the two orthogonal spaces spanned by the functions π_{ns} and ψ_{ns}.

The density function, expanded in the appropriate space, is represented by the series

$$\rho(r,\phi)=\sum_{n=-\infty}^{\infty}\sum_{s=0}^{\infty}\rho_{ns}\psi_{ns}(r,\phi), \tag{61}$$

whose Fourier coefficients are

$$\rho_{ns}=\frac{\lambda_{ns}}{\pi}\oint\int_{0}^{1}\rho(r,\phi)\overline{\psi}_{ns}(r,\phi)r\,dr\,d\phi. \tag{62}$$

Similarly, in the dual space, the normalized projected mass density will be represented by

$$\tau(x,\theta)=\sum_{n=-\infty}^{\infty}\sum_{s=0}^{\infty}\sigma_{ns}\pi_{ns}(x,\theta), \tag{63}$$

whose coefficients are found as

$$\sigma_{ns}=\frac{1}{\pi^2}\oint\int_{-1}^{1}\tau(x,\theta)\pi_{ns}(x,\theta)\sqrt{1-x^2}\,dx\,d\theta. \tag{64}$$

Inserting the expansion of $\rho(r, \phi)$ into the Radon integral, with its kernel also expanded, and performing the integrals over the orthogonal system's domain, we obtain the very simple relationship between the expansion coefficients of the density function and the projected density function. They are simply proportional to each other:

$$\sigma_{ns} = \frac{\rho_{ns}}{\lambda_{ns}}. \tag{65}$$

Therefore, the solution of the integral equation, i.e. the inversion of the experimental data into the sought-for information, is achieved by solving this relationship for ρ_{ns}:

$$\rho_{ns} = \lambda_{ns}\ \sigma_{ns}. \tag{66}$$

Herewith the problem is reduced to one of preparing the σ_{ns} values from the experimental raw material. Fortunately, the necessary integrals are all of the type ideally suited for Gaussian quadratures. If the experimentalist interpolates his discrete measurements onto abscissa points as demanded by the Gaussian quadrature, a reconstruction can be performed with ease, with high accuracy and, more importantly, with a reliable estimate of the errors.

The theory relates the data $\tau(x, \theta)$ in the double integral to

$$\sigma_{ns} = \frac{1}{\pi^2}\int_0^{2\pi} e^{in\theta}\, d\theta \int_{-1}^{+1} \tau(x,\theta) U_{n+2s}(x)\sqrt{1-x^2}\, dx. \tag{67}$$

In Gaussian quadrature, this double integral becomes a sum over the integrand at special 'spokes' θ_j and at special coordinates x_k, namely the zeros of the Chebyshev polynomial, multiplied by the appropriate weights v_j and w_k. Assume the option of $2l$ spokes and m zeros. Then, with the following discrete quantities

$$j = 1, 2, \ldots, 2l; \quad \theta_j = \frac{\pi}{l}; \quad v_j = \frac{\pi}{l};$$

$$k = 1, 2, \ldots, m; \quad \alpha_k = \frac{\pi}{m+1}k; \quad x_k = \cos\alpha_k; \quad w_k = \frac{\pi}{m+1}(1-x_k^2),$$

we obtain the formula

$$\sigma_{ns} = \frac{\pi}{l}\frac{\pi}{m+1}\sum_{j=1}^{2l} e^{in\theta_j}\sum_{k=1}^{m}\tau(x_k,\theta_j)U_{n+2s}(x_k)(1-x_k^2) + R. \tag{68}$$

Two simplifications are possible. One is merely a cancelation of the weights w_k in τ by those occurring in U_{n+2s}, which renders

$$\tau(x_k,\theta_j)U_{n+2k}(x_k)(-x^2) = \sin(\lambda_{ns}, \alpha_k)$$

The second simplification brings an experimental advantage. It relates to the fact that

$$\tau(x_k, \theta_j + \pi) = \tau(-x_k, \theta_j)$$
$$e^{in(\theta_j+\pi)} = (-1)^n e^{in\theta_j}$$

so that, for the same tilt, measurements are paired in a consistent fashion, thus increasing the accuracy of the procedure (this is why an even number of tilt angles should enter the analysis).

$$\sigma_{ns} = \frac{\pi}{l(m+1)} \sum_{j=1}^{l} e^{in\theta_j} \sum_{k=1}^{m} \left[\sigma(x_k, \theta_j) + (-1)^n \sigma(-x_k, \theta_j)\right] \sin(\lambda_{ns}, \alpha_k) + R. \quad (69)$$

Having obtained these values, it is a simple task to obtain the mass density, because we have

$$\rho(r, \phi) = \sum_{-m}^{m} \sum_{s=0}^{k} \lambda_{ns} \sigma_{ns} R^{|n|}_{|n|+2s}(r) e^{in\phi}. \quad (70)$$

The problem of finding a judicious truncation was highlighted and solved by Smith (1978) by resorting to the selection rules (see the following section.)

I would like to end this chapter with a quote from Strichartz (1982) in a paper on Radon inversion: 'What I like most about mathematics is the joy of discovery, when understanding overcomes confusion. What I like least about mathematics is the way clear and simple ideas tend to become muddy and murky and mystifying when committed to the printed page'.

9. REFERENCES TO THE RADON PROBLEM

This section is primarily intended as a guide through the electron microscopy literature, which was already rather complete by the mid-1970s. However, first I would like to briefly outline the general development.

The mathematical research has continued to develop along the lines typical for this field, i.e. towards generalization. Whereas the method presented here operates in a 2D space, mathematicians study the general *n*-dimensional problem. Deans (1979) describes a Radon inversion formula which holds in spaces of even or odd dimensions. As transformation pairs, he finds Gegenbauer polynomials, of which Chebyshev polynomials are a special type. Mathematically, one can construct various function pairs which span dual orthogonal spaces (see, for example, Lerche and Zeitler, 1976). As we have seen, the orthogonal functions are typified by the weight functions. Therefore, weight functions most suited for the particular Radon problem must be sought. In our case, we used a sharp cut-off in the weight-

ing function without investigating its consequences. One knows, however, from optics and signal theory that such measures will lead to ringing—to *Gibbs phenomena*—which in optics are avoided by masks with soft transitions (*apodization*). Mathematical investigations in this direction would be welcomed by the experimentalists. A fairly modern account of the mathematical status of the Radon problem is given by Helgason (1980) in the book *The Radon Transformation*.

Among the practical papers, we can only mention the large array from electrical engineers and computer specialists who are concerned with improving medical tomography. In this field, many restrictions typical for electron microscopic tomography are not operative. Furthermore, the projection schemes used in medicine differ from those in electron microscopy, the latter relying on a strictly parallel projection. In astronomy, image reconstruction had begun earlier than in electron microscopy. There the line integration (projection) is taken perpendicular rather than parallel to the propagation of the radiation, a fact that confirms the breadth of applicability of the Radon problem. An overview of the implementation and application of image reconstruction methods outside the field of electron microscopy is compiled and edited by Herman (1979).

We continue now with tomography for electron microscopy. This field was opened very suddenly in a paper published in *Nature* by DeRosier and Klug in 1968. These authors approached the problem as crystallographers who are accustomed to reconstructing regular structures from diffraction patterns and Fourier-transforming the information desired from reciprocal space into real space. Since the electron microscope forms an image of the object, this step should be unnecessary. Where is the mistake? The crux is that the electron microscopic image is a projection, and its Fourier transform is just a central section through the reciprocal space of the object. The idea of DeRosier and Klug was then to fill the entire reciprocal space by Fourier transforms of many tilted exposures and then perform the reconstruction just as crystallographers would do. This approach is called the *Fourier approach*.

Very early in the game, a group around Bates (Smith *et al.*, 1973) in New Zealand engaged in image reconstruction from finite numbers of projections. They were concerned with the fact that interpolation between data is required when the Fourier approach is to be applied. So they were looking for an appropriate and reliable interpolation scheme which also permitted economical usage of the computer. As a result of their research, they found polynomials which are identical to those of Cormack, identical to the orthogonal Chebyshev and Zernike polynomials, including the selection rule with which this chapter deals. They showed how one can use this rule as a consistency condition to estimate *a posteriori* the quality of the input data and the validity of the information derived from them, including statements about the resolution attained. The most recent paper along the same line of reliability is by Howard (1988);

it neither mentions the papers by Bates *et al*. nor does it surpass their results.

A number of years ago, a computer program based on these ideas was implemented and run by P. R. Smith at the Biozentrum in Basel, Switzerland, and it remains in use there and in many other laboratories. A comprehensive description of this program system for processing micrographs of biological structures, including a comparison with other systems, was prepared by Smith (1978).

An important trilogy on image reconstruction from projections by Bates *et al*. (Lewitt and Bates, 1978a,b; Lewitt *et al*., 1978) addresses the question of error sensitivity towards truncation and the influence of sampling. The basic difference between the Fourier approach and the orthogonal function method can be attributed to the difference of the expansion of the 2D Fourier kernel in cylindrical coordinates. The Fourier approach chooses Bessel functions and harmonic functions for its decomposition, of which only the latter are orthogonal, whereas the Cormack system of orthogonal polynomials is readily introduced by expanding the kernel right from the beginning into orthogonal functions, namely into Airy functions and Chebyshev polynomials (see, for example, Zeitler, 1974). Also the convolution method of Ramachandran and Lakshiminarayanan (1971) becomes unnecessary in this system since the Fourier transform of a projection leads to the same Airy function as does the 2D Fourier transform of the structure.

The most obvious extension of the expansion method into three dimensions has been implemented by Provencher and Vogel (1988). They describe the 3D structure of single particles by the orthogonal system of spherical harmonics, which are generally known from quantum mechanics through their application in calculating the electronic states of atoms. Instead of tilting a single object, they start out with an array of statistically oriented particles and study in an iterative fashion how the projections obtained fit the proposed structure. They obtained very impressive results for a globular virus (Semliki Forest virus) suspended in vitrified water. It must be pointed out, however, that the functions of the projections are not orthogonal.

Discrete Radon transform methods have been introduced into electron microscopy by Radermacher (Radermacher, 1994, 1997, Radermacher *et al*., 2001), particularly in the context of single-particle reconstruction of macromolecular complexes. Another group with active interest in the development of Radon transform-based image processing procedures has been that of Lanzavecchia in Milan (Lanzavecchia and Bellon, 1998; Lanzavecchia *et al*., 1999, 2002). Radon-based projection alignment and 3D reconstruction were implemented as a suite of programs in the SPIDER program system (Frank *et al*., 1996; http://www.wadsworth.org/spider_doc), and there are examples of structures solved entirely by Radon techniques (Ruiz *et al*., 2003).

Since the first edition of this chapter, new books on orthogonal polynomials have been published: a survey on *Application and Computation* by Gautschi *et al*. (1999) and the *Classical Polynomial of Discrete Variables* by Nikiforov et al. (1992).

More exciting, however, is the import and the standing tomography has gained in the intervening years. From a general public point of view, tomography is considered part of the fast growing and greatly supported field of biomedical technologies and biomedical engineering. The universities are aware of this trend as a challenge and an opportunity as well; they offer curricula, graduate programs and faculty positions to create foundations and professionals in both sciences, the biomedical and the engineering physical sciences. Thus, it is entirely appropriate for the department of electrical engineering, of computer science or of applied mathematics to propose a new program in computer tomography. For instance, G. T. Herman, a pioneer in reconstruction methods and a Professor at the City University of New York, has offered the following graduate seminars:

Reconstruction from Projections, 2002
Discrete Tomography, 2003
Algorithms for Image Reconstruction, 2004

This development will also help in avoiding the infertile duplications of tomographic methods in the various fields such as astronomy, geology, medicine and biology. The awareness and the common teaching will be beneficial and economical. To see an example for the present compartmentalization, look in *IEE Transactions on Information Theory*, 2002 for a paper 'on the recovery of a function on a circular domain' by Pawlak and Liao (2002) which, found accidentally, covers a similar ground as the present chapter.

Young scientists looking for an important problem should turn to the still lacking quantitative comparison of the calculated data with the original physical data.

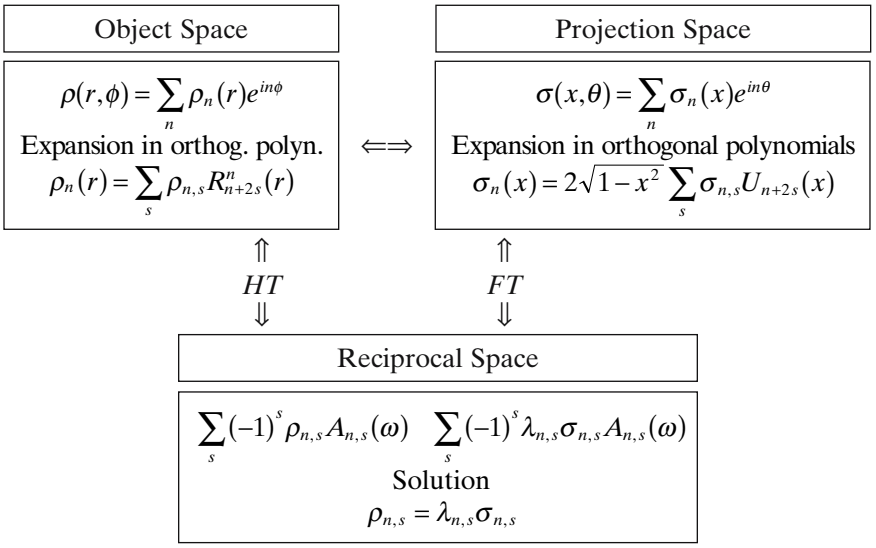
More's law, still effective, has advanced the power and the speed of computers to incredible levels. Large populations of single particles can by analyzed with statistical significance and refinement loops can be run until successful convergence is achieved. One more item that attests to the vitality of our field is Eric Weisstein's list (www.treasure-troves.com/books Radon Transforms.html) of books about Radon transforms on the worldwide web!

Let's see what's next.

10. SUMMARY

The premise of this chapter is that the experimental results are given as numerical data. The mathematical solution of reconstructing a 2D object

from its 1D projections is based on the orthogonal function expansion method such that the experimental data can be utilized with ease and least error.



The necessary steps towards this goal are summarized in the above scheme. Arrows pointing from structure functions ρ to projection functions σ indicate the *projection operation*, whereas arrows of opposite sense denote the inverse operation, i.e. *reconstruction* and hence the solution of the problem. Hankel transforms applied to structure functions and Fourier transforms applied to projection functions lead into the reciprocal or frequency space. Here the most suitable representation of the various Fourier components is found as an expansion into orthogonal Airy functions $A_{ns}(\omega)$. From the object space of structures, orthogonal Zernike polynomials $R^n_{n+2s}(r)$ transform into Airy functions, and from the space of projections orthogonal Chebyshev polynomials $U_{n+2s}(x)$ transform into the same Airy functions. When the Fourier components ρ_n and σ_n are expanded in terms of the pertinent orthogonal sets, the expansion coefficients ρ_{ns} and σ_{ns} are proportional to each other. Hence if the σ_{ns} are determined from the experimental data, the reconstruction is effected by merely inserting the proper ρ_{ns} into the structure expansion.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Dover.
Born, M. and Wolf E. (1975). *Principles of Optics*. Pergamon, Oxford, UK.
Bracewell, R. (1999). *The Hankel Transform*, 3rd edn. McGraw-Hill, New York.

Cormack, A. M. (1963). Representation of a function by its line integrals with some radiological applications. *J. Appl. Phys.* **34**:2722–2727.

Cormack, A. M. (1964). Representation of a function by its line integrals with some radiological applications. II. *J. Appl. Phys.* **35**:2908–2912.

Deans, S. R. (1979). Gegenbauer transforms via Radon transforms. *Siam J. Math. Am*. **10**:577–585.

DeRosier, D. J. and Klug, A. (1968). Reconstruction of three-dimensional structures from electron micrographs. *Nature* **217**:130–134.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M. and Leith, A. (1996). Spider and web: processing and visualization of images in 3d electron microscopy and related fields. *J. Struct. Biol.* **116**:190–199.

Gautschi, W., Golub, G. N. and Opfer, G. (1999). *Application and Computation of Orthogonal Polynomials*. Birkhaueser, Basel, Switzerland.

Gradshteyn, J. S. and Ryzhik, L. M. (1994). *Tables of Integrals, Series and Products*, 5th edn. Academic Press.

Helgason, S. (1980). *The Radon Transform*. Birkhäuser, Boston.

Herman, G. T. (1979). *Image Reconstruction From Projections*. Springer, Berlin.

Howard, J. (1988). Tomography and reliable information. *JOSA* **5**:999–1014.

Lanzavecchia, S. and Bellon, P. L. (1998). Fast computation of 3D radon transform via a direct fourier method. *Bioinformatics* **14**:212–216.

Lanzavecchia, S., Bellon, P. L. and Radermacher, M. (1999). Fast and accurate three-dimensional reconstruction from projections with random orientations via radon transforms. *J. Struct. Biol*. **128**:152–164.

Lanzavecchia, S., Cantele, F., Radermacher, M. and Bellon P. L. (2002). Symmetry embedding in the reconstruction of macromolecular assemblies via discrete radon transform. *J. Struct. Biol*. **17**:259–272.

Lerche, I. and Zeitler, E. (1976). Projections, reconstructions and orthogonal functions. *J. Math. Anal. Apppl*. **56**:634–649.

Lewitt, R. M. and Bates, R. H. T. (1978a). Image reconstruction from projections: I. general theoretical considerations. *Optik* **50**:19–33.

Lewitt, R. M. and Bates, R. H. T. (1978b). Image reconstruction from projections: III. projection completion methods (theory). *Optik* **50**:189–204.

Lewitt, R. M., Bates, R. H. T. and Peters, T. M. (1978). Image reconstruction from projections: II. modified back-projection methods. *Optik* **50**:85–109.

Nikiforov, A. F., Uvarov, V. B. and Suslov, S. S. (1992). *Classical Orthogonal Polynomials of a Discrete Variable*. Springer, New York,

Pawlak, M. and Liao, S. X. (2002). On the recovery of a function on a circular domain. *IEEE Trans. Inform. Theory* **48**:2736–2753.

Radermacher, M. (1994). Three-dimensional reconstruction from random projections: orientational alignment via radon transforms. *Ultramicroscopy* **53**:121–136.

Radermacher, M. (1997). Radon transform technique for alignment and 3d reconstruction from random projections. *Scanning Microsc. Int. Suppl*.:169–176.

Radermacher, M., Ruiz, T., Wieczorek, H. and Gruber, G. (2001). The structure of the v(1)-atpase determined by three-dimensional electron microscopy of single particles. *J. Struct Biol.* **135**:26–37.

Radon, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Ber. Verh. Sächs. Akad*. **69**:262–277.

Ramachandran, G. N. and Lakshiminarayanan, A. V. (1971). Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. *Proc. Natl Acad. Sci. USA* **68**:2236–2240.

Ruiz, T., Mechin, I., Br, J., Rypniewski, W., Kopperschlger, G. and Radermacher, M. (2003). The 10.8-structure of *Saccharomyces cerevisiae* phosphofructokinase determined by cryoelectron microscopy: localization of the putative fructose 6-phosphate binding sites. *J. Struct. Biol.* **143**:124–134.

Smith, P. R. (1978). An integrated set of computer programs for processing electron micrographs of biological structures. *Ultramicroscopy* **3**:153–160.

Smith, P. R., Peters, T. M. and Bates, R. H. T. (1973). Image reconstruction from finite numbers of projections. *J. Phys. A* **6**:319–381.

Strichartz, R. S. (1982). Radon inversion—variation on a theme. *Am. Math. Monthly* **89**:377–384.

Szegö, G. (1975). *Orthogonal Polynomials*. American Mathemeatics Society, Providence RI.

Vogel, R. H. and Provencher, S. W. (1988.) Three-dimensional reconstruction from electron micrographs of disordered specimens, II. Implementation and results. *Ultramicroscopy* **25**:223–240.

Zeitler, E. (1974). The recostruction of objects from their projections. *Optik* **39**:396–415.

10

Resolution in Electron Tomography

Pawel A. Penczek and Joachim Frank

1. INTRODUCTION

Traditionally, in computed tomography practiced in radiology, the resolution of the reconstruction is expressed in terms of the number of evenly spaced projections required for the faithful reconstruction of an object that has a given diameter (see equation (10) below). The tacit assumption is that projection data have a sufficient spectral signal-to-noise ratio (SSNR) in the whole frequency range in order to reproduce the object faithfully. In electron microscopy, the situation is dramatically different, as the electron dose limitations result in very low SSNR in the individual projections. The suppression of signal is particularly severe in high spatial frequencies, where the signal is affected by the envelope function of the microscope and the high amount of ambient noise, as well as in some low spatial frequency

Pawel A. Penczek • The University of Texas-Houston Medical School, Department of Biochemistry and Molecular Biology, 6431 Fannin, MSB 6.218, Houston, TX 77030, USA

Joachim Frank • Howard Hughes Medical Institute, Wadsworth Center, Albany, NY 12201-0509, USA

regions (due to the influence of the contrast transfer function (CTF) of the electron microscope). In single-particle reconstruction, a satisfactory level of the SSNR in the 3D reconstruction is achieved by including a large number of 2D projections (tens to hundreds of thousands) that are averaged during the reconstruction process. Except for rare cases (Boisset *et al.*, 1998), the angular distribution of projections is not an issue, as the large number of molecules and the randomness of their orientations on the support grid all but guarantee uniform coverage of angular space. The concern is whether the number of projections per angular direction is sufficient to yield the desired SSNR or whether the angular distribution of projections is such that the oversampling of the 3D Fourier space achieved during the reconstruction process will yield the desired SSNR. The resolution measures used in single-particle reconstruction are designed to evaluate the SSNR in the reconstruction as a function of spatial frequency (Penczek, 2002). The 'resolution' of the reconstruction is reported as a spatial frequency limit beyond which the SSNR drops below a selected level, for example below one.

The most commonly used resolution measure in single-particle reconstruction is the Fourier shell correlation (FSC), the 3D equivalent of the Fourier ring correlation (Saxton and Baumeister, 1982). Its advantages are the ease of use and its direct relationship to the SSNR in the data, thus leading to straightforward interpretation of the results. The FSC is evaluated by taking advantage of the large number of single-particle images available: the total data set is randomly split into halves; for each subset a 3D reconstruction is calculated (in 2D, a simple average); and two objects are compared in Fourier space:

$$\mathrm{FSC}(f,g;r)=\frac{\sum_{|\|\mathbf{y}_n\|-r|\le\varepsilon}\hat{f}(\mathbf{y}_n)\hat{g}^*(\mathbf{y}_n)}{\left\{\left(\sum_{|\|\mathbf{y}_n\|-r|\le\varepsilon}\left|\hat{f}(\mathbf{y}_n)\right|^2\right)\left(\sum_{|\|\mathbf{y}_n\|-r|\le\varepsilon}\left|\hat{g}(\mathbf{y}_n)\right|^2\right)\right\}^{1/2}} \tag{1}$$

In equation (1), $\hat{f}$ is the Fourier transform of f, the asterisk indicates complex conjugation, 2ε is a pre-selected shell thickness, the $\mathbf{y}_n$ form a uniform grid in Fourier space and $r = \|\mathbf{y}_n\|$ is the magnitude of the spatial frequency. The FSC yields a 1D curve of correlation coefficients as a function of r. Note that the FSC is insensitive to linear transformations of the objects' densities. An FSC curve everywhere close to one reflects strong similarity between f and g; while an FSC curve with values close to zero indicates the lack of similarity between f and g. Particularly convenient for the interpretation of the results in terms of 'resolution' is the relationship between the FSC and SSNR (Frank and Al-Ali, 1975):

$$SSNR = 2\frac{FSC}{1-FSC}. \tag{2}$$

The factor of two is inserted because the data set was split into halves (Unser *et al.*, 1987); thus, the SSNR of the full reconstruction is twice that derived from half data sets. Equation (2) yields an interpretation of the FSC in terms of 'resolution', understood as the spatial frequency at which the SSNR falls off below a pre-defined (although arbitrary) level. A good choice of the cut-off level is $SSNR(r) = 1.0$ — a level at which the power of the signal in the reconstruction is equal to the power of the noise. According to (2), this corresponds to FSC = 0.333. Another often used cut-off level is FSC = 0.5, at which SSNR in the reconstruction is 2.0.

It is important to keep in mind that to call the FSC a 'resolution measure' is a misnomer as there is no simple relationship between the FSC and the resolution concept used in optics. The latter is based on the notion of *resolvability* of the imaging system and is usually related on the one hand to the wavelength of the radiation used to image the object, and on the other to the point-spread function (PSF) of the imaging system. To appreciate the difference, it is sufficient to notice that an object blurred with a radially symmetric PSF has, according to the FSC, resolution extending to the maximum spatial frequency (this is because FSC is invariant with respect to multiplication of the Fourier transform of the object by a radially symmetric function, see (1)), while the same object can have very poor resolution according to optical criteria. Conversely, an object imaged with ideal PSF (the delta function), but corrupted by strong noise, will generally have low resolution as evaluated by the FSC, while the influence of noise is not considered in the context of optical resolution. For electron microscopic data, the FSC reflects mainly phase errors. Although in general the FSC would also decrease for data that have phases in perfect agreement but errors in amplitudes, this does not normally occur to a significant extent in practice in electron microscopy. Thus, it would be more proper to call the FSC the 'phase consistency measure', as this is what the FSC primarily reflects.

Despite its popularity, the FSC has a number of well-known shortcomings. The relationship between FSC and SSNR (equation 2), and thus the interpretation of the results, is only valid under the assumption that the noise in the data is additive and statistically independent from the signal. With that, it is possible to calculate the variance of the FSC (Saxton, 1978) and also of the SSNR (Penczek, 2002; Unser *et al.*, 1987). For the FSC equal to zero, the variance of its estimate is particularly simple, and equal to the inverse of the number of Fourier coefficients within a given Fourier shell. This is the basis for the 3σ resolution criterion (van Heel, 1987), according to which the resolution is set to the spatial frequency at which the FSC is significantly different from zero, at a significance level corresponding to three standard deviations. However, the exact number of independent Fourier coefficients in the data is all but impossible to estimate, as in addition to the likely reduction of this number due to alignment of the noise component in the data, it is also reduced by the interpolation of the images during various

image processing operations and by the masking of the results in real space; it is further influenced by the size of the window in which the object of interest is embedded, and by other image processing operations performed on the data, which are often difficult to identify. In addition, we have to consider that the variance of the FSC decreases to zero when FSC increases to one (Penczek, 2002), so the error of the assessment of the FSC value is much less for high FSC values. Overall, because of these properties of the FSC, the danger of overestimating the resolution of the results (and, in effect, interpretation of the structure corrupted by artifacts) can be reduced by adopting as a cut-off level a value of the FSC larger then zero, for example 0.5 (Böttcher *et al.*, 1997; Conway *et al.*, 1997; Penczek, 1998).

In addition to the listed shortcomings of the FSC, which in practical applications make it difficult to decide precisely what the resolution of the results might be, one has to be aware that FSC yields a 1D resolution that results from rotational averaging. This makes it impossible to detect anisotropy in the 3D distribution of the SSNR in objects reconstructed from the set of their projections. For example, in the case of overabundant projections, one can expect the resolution to be strongly anisotropic, being higher in Fourier planes perpendicular to the over-represented directions (Boisset *et al.*, 1998). A directional resolution can be calculated by comparing reprojections of the reconstructed structure with the original projection data (Cardone *et al.*, 2005; Unser *et al.*, 2005), but the methods based on this concept are not general, as they will not yield information about resolution in Fourier planes that do not have corresponding projections. A general approach requires calculation of the 3D SSNR, which, as shown in Penczek (2002), can be done to a good degree of accuracy for a class of reconstruction algorithms that are based on interpolation in Fourier space. Although the estimation is not exact, it yields the distribution of SSNR per Fourier voxel, making it possible to detect and quantify any anisotropy in resolution.

The principle of data collection and calculation of the 3D reconstruction in electron tomography is dramatically different from those used in single-particle reconstruction, being more similar to those in basic computed tomography in radiology. There is only one object—a thin section of a biological specimen—and a tilt projection series is collected in the microscope by appropriately tilting the stage holding the specimen. This immediately imposes severe limitations on the total dose that can be applied, making individual projection images rather noisy. The projection images have to be aligned, but in tomography the problem is somewhat simpler than in single-particle reconstruction because fiducial markers can be used to aid the procedure. Moreover, the presence of the markers makes the problem mathematically relatively well defined, although possible distortions of the specimen due to radiation damage can complicate the otherwise straightforward task considerably. Significantly, the data collection geometry in electron tomography can be controlled (within the limits imposed by the mechanics of the specimen holder and by the specimen dimensions).

Two commonly used data collection geometries are single-axis tilting and double-axes tilting, i.e. two single-axis series, with the second collected after rotating the specimen by 90° around the axis coinciding with the direction of the electron beam (usually set as the z-axis of the system) (see also the Introduction by Frank in this volume). In either case, the maximum tilt angle is 60–70° (limited by the effectively increased thickness of the object at high tilt angles and by the geometry of the specimen holder) and the number of projections (limited by the overall electron dose that the specimen can sustain). Another possibility is to use conical tilting, where the specimen tilt angle remains constant while the stage is being rotated in equal increments around the axis perpendicular to the specimen plane. In all three geometries, part of the Fourier transform of the reconstructed object is undetermined. In single-axis tilting, the undetermined region is called the missing wedge. The double-tilt tomography was originally conceived to alleviate the problem and, although the quality of double-tilt reconstructions is markedly improved, a large portion of the Fourier space still remains undetermined, forming the so-called missing pyramid. Particularly discomforting is the fact that in all data collection geometries, Fourier information along the z-axis is missing. This all but eliminates from the 3D reconstruction those features that are planar in x–y planes, making it difficult to study objects that are dominated by such features, such as membranes.

In electron tomography, there is only one projection per projection direction, so the evaluation of resolution based on the availability of multiple projections per angular direction, as practiced in single-particle reconstruction, is not applicable. It was recognized early that splitting of the available set of projections into halves and comparing the two reconstructions in Fourier space using the FSC technique would lead to an underestimation of the resolution, as each of the data sets would have projections spaced with an increased angular step, which—as follows from Crowther's formula (Crowther *et al.*, 1970) (see equation (10))—would necessarily decrease the resolution. The recently published papers on various aspects of resolution measures do not address the core of the issue, nor do they provide operational solutions (Cardone *et al.*, 2005; Penczek, 2002; Unser *et al.*, 2005).

The assessment of resolution in tomography has to include a combination of two key aspects of resolution evaluation in reconstructions of objects from their projections: (i) the distribution of projections should be such that the Fourier space is, as much as possible, evenly covered to the desired maximum spatial frequency; and (ii) the SSNR in the data should be such that at the resolution claimed, the signal is sufficiently high. In electron microscopy, both in single-particle reconstruction and in tomography, there is an additional loss of SSNR in the reconstructed object due to errors in alignment of projections. Unlike in single-particle reconstruction, in tomography these errors are relatively small and, if fiducial markers are used to align the projections, they are uncorrelated with the signal component in the data. In this chapter, we will demonstrate that in the case of 3D reconstructions from 2D projections, an assessment of resolution that fulfills the

two requirements listed is indeed possible. The principle of the method is based on the observation that in this case, the projections will always have redundant components in Fourier space extending to maximum spatial frequency (incidentally, this statement is not true for reconstructions of multidimensional objects from 1D projections). By taking advantage of these redundancies and by assuming isotropy of SSNR in projections, it is possible to estimate the overall resolution in the reconstruction.

2. 3D RECONSTRUCTION FROM PROJECTIONS AS AN INVERSE PROBLEM

In its design, the 3D single-axis tilt tomography is very similar to the early applications of computed tomography to 2D scanning of the interior of the human body (for a review, see Natterer and Ritman, 2002). The reason is that in the case of the single-axis tilt data collection geometry, the 3D problem is reduced to a set of essentially independent 2D problems, the solution of each yielding one 'slice' of the structure in a plane perpendicular to the tilt axis. A 'stack' of these slices forms the final 3D object. Thus, principles of single-axis tilt tomography can be conveniently discussed in terms of 2D computed tomography. Whenever necessary, we will refer to fully 3D tomography, i.e., double-tilt tomography (Penczek *et al.*, 1995) or the recently introduced conical tomography (Lanzavecchia *et al.*, 2005).

Within the linear, weak-phase-object approximation of the image formation process in the microscope (Wade, 1992), the images (projections) observed represent line integrals (ray transforms) of the Coulomb potential of the specimen under examination, modified by the CTF of the microscope. Although the CTF decreases toward the origin of Fourier space, there is no loss of information in this region in higher frequencies, the information is lost at frequencies at which CTF is equal zero. To prevent the loss of high-frequency information, in electron tomography the underfocus setting of the microscope is usually selected such that the first zero of the CTF coincides with the maximum expected resolution. In effect, it is justified to ignore the influence of the CTF on the resolution.

Considering the general d-dimensional problem, we denote by S^{d-1} a unit sphere in $\mathbf{R}^d$. A direction of projection onto a $(d\text{-}1)$-dimensional hyperplane $\mathbf{x} \in \boldsymbol{\theta}^{\perp}$ is given by a vector $\boldsymbol{\theta} \in S^{d-1}$, where $\boldsymbol{\theta}^{\perp} = (\mathbf{y} \in \mathbf{R}^d : \boldsymbol{\theta} \cdot \mathbf{y} = 0\}$; $\boldsymbol{\theta} \neq \mathbf{0}$, then $\boldsymbol{\theta}^{\perp}$ is the central hyperplane perpendicular to $\boldsymbol{\theta}$.

With this notation, the d-D ray transform of function f is defined by (Natterer and Wübbeling, 2001):

$$(Ph)(\boldsymbol{\theta}, \mathbf{x}) = \int_{\mathrm{R}} f(\mathbf{x} + t\boldsymbol{\theta}) dt, \quad \boldsymbol{\theta} \in S^{d-1}, \quad \mathbf{x} \in \boldsymbol{\theta}^{\perp} \tag{3}$$

The function

$$g_{\theta} = P_{\theta} f = (Pf)(\theta, \cdot) \tag{4}$$

is called a (d-1)-dimensional parallel beam projection of f with direction $\boldsymbol{\theta}$. In single-axis tilt tomography, we parameterize $\boldsymbol{\theta}$ by the tilt angle θ[1] (the angle between the direction of the electron beam and the normal to the microscope stage plane): $\boldsymbol{\theta} = (\sin\theta, \cos\theta)^T$. The ray transform is related to the Fourier transform by the central section theorem (Natterer and Wübbeling, 2001), which states that

$$(P_\theta f)\hat{}\,(\mathbf{x}) = \hat{f}(\mathbf{x}), \quad \mathbf{x} \in \boldsymbol{\theta}^\perp, \tag{5}$$

i.e., each Fourier transform $(P_\theta f)\hat{}$ of projection $P_\theta f$ yields the Fourier transform of f on the central hyperplane, whose normal is parallel to the projection direction. Note that we also write Ff instead of $\hat{f}$.

Pf is an integral transform, and the reconstruction problem belongs to a class of inverse problems. The invertibility of the ray transform has been extensively studied (Faridani, 2003; Natterer, 1986) and, for continuous functions, analytical solution have been derived (see Natterer, 1986). For electron tomography, the two important questions are: (i) what should be the distribution of projection directions $\boldsymbol{\theta}$ for the ray transform to be invertible; and (ii) is the solution unique? The required properties of the distribution of projections can be deduced from the central section theorem: the distribution of projections should be such that their Fourier transforms cover the whole Fourier space continuously. Indeed, since the Fourier transform of each projection $\hat{g}_\theta$ is the cross-section of $\hat{f}$ that is perpendicular to $\boldsymbol{\theta}$ in Fourier space, their continuous set yields $\hat{f}$ and thereby determines $f = F^{-1}\hat{f}$. In 2D, it follows that (for asymmetric objects) projection directions should continuously fill the angular range $[0, \pi[$ (because f is a real function, its Fourier transform is Friedel-symmetric, and it follows that a projection in projection in direction θ is Friedel-related to the projection in direction $\theta + \pi$: $\hat{g}_\theta(u) = \hat{g}^*_{\theta+\pi}(u)$; thus, in real space, the two are mirror-related about the origin of the system of coordinates $g_\theta(x) = g_{\theta+\pi}(-x)$). For the 3D ray transform, the requirement for its invertibility is known as Orlov's condition (Orlov, 1976), which states that the inversion of the 3D ray transform is possible, and the solution is unique, if there is a continuous set of projections g_θ corresponding to the motion of the vector $\boldsymbol{\theta}$ over any continuous line connecting the opposite points on the unit sphere. Again, as in 2D, projections in 3D are mirror-related. If we parameterize $\boldsymbol{\theta}$ by the tilt angle θ and the azimuthal angle φ as

$$\boldsymbol{\theta} = (\cos\varphi\sin\theta, \sin\varphi\sin\theta, \cos\theta)^T \tag{6}$$

and use Friedel symmetry of the 3D Fourier transform of f, we find that $g_{\theta,\varphi}(\mathbf{x}) = g_{\theta+\pi,\,\varphi+\pi}(-\mathbf{x})$.

Unlike in 2D, in 3D there is an infinite number of different data collection geometries that fulfill Orlov's condition. In electron tomography, we

[1] For the convention of Euclerian angles used, see Frank (2006).

use three data collection geometries but, due to the technical limitations of the microscope (design of the stage), none fulfills Orlov's condition.

1. *Single-axis tilt geometry*: the vector $\boldsymbol{\theta}$ moves on a grand circle covering a fragment of its length. Assuming that the tilt axis coincides with the y-axis of the coordinate system in which the z-axis corresponds to the direction of the electron beam, we have

$$\varphi = 0, \quad -\theta_{\max} \le \theta \le \theta_{\max}, \tag{7}$$

where $0 < \theta_{\max} < 90°$[2] is the maximum tilt angle (usually 60° or 70°). The portion of Fourier space that is left unmeasured is referred to as the *missing wedge*.

2. *Double-tilt geometry* (Penczek *et al.*, 1995): the vector $\boldsymbol{\theta}$ moves on two perpendicular grand circles, in each case covering a fragment of the grand circle's length:

$$\begin{aligned} &\text{first series} \quad \varphi = 0, \quad -\theta_{\max} \le \theta \le \theta_{\max}, \\ &\text{second series} \quad \varphi = 90°, \quad -\theta_{\max} \le \theta \le \theta_{\max}. \end{aligned} \tag{8}$$

The portion of Fourier space that is left unmeasured is referred to as *missing pyramid*.

3. *Conical geometry* (Lanzavecchia *et al.*, 2005): the vector $\boldsymbol{\theta}$ covers the whole length of the circle parallel to the x–y plane that forms a cone with its tip at the origin of the coordinate system:

$$0 \le \varphi < 90°, \quad \theta = \theta_{\text{tilt}}, \tag{9}$$

where θ_{tilt} is the tilt angle. Thus, in this data collection geometry, all images are collected at the same tilt, while in the two previous geometries the tilt angle varies. The portion of Fourier space that is left unmeasured is referred to as the *missing cone*.

The reconstruction problems for data collections geometries stated in equations—and used in electron tomography are known as *limited-angle problems*. If the number of projections within the limited angular range were infinite, it would be possible to recover f uniquely (Keinert, 1989). Recovery of f from limited angular-range projections is equivalent to analytical continuation of $\hat{f}$ and is very unstable and sensitive to errors in the data. More practical approaches are based on the availability of *a priori* knowledge about the reconstructed object, such as limited spatial support, non-negativity, similarity to known objects, etc. Formally, the process of enforcing selected constraints is best described in the framework of the theory of Projections Onto Convex Sets (POCS) (Sezan, 1992; Sezan and

[2] In electron microscopic tomography, the maximum tilt angle is usually given in degrees. In agreement with this convention, we give all angles and arguments of trigonometric functions in degrees instead of radians, hence the occasional appearance of the radians-to-degrees factor $180°/\pi$ in some equations.

Stark, 1982; Youla and Webb, 1982), introduced into electron microscopy by Carazo and co-workers (Carazo, 1992; Carazo and Carrascosa, 1986, 1987). However, convincing results in the application of POCS to low- and intermediate-resolution electron microscopic data are scarce; therefore, in the following, we will assume that meaningful statements can be made about the resolution only within regions of Fourier space that are covered by Fourier transforms of the measured projections.

The discussion so far concerned an infinite number of projections, while in practice only a finite number can be measured. It is well known that a function f cannot be uniquely recovered from a finite number of its projections (Faridani, 2003). The reason is that a function reconstructed from a finite number of its projections can contain so-called null functions, i.e., functions whose projections calculated in directions of measured data are exactly zero (this is why the null functions are also referred to as ghosts (Louis, 1984)). It follows that a reconstructed function can be significantly different from the original function, but it can still agree perfectly well with each of its measured projections. The ghosts will appear as artifacts and they can be introduced into the reconstructed object by various image processing steps, such as interpolation, suboptimal weighting functions, inconsistencies in the data, and so on. However, in practice, the situation is not as difficult as it would seem from these statements. It can be shown that for the ray transform, the null functions are high-frequency objects (Louis, 1984; Maass, 1987) and, in practical applications, they can be suppressed by appropriate low-pass filtration. Nevertheless, even if ghosts do not constitute a serious problem, their potentially adverse influence underscores the necessity to estimate the resolution of the reconstructed object properly and to apply a Fourier filter that accounts for the uneven distribution of the SSNR.

3. ANGULAR DISTRIBUTION OF PROJECTION ANGLES IN TOMOGRAPHY

In order to reconstruct an object from a finite number of its projections, it is necessary to establish what is the minimum number of projections required (in the absence of noise) and what should be their angular distribution. In the *single-axis tilt geometry*, the problem of 3D reconstruction is reduced to a series of (in principle) independent 2D reconstruction problems for the x–z slices. If the object has a circular support, lengths x and z of the reconstructed object are the same and each is equal to L pixels, with pixel dimensions being $p \times p$ nm. For evenly spaced projections in 2D, the requirement for sufficient sampling in Fourier space yields the following angular step (Bracewell and Riddle, 1967):

$$\Delta\theta = \frac{180^\circ}{\pi}\frac{2}{L}. \tag{10}$$

In electron microscopy, equation (10) is often referred to as Crowther's criterion (Crowther *et al.*, 1970).

However, in order to facilitate the penetration by the electron beam, the objects imaged by electron tomography often are made to have slab geometry, i.e., they are obtained by thin sectioning of larger volumes of biological material and, in effect, their z-dimension is much less than the dimensions in x and y. In effect, in terms of pixels we have:

$$L_z << L_x = L_y = L. \tag{11}$$

Because of the slab geometry of the system (relationship (11)), the angular step between projections should no longer be even. In the context of 3D reconstruction from 2D crystallographic data, a non-even distribution of angles was proposed by Saxton *et al.* (1984) as:

$$\Delta\theta = \frac{180^\circ}{\pi}\frac{2}{L_z}\cos\theta, \tag{12}$$

where we assumed that the projections are sampled at Nyquist frequency and the resolution of the reconstruction is 2 pixels. However, the angular spacing computed according to this equation has the following disadvantages (i) it does not depend on the x–y dimensions of the slab; and (ii) it would approach 0° as the tilt angle approaches 90°. Another distribution of angles can be proposed based on the simple observation that a 2D reconstruction for dimensions given by equation (11) can consistently recover information only within an ellipse with main axes equal to L and L_z, respectively (similarly to the situation for a 2D reconstruction carried out within a square and using 1D projections of equal length and constant sampling step, one can recover only information restricted to a circle):

$$\Delta\theta = \frac{180^\circ}{\pi}\frac{2}{L_z}\sqrt{1-\left(1-L_z^2/L^2\right)\sin^2\theta}. \tag{13}$$

For infinite x–y dimensions of the slab ($L\to\infty$), equations (13) and (12) are equivalent, and for $L_z = L$ (reconstruction within a square) equation (13) yields the familiar dependence on the angular step on the radius of the structure (equation (10)). Generally, equation (13) will yield fewer angles than equation (12) (Table 1), although for large L and small tilt angles the differences are negligible.

The distribution of projection angles in the case of double-axis tilt geometry can be conveniently considered as a natural extension of the single-axis tilt geometry. If the first series is collected using the scheme given by equation (13), then for the second tilt series one can use the same scheme with exclusion of projections — in order to minimize redundancies — that have tilt angles < ϑ. The value of this angle is obtained based on a simple geometrical consideration: if the maximum tilt angle for the first tilt series

TABLE 1. (a) The sequence of projection angles for single-axis tomography as proposed by Saxton *et al.* (1984), taking into account only the thickness of the specimen and (b) taking into account both the thickness and x–y dimensions. The sequence of projection angles for the second tilt series in double-axis tomography assuming a maximum tilt angle (c) $\theta_{max} = 60°$ and (d) $\theta_{max} = 70°$

	(a) Equation (12)	(b) Equation (13)	(c) Equation (15) $\theta_{max} = 60°$	(d) Equation (15) $\theta_{max} = 70°$
0	0.0	0.0		
1	10.0	10.0		
2	19.4	19.4	16.8	
3	28.5	28.5	25.9	25.4
4	36.9	37.0	34.6	34.1
5	44.6	44.8	42.6	42.1
6	51.4	51.8	49.9	49.4
7	57.5	58.0	56.3	55.9
8	62.7	63.5	62.0	61.7
9	67.1	68.3		66.7
10	70.9	72.5		71.1
11	74.1	76.2		
12	76.8	79.4		
13	79.0	82.3		
14	80.9	84.9		
15	82.5	87.4		
16	83.8	89.7		
17	84.9			
18	85.9			
19	86.6			
20	87.3			
21	87.8			
22	88.2			
23	88.6			
24	88.9			
25	89.2			
26	89.4			

All angles were calculated using thickness $L_z = 12$ and $x = y$ dimension, $L = 50$ voxels and assuming resolution of two voxels.

is θ_{max}, ϑ should be such that the angular gap β between second series projections collected at ϑ is no larger than required by the sampling distance at this angle (Fig. 1), yielding the following relationship:

$$\cos^2(\vartheta) = \frac{\cos(\beta) + \cos(2\theta_{max})}{(1 + \cos(2\theta_{max}))\cos^2\left(\beta/2\right)}. \tag{14}$$

If the requested angular gap β is equal to the maximum angular gap due to missing edge, equation yields $\vartheta = 90°$, so there is only one projection from the second tilt series, as expected. For the requested angular gap β

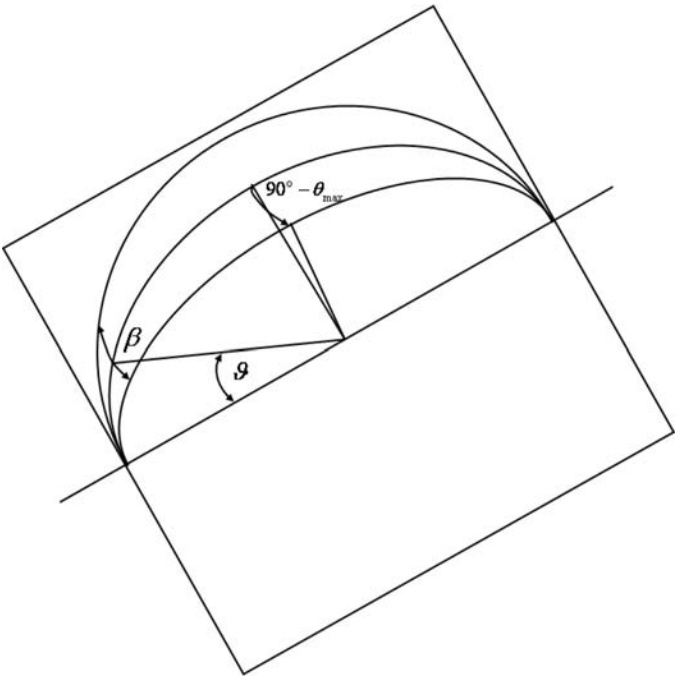


FIGURE 1. Double-tilt data collection geometry. θ_{max} = maximum tilt angle of both series of projections; ϑ = minimum tilt angle of the second series of projections; β = the angular gap of the missing wedge of the first series of projections.

larger than the missing gap wedge, equation (14) has no solution. For a reconstruction within a square, we have $\beta = \frac{180^\circ}{\pi}\frac{2}{L_z}$, and the minimum tilt angle of the second series can be easily found. For example, for L_z = 50, β = 2.3° and θ_{max} = 70°, we obtain ϑ = 6.3°, while for θ_{max} = 60° we have ϑ = 4.0°. For the slab geometry, the problem is more difficult, as the angular gap β depends on the angle ϑ via equation (13):

$$\cos^2(\vartheta) = \frac{\cos\left(\frac{180^\circ}{\pi}\frac{2}{L_z}\sqrt{1-\left(1-L_z^2/L^2\right)\sin^2\vartheta}\right)+\cos(2\theta_{max})}{(1+\cos(2\theta_{max}))\cos^2\left(0.5\frac{180^\circ}{\pi}\frac{2}{L_z}\sqrt{1-\left(1-L_z^2/L^2\right)\sin^2\vartheta}\right)}. \quad (15)$$

This relationship is non-linear; however, for given L, L_z, and θ_{max}, the solution can be found numerically. For example, for L_z = 50 and θ_{max} = 70°, we obtain as the value of the minimum angle of the second tilt series ϑ = 6.2°, while for θ_{max} = 60° we have ϑ = 4.0°. The remaining angles for the second tilt series are found using equation (13). Table 1 contains examples of sequences of projection angles generated using equations (12), (13) and (15).

4. INFLUENCE OF MISSING FOURIER INFORMATION ON RESOLUTION IN TOMOGRAPHIC RECONSTRUCTIONS

The quality of tomographic reconstructions depends on the maximum tilt angle used. The fraction q of missing Fourier space information can be easily calculated for the three data collection geometries, taking as a reference the volume of the slab:

(i) single-axis tilt; missing wedge: $$q_1 = \frac{1}{2}\frac{L_z}{L}\cot\theta_{\max}, \quad (16)$$

(ii) double-axis tilt; missing pyramid: $$q_2 = \frac{1}{3}\left(\frac{L_z}{L}\cot\theta_{\max}\right)^2, \quad (17)$$

(iii) conical; missing cone: $$q_3 = \frac{\pi}{12}\left(\frac{L_z}{L}\cot\theta_{\max}\right)^2. \quad (18)$$

The equations are valid for $L_z \le L$ and $\theta_{\max} > 45°$. Examples of fractions of missing Fourier information resulting from equations (16)–(18) are given in Table 2. In the case of a reconstruction within a square ($L_z = L$), the fractions of missing Fourier space information appear to be large, although it helps to put them in perspective. For an object that has a spatial resolution of $\frac{1}{20.0}\frac{1}{\text{Å}}$, isotropic loss of 29% of Fourier information would result in a spatial resolution of $\frac{1}{22.4}\frac{1}{\text{Å}}$ — it certainly would be close to impossible to notice the resulting deterioration in the otherwise noisy tomograms. For the more realistic slab geometry, the results are even more striking as the loss of information is very small, indeed. One has to conclude that the problems with the missing information in tomography are mainly caused by the fact that the loss is entirely along (and in directions that form oblique angle with) one of the axes of the coordinate system, resulting in anisotropic, object-dependent artifacts. This analysis also indicates that the occasionally quoted assessment of the loss of resolution due to the missing wedge or cone might be to a certain extent misleading. It is possible to calculate

TABLE 2. Examples of fractions of missing Fourier information for three data collection geometries calculated using equations (16)–(18)

	$L_z/L = 1.0$		$L_z/L = 0.2$	
	$\theta_{\max} = 60°$	$\theta_{\max} = 70°$	$\theta_{\max} = 60°$	$\theta_{\max} = 70°$
Single-axis	29%	18%	5.80%	3.6%
Double-axis	11%	4%	0.44%	0.16%
Conical	9%	3%	0.36%	0.12%

$\theta_{\max}$ is the maximum tilt angle. Reduction of the section thickness to the ratio of 0.2 results in a dramatic reduction of the fraction of missing Fourier information and in a seeming increase in resolution.

analytically the PSF corresponding to the particular data collection geometry. These PSFs are elongated in the direction of the missing wedge or cone (in our geometry, the z-axis) and one can interpret the width of the central lobe as the expected elongation point-like features in the tomogram (Radermacher, 1988). Although this is certainly true for small spherical objects, for more complicated objects the distortions will be object dependent. In general, it is possible to predict that flat objects extending in a plane perpendicular to the missing wedge or cone axis will be severely deteriorated, or that it might be entirely missing in the reconstruction. This is illustrated in Fig. 2, in which we demonstrate that elongated objects that are oriented perpendicularly to the direction of a missing wedge vanish entirely after Fourier information is removed systematically from the Fourier transform of the original image. Moreover, this effect and, more generally, the extent of deterioration of images are only marginally affected by the change of the missing wedge angle from 60 to 70° (Fig. 2).

5. RESOLUTION ESTIMATION IN TOMOGRAPHY

Although there is a consensus on the general concept of resolution in electron tomographic reconstructions, i.e., that that the resolution should be derived from (i) the number and angular distribution of projections and (ii) the noise level in the data, a systematic study of the problem is lacking. One of the reasons is that an electron tomographic study of any particular object is unique. What we mean by uniqueness is that biological objects visualized by tomography, although they are usually members of certain broader categories (e.g. mitochondria), are not reproducible and they have inherent variability. Thus, unlike in crystallography or in single-particle studies, repeated reconstructions of the object from the same category will yield structures that have similar overall features, but are also significantly different. In effect, this makes it impossible to study resolution of tomographic reconstructions in terms of statistical reproducibility. Moreover, because of the dose limitations, there is only one projection per each angular direction, so the standard approach to SSNR estimation based on dividing data set into halves is not applicable. In order to develop a working approach, one has to consider two aspects of the resolution estimation in tomography: (i) angular distribution of projections; and (ii) estimation of the SSNR in the data. The first aspect together with the conditions for proper distributions of projection angles in slab geometry were discussed in previous sections. In this section, we will introduce a method for estimation of the SSNR that takes advantage of inherent Fourier space redundancies in the inversion of the 3D ray transform.

Three recently published papers devoted to resolution measures claimed prospective applications to tomography, but were focused only on the second aspect of the problem, i.e., estimation of the SSNR. (Cardone *et al.*, 2005; Penczek, 2002; Unser *et al.*, 2005). In all cases, the proposed

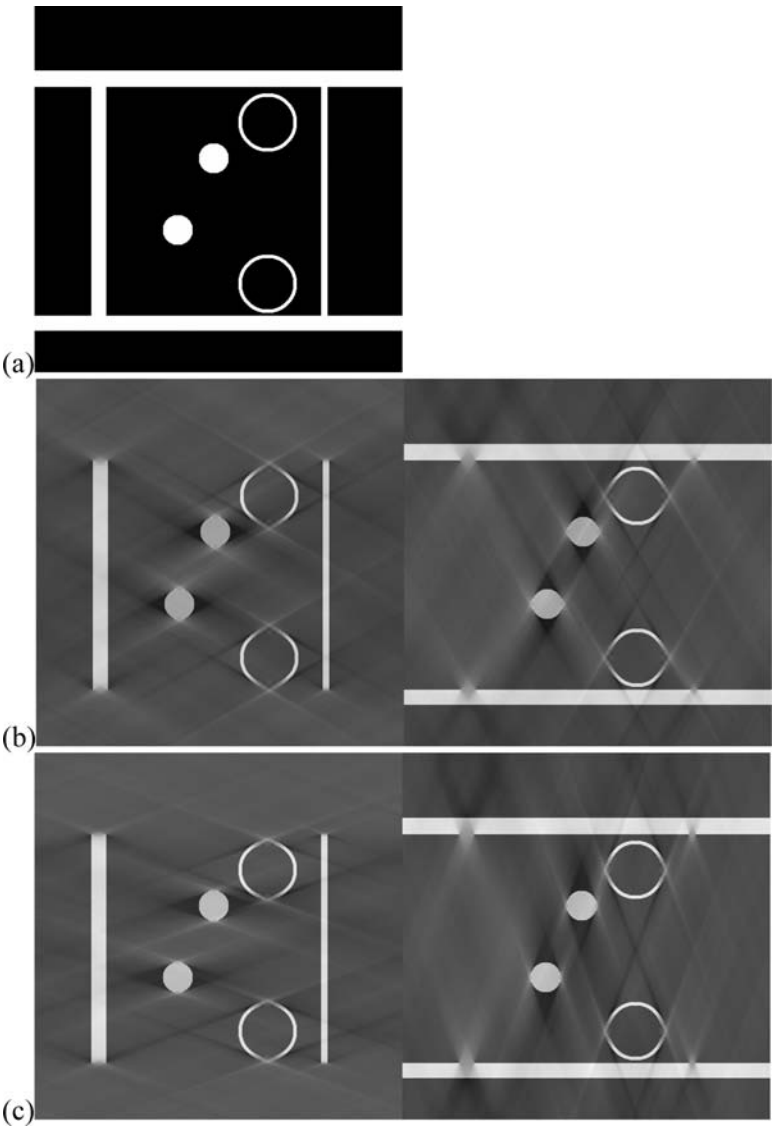


FIGURE 2. Effects of the missing wedge on a 2D image. (a) Model image. (b) Maximum tilt angle of 60° and directions of the missing wedge in the *y* and *x* directions, respectively. (c) Maximum tilt angle of 70° and directions of the missing wedge in the *y* and *x* directions, respectively.

solution was based on extensions of resolution measures routinely used in single-particle reconstruction. In Penczek (2002), the author developed a 3D SSNR for a class of 3D reconstruction algorithms that are based on interpolation in Fourier space. The work is a direct extension of an earlier work by Unser *et al*. (1987), who initially proposed a 2D SSNR. In 3D, the

problem is more complicated, as the resolution measure has to account for the Fourier space interpolation between polar and Cartesian coordinates necessary during the reconstruction step. For a nearest-neighbor interpolation, the extension of 2D definition is straightforward, but the SSNR is underestimated due to the very simple interpolation scheme used. A marked improvement can be achieved by using the gridding-based interpolation scheme (Penczek *et al.*, 2004). The 3D SSNR works well for isolated objects and within a limited range of spatial frequencies. The measure requires calculation of the Fourier space variance, so it will yield correct results only to the maximum frequency limit within which there is sufficient overlap between Fourier transforms of projections.

An alternative approach to the estimation of the 3D SSNR was recently proposed by Unser *et al*. (2005). In order to avoid the reliance on a particular reconstruction algorithm, the authors proposed to estimate the SSNR in 2D by comparing reprojections of the reconstruction structure with the original input projection data, and to average the contribution in 3D Fourier space to obtain the 1D dependence of the SSNR on spatial frequency. The process of 3D reconstruction involves averaging of the data in 3D Fourier space whose extent is difficult to assess. The authors propose to estimate it by repeating the calculation of the 3D SSNR for simulated data containing white Gaussian noise. The ratio of the two curves, i.e., the one obtained from projection data and the other obtained from simulated noise, yields the desired true SSNR of the reconstructed object. Although the method is appealing in that it can be applied to any (linear) 3D reconstruction algorithm, the application to tomography is of doubtful validity since, similarly to the method of Penczek (2002), it requires sufficient oversampling to yield correct results. A more serious disadvantage is that the calculation of the 3D SSNR as proposed by Unser *et al*. (2005) actually yields the 2D SSNR of the input data, not the 3D SSNR of the reconstruction. This can be seen from the fact that in averaging of 2D contributions to the 3D SSNR, the authors neglect weighting due to uneven distribution of projections in Fourier space (see Vainshtein and Penczek (2006) for the role of weighting in reconstruction from projections). In effect, the SSNR calculated according to their recipe will be dominated by contributions from projections data that have similar angular directions (for example, in the case of double-tilt, by untilted projections).

An interesting approach to the resolution estimation was introduced by Cardone *et al*. (2005), who proposed to calculate, for each available projection, two 2D Fourier ring correlation (FRC) curves: (i) between selected projections and the reprojections of the volume reconstructed using the whole set of projections; and (ii) between selected projections and reprojections of the volume reconstructed with the selected projection omitted. The authors showed that the ratio of these two FRC curves is related to the SSNR in the volume in the Fourier plane lying perpendicular to the projection direction, as per the central section theorem (5). The authors

propose to calculate the SSNR of the whole tomogram by summing the independent contributions from respective FRC curves, normalizing the results, and calculating the ratio. However, it is straightforward to note that the method suffers from the same disadvantages as the method of Unser *et al.*, i.e., (i) it does not account for the SSNR in the data lying in non-overlapping regions and (ii) because of the omission of reconstruction weights, it does not yield the proper 3D SSNR.

In order to introduce our approach to the estimation of resolution in tomograms, we begin by noting that, given a set of noisy projections collected such that the Fourier space is sufficiently covered (as given by equations (12), (13), and (15)), it is impossible to state the resolution of the reconstruction unless the SSNR of the data is known. However, unlike in 2D, any data collection geometry in a 3D reconstruction from 2D projections (inversion of the ray transform) will necessarily have Fourier space redundancies within a whole range of Fourier space. This is a simple consequence of the central section theorem (5): any non-trivial arrangement of central sections will share at least one common line and, as a result, will yield oversampled regions in Fourier space. By exploring these redundancies, and by using the standard FSC approach, it is possible: (i) to calculate the SSNR in certain regions of Fourier space; (ii) to calculate the SSNR in individual projections in the entire range of spatial frequencies; and (iii) assuming isotropy of the data, to infer/deduce the resolution in non-redundant regions of Fourier space. Given the SSNR in projections and equations (12), (13), and (15), which tell us what the angular spacing of the projections is, it becomes possible to calculate the distribution of the SSNR in the reconstructed 3D object.

In order to illustrate the method, we will analyze the single-axis tilt data collection geometry with the simplifying assumptions that the reconstruction was performed within a square ($L_z = L$) and that the angular step of projection data was selected in agreement with equation (10). In effect, given sufficient SSNR in the projection data, it should be possible to reconstruct the object within the entire range of spatial frequencies. We note that in the direction perpendicular to the tilt axis (y), the oversampling decreases monotonically with the spatial frequency, while along the tilt axis it remains constant and equal to the factor given by the number of projections (Fig. 3). Moreover, in the regions adjacent to the tilt axis, there is a high degree of oversampling within a cylinder extending to the maximum spatial frequency, which makes it possible to calculate the SSNR in the data within this region by comparing 3D reconstructions calculated from even- and odd-numbered projections. Finally, by assuming that the distribution of SSNR in 2D projections is isotropic, we can infer the SSNR in the perpendicular, undersampled direction. Based on this information, we can state (using a selected cut-off level) what the resolution will be in the undersampled direction. The method is even more easily applicable to the double-tilt geometry, in which there are two perpendicular redundant directions, and to conical tomography, in which there are numerous pair-wise common lines

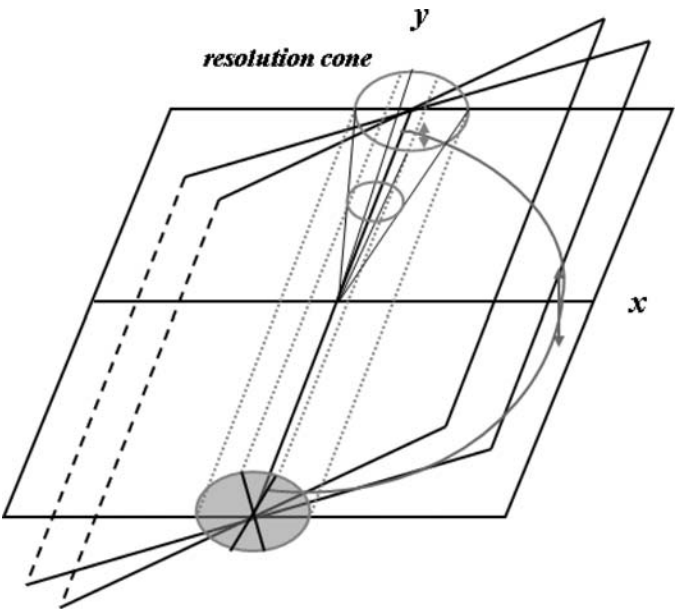


FIGURE 3. Estimation of SSNR of projection data and resolution of tomographic reconstruction. The half-circle = the region in 3D Fourier space that has the same modulus of spatial frequency $|\boldsymbol{k}|$. Long vertical arrow = the larger gap in Fourier space in the direction perpendicular to the tilt axis at the spatial frequency $|\boldsymbol{k}|$. Short arrow = the smaller gap in Fourier space close to the position of the tilt axis *y*, also at the spatial frequency $|\boldsymbol{k}|$. The resolution is calculated using the modified FSC approach by carrying out the summation of Fourier coefficients located on circular cross-sections of the resolution cone. For a small opening angle of the resolution cone, the oversampling of the Fourier space is approximately equal to the factor given by the number of projections.

between Fourier space projection planes. It is also important to note that, as long as the alignment of the projection data was done using fiducial markers, the SSNR analysis will not be adversely affected by the problem of 'alignment of noise' common to reconstructions aligned or refined using correlation techniques (Stewart and Grigorieff, 2004).

The proposed method of calculation of the resolution in tomography was tested using single-tilt series of projections of a test 3D object placed in a cube with 128^3 voxels. In this case $L_z = L$, so the angular step was selected according to Crowther's criterion (10) and set to $\Delta\theta = 0.5°$, resulting in 360 equispaced projections. The projections were corrupted by Gaussian noise such that the SNR in real space was ~200 (the relatively low level of noise was selected to demonstrate the performance of the method in the high-frequency range). In order to simplify the analysis, we generated the full range of projections, so there was no missing wedge in the reconstruction. We calculated two 3D reconstructions using odd- and even-numbered projections, respectively.

We compared the two test reconstructions in Fourier space using a modified FSC approach. Instead of carrying summations in equation (1) over spheres, we included in summations only those Fourier coefficients that were located on cones with the axes placed on the *y*-axis of the coordinate system (which coincides with the tilt axis, see Fig. 3). Thus, for a given opening angle of this *resolution cone* and for a given modulus of spatial frequency, the resulting FSC curve was calculated using only Fourier coefficients from a region of the 3D Fourier space that had uniform overlap between 2D Fourier transforms of projections. Since the number of projections was known, this overlap could be calculated based on the central section theorem (5) using simple geometrical considerations. Depending on the opening angle of the resolution cone (measured from the tilt axis *y* of the coordinate system), we obtained the resolution estimate for a given conical region in Fourier space. We performed the calculation for cone angles varying from 2 to 90° (Fig. 4). The FSC curve for the cone of 2° reflects the resolution along the *y*-axis, which is oversampled nearly 360 times. With increased opening angle of the cone, the values of FSC coefficients decrease, down to the lowest values for the opening angle of 90°. The FSC curve calculated using the opening angle of 90° reflects the resolution in the grand circle in the *x*–*z* central section of the 3D Fourier space,

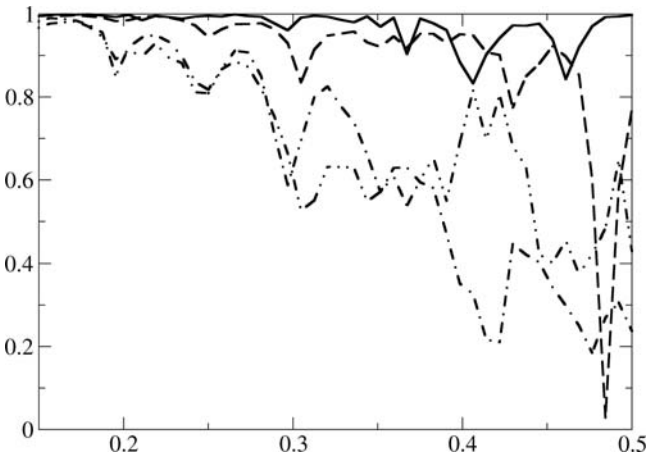


FIGURE 4. Calculation of directional resolution using the modified FSC approach that takes advantage of redundancies in Fourier space in tomographic reconstructions from projections. The FSC curves are calculated using Fourier coefficients that are located in 3D on cones that have axes placed on the single-tilt axis (*y*) of the system. Depending on the opening angle of the cone, the resulting FSC curves reflect resolution in a given region of Fourier space: (——) 2°, the resolution on circles along the direction of the tilt axis with the highest degree of the oversampling; (— —) 10°; (— · —) 40°; (— ·· —) 90°, the resolution on the great circle perpendicular to the tilt axis, thus with the lowest degree of oversampling.

thus perpendicular to the direction of the tilt axis y. This is the direction with the least degree of oversampling, and thus with the least favorable resolution.

The SSNR in the data was calculated using the FSC results calculated for the cone with 2° opening angle. The SSNR in the reconstruction was calculated using the relationship between SSNR and FSC (equation (2)) and then divided by the number of projections included in respective reconstructions to yield finally the SSNR in a single projection. The actual SSNR level in the data was derived from the FSC curve calculated between two projections of the test structure, generated using the same angular direction, but corrupted by two different realizations of Gaussian noise. The agreement between the SSNR calculated using the modified FSC approach and the actual SSNR level is excellent (Fig. 5).

The final step of analysis is to combine the information about the calculated level of SSNR in the data with the knowledge of the angular step of the single-axis tilt series. The latter was set to 0.5° — sufficient for the full coverage of Fourier space. Based on the FSC comparison of 3D reconstructions calculated from data sets split into halves, thus each using a set of projections spaced by 1°, one would have to set the resolution of the test

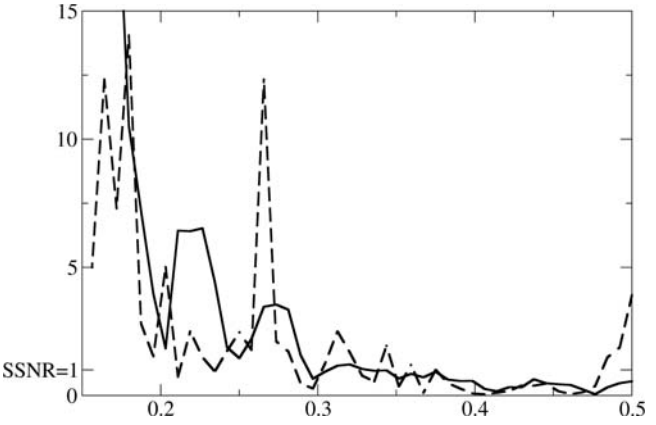


FIGURE 5. Calculation of the SSNR in the data based on oversampled region of Fourier space in a 3D tomographic reconstruction. (——) an SSNR curve derived from an FRC curve calculated between two projections of the test structure generated using the same angular direction, but corrupted by two different realizations of the Gaussian noise. This curve reflects the true distribution of the SSNR in the projection data that could be calculated if redundant projections of the specimen could be collected in the electron microscope. In the practice of electron microscopy, this information is not available; (— —) SSNR in the data derived from the modified FSC calculation with the cone angle of 2° (see Fig. 4). The two curves agree very well, demonstrating that it is possible to derive the level of SSNR in the data by taking advantage of redundancies in the coverage of 3D Fourier space by Fourier transforms of 2D projection data.

reconstruction in the direction perpendicular to the direction of the tilt axis to spatial frequency ~0.28, if a cut-off level of SSNR = 1 was selected (see Fig. 4). However, the calculation of the SSNR using the modified FSC approach yields SSNR > 1 extending to a spatial frequency of ~0.38, thus significantly higher.

6. CONCLUSIONS

The determination of resolution in electron tomography requires a multifaceted approach. In order to decide whether a given reconstruction from electron microscopic projection data constitutes an acceptable approximation of the original (albeit unknown) structure, it is necessary to consider a number of general issues: (i) the invertibility of the ray transform, particularly in the context of the slab geometry commonly used in electron tomography; (ii) the dependence of the distribution of projection angles on the data collection geometry; (iii) the influence of the limited angular range of the projection data; and (iv) the determination of the SSNR in the data.

Although the reconstruction of a function from a finite set of its projections is a difficult inverse problem that does not have a unique solution, in practice the uncertainty due to possible introduction of artifacts is low, as these artifacts mainly affect high spatial frequencies in the reconstructed object and can be suppressed by the appropriate low-pass filtration. Much more serious difficulties are caused by the limited tilt range of the specimen stage. In effect, the reconstruction problems for data collection geometries used in electron tomography fall into the category of limited-angle problems. Theoretically, given an infinite number of projections within the limited angular range, it would still be possible to reconstruct the object uniquely. However, in practice, only a limited number of projections is available, the data are noisy and the orientations of projections are known to a limited accuracy. Although it is possible improve — to a degree — the quality of the reconstruction using approaches that are based on the availability of *a priori* knowledge about the reconstructed object (POCS), in the absence of objective and independent standards the success of such methods is difficult to assert. In addition, these approaches often depend on non-linear operations of the data that will result in apparent, but difficult to verify improvements in resolution, particularly if the latter is assessed using Fourier space-based criteria.

The distribution of projection angles is derived based on the requirement for a possibly uniform coverage of the Fourier space by finite-thickness central sections corresponding to Fourier transforms of projections. For standard slab geometry of the sample in electron tomography, for single-axis tilt, the angular step of projections has to decrease with the increased tilt angle. If the double-tilt data are collected, the same scheme has to be

applied to the second tilt series, although low-tilt projections are redundant and should be omitted. The minimum tilt angle of the second tilt series is derived based on simple geometrical consideration. As a consequence, we showed that for a maximum tilt angle of 70°, the number of second series projections that have to be collected is significantly reduced.

For the estimation of the SSNR in the reconstruction, we introduced a novel approach based on the modified FSC analysis. We demonstrated that in the case of a 3D reconstruction from 2D projections (inversion of the ray transform), the object will necessarily have, as a consequence of the central section theorem, Fourier space redundancies within the entire range of Fourier space. In the proposed approach, these redundancies are used to calculate FSC coefficients within oversampled regions of Fourier space. This is done by comparing Fourier transforms of objects reconstructed from even- and odd-numbered projections, respectively. Because the degree of the oversampling in the region adjacent to the tilt axis (or axes, for double-tilt geometry) is high, it is possible to obtain a robust estimate of the FSC and, in consequence, the SSNR. Next, we have to assume that the SSNR in the data is isotropic and this allows us to infer what the SSNR in the under-sampled regions of Fourier space is. Finally, given the known distribution of projections angles, it becomes possible to calculate the distribution of the SSNR in the reconstructed 3D object. We demonstrated the efficacy of the new method using a simple single-axis tilt simulated data set. However, the method is quite general and applicable to all three data collection geometries used in electron tomography.

REFERENCES

Boisset, N., Penczek, P. A., Taveau, J. C., You, V., Dehaas, F. and Lamy, J. (1998). Overabundant single-particle electron microscope views induce a three-dimensional reconstruction artifact. *Ultramicroscopy* **74**:201–207.

Böttcher, B., Wynne, S. A. and Crowther, R. A. (1997). Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature* **386**:88–91.

Bracewell, R. N. and Riddle, A. C. (1967). Inversion of fan-beam scans in radio astronomy. *Astrophys. J.* **150**:427–434.

Carazo, J. M. (1992). The fidelity of 3D reconstruction from incomplete data and the use of restoration methods. In *Electron Tomography* (J. Frank, ed.). Plenum, New York, pp. 117–166.

Carazo, J. M. and Carrascosa, J. L. (1986). Information recovery in missing angular data cases: an approach by the convex projections method in three dimensions. *J. Microsc.* **145**:23–43.

Carazo, J. M. and Carrascosa, J. L. (1987). Restoration of direct Fourier three-dimensional reconstructions of crystalline specimens by the method of convex projections. *J. Microsc.* **145**:159–177.

Cardone, G., Grünewald, K. and Steven, A. C. (2005). A resolution criterion for electron tomography based on cross-validation. *J. Struct. Biol.* **151**:117–129.

Conway, J. F., Cheng, N., Zlotnick, A., Wingfield, P. T., Stahl, S. J. and Steven, A. C. (1997). Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. *Nature* **386**:91–94.

Crowther, R. A., DeRosier, D. J. and Klug, A. (1970). The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. *Proc. R. Soc. A* **317**:319–340.

Faridani, A. (2003). Introduction to the mathematics of computed tomography. In *Inside Out: Inverse Problems and Applications*, vol. 47 (G. Uhlmann, ed.). Cambridge University Press, Cambridge, pp. 1–46.

Frank, J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press, New York.

Frank, J. and Al-Ali, L. (1975). Signal-to-noise ratio of electron micrographs obtained by cross correlation. *Nature* **256**:376–379.

Keinert, F. (1989). Inversion of *k*-plane transforms and applications in computer-tomography. *SIAM Rev.* **31**:273–298.

Lanzavecchia, S., Cantele, F., Bellon, P., Zampighi, L., Kreman, M., Wright, E. and Zampighi, G. (2005). Conical tomography of freeze-fracture replicas: a method for the study of integral membrane proteins inserted in phospholipid bilayers. *J. Struct. Biol*. **149**:87–98.

Louis, A. K. (1984). Nonuniqueness in inverse Radon problems—the frequency-distribution of the ghosts. *Math. Z.* **185**:429–440.

Maass, P. (1987). The X-ray transform—singular value decomposition and resolution. *Inverse Probl.* **3**:729–741.

Natterer, F. (1986). *The Mathematics of Computerized Tomography*. John Wiley & Sons, New York.

Natterer, F. and Ritman, E. L. (2002). Past and future directions in x-ray computed tomography (CT). *Int. J. Imaging Syst. Technol*. **12**:175–187.

Natterer, F. and Wübbeling, F. (2001). *Mathematical Methods in Image Reconstruction*. SIAM, Philadelphia.

Orlov, S. S. (1976). Theory of three-dimensional reconstruction 1. Conditions for a complete set of projections. *Soviet Phys. Crystallogr.* **20**:312–314.

Penczek, P. (1998). Measures of resolution using Fourier shell correlation. *J. Mol. Biol*. **280**:115–116.

Penczek, P., Marko, M., Buttle, K. and Frank, J. (1995). Double-tilt electron tomography. *Ultramicroscopy* **60**:393–410.

Penczek, P. A. (2002). Three-dimensional spectral signal-to-noise ratio for a class of reconstruction algorithms. *J. Struct. Biol*. **138**:34–46.

Penczek, P. A., Renka, R. and Schomberg, H. (2004). Gridding-based direct Fourier inversion of the three-dimensional ray transform. *J. Opt. Soc. Am. A* **21**:499–509.

Radermacher, M. (1988). Three-dimensional reconstruction of single particles from random and nonrandom tilt series. *J. Electron Microsc. Tech*. **9**:359–394.

Saxton, W. O. (1978). *Computer Techniques for Image Processing of Electron Microscopy*. Academic Press, New York.

Saxton, W. O. and Baumeister, W. (1982). The correlation averaging of a regularly arranged bacterial envelope protein. *J. Microsc*. **127**:127–138.

Saxton, W. O., Baumeister, W. and Hahn, M. (1984). Three-dimensional reconstruction of imperfect two-dimensional crystals. *Ultramicroscopy* **13**:57–70.

Sezan, M. I. (1992). An overview of convex projections theory and its application to image recovery problems. *Ultramicroscopy* **40**:55–67.

Sezan, M. I. and Stark, H. (1982). Image restoration by the method of convex projections. II. Applications and numerical results. *IEEE Trans. Med. Imaging* **1**:95–101.

Stewart, A. and Grigorieff, N. (2004). Noise bias in the refinement of structures derived from single particles. *Ultramicroscopy* **102**:67–84.

Unser, M., Sorzano, C. O., Thevenaz, P., Jonic, S., El-Bez, C., De Carlo, S., Conway, J. F. and Trus, B. L. (2005). Spectral signal-to-noise ratio and resolution assessment of 3D reconstructions. *J. Struct. Biol.* **149**:243–255.

Unser, M., Trus, B. L. and Steven, A. C. (1987). A new resolution criterion based on spectral signal-to-noise ratios. *Ultramicroscopy* **23**:39–51.

Vainshtein, B. K. and Penczek, P. A. (2006). Three-dimensional reconstruction. In *International Tables for Crystallography* 3rd edn., vol. B *Reciprocal Space* (U. Shmueli, ed.).
van Heel, M. (1987). Similarity measures between images. *Ultramicroscopy* **21**:95–100.
Wade, R. H. (1992). A brief look at imaging and contrast transfer. *Ultramicroscopy* **46**:145-156.
Youla, D. C. and Webb, H. (1982). Image restoration by the method of convex projections. 1. Theory. *IEEE Trans. Med. Imaging* **1**:81–94.

11

Denoising of Electron Tomograms

Reiner Hegerl and Achilleas S. Frangakis

1. INTRODUCTION

The crucial problem inherent to electron tomography is radiation damage or, related to this, the choice of the correct electron dose: an excessive dose destroys the specimen, especially biological ones, while an insufficient dose results in images that are noisy and lack information. Sophisticated and highly automated techniques have been developed both for data acquisition with the aim of keeping the electron dose as low as possible, and for image processing, in order to extract reliable information from the

Reiner Hegerl • Max Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany

Achilleas S. Frangakis • EMBL, European Molecular Biology Laboratory, Meyerhofstr. 1, D-69117 Heidelberg, Germany

recorded data. However, the tolerable dose is very small, especially for unstained, frozen-hydrated specimens. As a rule of thumb, 5000 e/nm^2 are tolerable for such specimens. According to the dose fractionation theorem (Hegerl and Hoppe, 1978), the total tolerable dose has to be divided by the number of projection views in order to find the dose allowed for each image of a tilt series. In addition, the low scattering power of biological material results in low-contrast images. For instance, assuming a tilt series of 50 images, a pixel size of 1 nm^2, phase contrast imaging with a contrast of 10%, and considering only the shot noise of the electrons, the signal-to-noise ratio (SNR defined as energy of signal over energy of noice) in the projection images is in the order of 1. An increase in the number of projection images, a decrease of the pixel size and additional noise arising from the image recording system push the SNR below 1. The noise in the raw images propagates into the final tomogram; the resulting noise distribution in the tomogram strongly depends on the type of reconstruction algorithm used, e.g. the choice of weighting function in the case of weighted back-projection. Due to the incompleteness of information typical for tomography of the cell, the SNR of the final density map may be of the same order of magnitude as that of the raw images, at least when no additional filtering is used.

The low SNR creates severe problems for the visualization and interpretation of the 3D density maps reconstructed from tilt series. Typically, the maps are inspected first by displaying slices through the reconstruction. Even though our brain is very skillful in recognizing 2D structures in a noisy environment, we have difficulties recognizing small, arbitrarily sliced, 3D structures. Iso-surface representation and volume rendering are powerful tools to represent 3D structures; however, the interpretation of the output is frequently impossible without further processing of the map. Quantitative analysis of the reconstruction, e.g. by automatic segmentation (see Chapter 12 of this volume), and subsequent measurement of structural parameters is an even more demanding task and requires powerful tools for noise reduction.

There is a growing interest in developing new techniques for noise reduction. Looking around at other fields concerned with image analysis, for instance medical imaging or astronomy, we see much activity. Studying the literature is inspiring and helps in developing new ideas or adapting existing approaches to the needs of electron tomography, but it also has to be pointed out that most existing techniques require data with a SNR much higher than 1, and therefore are of no use for electron tomography. Furthermore, most of the presently available techniques have been developed for 2D images and need to be extended to 3D.

For structure analysis based on electron microscopy with a resolution in the nanometer range, noise reduction is absolutely essential. The impressive results obtained by electron crystallography (Fujiyoshi, 1998) as well as by single-particle analysis (Frank, 2002) are attributed to noise reduction by averaging. In both cases, the signal is highly redundant and averaging is

either an intrinsic part of the measurement or it is achieved by appropriate data processing. Assuming the noise to be additive, averaging is the optimum technique because it is—in principle—free from artifacts. Occasionally, redundant signals are also encountered in electron tomography, e.g. by direct imaging of single molecules in suspension, by imaging cells which contain multiple copies of a stable molecular complex or by imaging quasi-periodic structures such as some types of cell wall. In the first two cases, 3D extensions of the methods for single-particle averaging can be applied, whereas for the third case, modifications of the correlation averaging technique might help. Signals originating from whole cells or cell organelles, however, are unique and cannot be averaged.

Imaging of unique, non-repetitive structures is in fact the main application field of electron tomography. In those cases where averaging is not feasible, classical filter techniques are frequently applied to improve the visibility of the reconstructed density maps: real-space filters, such as Gaussian or median, or filters operating in reciprocal space producing a characteristic dampening of high frequencies. All of these filters smooth the data, but they smooth the signal along with the noise. With such filters, some improvement can be achieved as long as the noise has more energy in the high-frequency domain than the signal. Advanced filters should differentiate in a more elaborate way between noise and signal; however, the true signal is usually unknown. Several attempts have been made to develop appropriate methods, using very general properties of signals. In contrast to noise, features of signals, such as for instance boundaries of cellular structures, are extended and may also be characterized by their orientation. Special operations or transformations make use of relationships or affinities among pixels that exist in signals but not in noise. The examples discussed in Section 3 are filtering of coefficients obtained by wavelet transformation (WT), non-linear anisotropic diffusion and the bilinear filter. The methods mentioned so far are designed for post-processing, which means that they are applied *after* the 3D reconstruction. On the other hand, several reconstruction techniques have been proposed in which denoising is an integral part of the reconstruction procedure itself. Section 3.4 presents a brief discussion of such techniques, e.g. 3D reconstruction with minimum energy (minimum variance) or maximum entropy. Practical realizations of such techniques, however, are frequently achieved as a refinement procedure applied to an initial reconstruction, and can therefore also be considered as post-processing methods.

2. CHARACTERIZATION OF NOISE AND SIGNAL

2.1. Noise

In electron microscopy, the answer to the question ‘what is noise?’ depends on the experiment. Two noise sources are permanently present:

the quantum noise of the electron beam and the noise originating from the image recording system. In single-particle averaging, the background fluctuations caused by the specimen support are also considered noise. This is in contrast to electron tomography, where both the object under scrutiny and the specimen support or the surrounding ice matrix are signals, which have to be jointly reconstructed.

The quantum noise originates from random fluctuations of the number n_i of electrons that hit the pixel with index i. When considered as a random variable, this number is subject to a Poisson distribution specified by expectation and covariance. The statistical properties of these variables, with accuracy up to the second order, are:

$$\mathrm{E}\{n_i\} = v_i, \quad \mathrm{Cov}\{n_i, n_j\} = \delta_{ij} v_i \quad \text{with } \delta_{ij} = \begin{cases} 1 & for \quad i = j \\ 0 & for \quad i \neq j \end{cases} \tag{1}$$

v_i is the expected signal at pixel i, and the whole set $\{v_i\}$ describes a hypothetical image that has been recorded without noise. Following the covariance term, the noise is characterized by two statements: its variance $\mathrm{Var}\{n_i\} = v_i$ depends on the signal, and the noise at pixel i is statistically independent of the noise at all other pixels. If we regard phase contrast (see Chapter 3 of this volume) as the prevailing imaging mode in electron tomography, the signal n_i is mainly composed of two terms, the primary beam and a weak scattering wave. The signal I_i of interest, produced by the object structure, is then given by a usually small deviation from the mean:

$$I_i = n_i - \bar{n} \quad \text{with} \quad \bar{n} = \frac{1}{N} \sum_{i=1}^{N} n_i \tag{2}$$

After some calculations

$$\mathrm{E}\{I_i\} = v_i - \bar{v}, \quad \mathrm{Cov}\{I_i, I_j\} = \delta_{ij} v_i + \frac{1}{N}(\bar{v} - v_i - v_j) \approx \delta_{ij} \bar{v} \quad \text{with} \quad \bar{v} = \frac{1}{N} \sum_{j=1}^{N} v_i \tag{3}$$

The covariance term can be approximated by $\delta_{ij}\bar{v}$ due to the fact that the number N of pixels is typically large, and—assuming a thin and weakly scattering specimen—due to small deviations of all v_i with respect to the mean $\bar{v}$. In other words, the noise level is determined mainly by the intensity of the primary beam and, therefore, the concept of signal independence and additivity of noise is an acceptable approximation.

The image recording system creates additional noise and induces correlation between the pixels. Film or CCD cameras are usually characterized by a modulation transfer function (MTF) and a detection quantum efficiency (DQE). The first quantifies the broadening of a point signal and the coupling between neighboring pixels; the second describes the enhancement of the noise level. For a given electron microscope, it is easy to get an

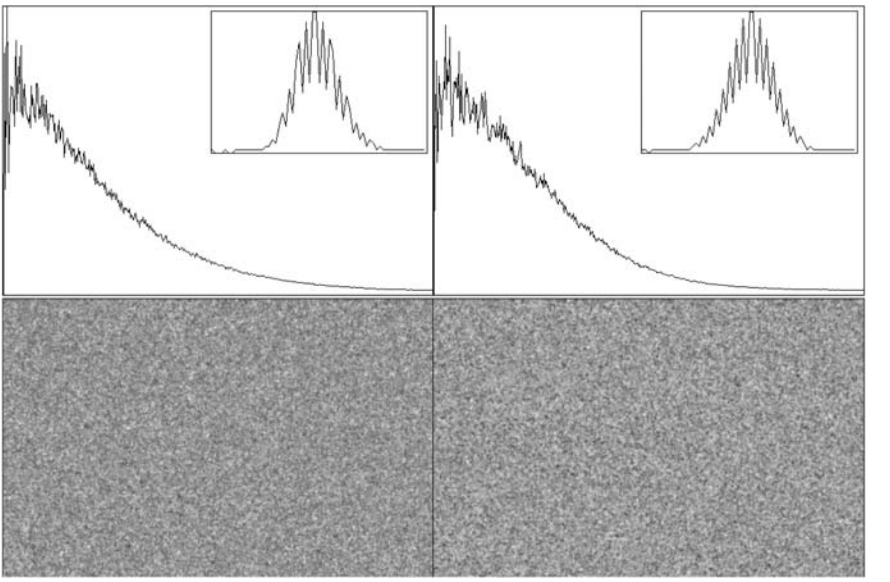


FIGURE 1. Real (left hand) and simulated (right hand) noise. Upper row: rotationally averaged power spectra with histograms of the image as an insert. Lower row: image data.

idea how quantum noise and detector noise add up by recording an image without a sample. The image shows pure noise and can be used to calculate a power spectrum, as shown in Fig. 1. There is a strong dampening of high frequencies corresponding to the correlation between pixels. The curve can be approximated by superposition of a constant part with a Gaussian function. When test calculations are made to evaluate the efficiency of denoising methods, such a model should be used for a more realistic simulation of noise.

The propagation of noise from the raw images into the 3D reconstruction is a complicated process. Assuming a linear reconstruction approach and signal-independent noise in the projection images, the noise in the tomogram is also independent from the reconstructed signal. Its spatial distribution strongly depends on the experimental conditions. In the case of an ideal single-axis tilt series, the variance of noise is constant within the volume of isotropic resolution defined by the Crowther condition.

2.2. Signal

For the interpretation of electron tomograms, it is usually assumed that the signal is obtained by a linear mapping of the object's mass density. The signal, however, is distorted due to the complex process of image formation

in the electron microscope (for details, see Chapter 3 of this volume) and due to imperfections of the 3D reconstruction, the latter arising from missing data (the 'missing wedge' problem) and imperfect alignment of the projection images. It might happen that signal distortions in the high-frequency domain cannot be differentiated from the noise. Missing wedge artifacts may also modify the signal in a lower frequency range such that the effect of noise filters is influenced, as mentioned below in Section 3.2. Typically for ice-embedded biological specimens, the signal is extremely weak and, therefore, noise filters are needed particularly for such specimens.

3. NOISE FILTERS

Filtering an image means that the value of each pixel is replaced by a new value, which is calculated from a combination of the current value of the pixel with those of its neighbors. The calculation scheme, usually represented by a filter kernel, has to be chosen according to the desired filter response; i.e., the task the filter has to achieve. Filters intended for noise reduction should remove random fluctuations superimposed on the signal. A simple approach is to replace each pixel value by the average of its near neighbors and itself. The Gaussian filter is another frequently used example. Both filters remove high-frequency noise; however, they also smooth edges. These filters are easy to apply because they belong to a class of filters with some favorable properties: the new pixel value is linearly related to the old pixel values with weights that do not depend on these values. Since the weights do not change when moving from one processed pixel to the next one, the kernel can be represented as a small image composed of these weights, and the filter operation can be realized as a convolution of the whole image with this kernel. In contrast to such linear filters, the median filter is an example of a non-linear filter, sometimes also classified as a neighborhood-ranking filter versus the above-mentioned neighborhood-averaging filters (Russ, 1995). The median filter operates by analyzing the statistical distribution of neighboring pixel values such that the most probable value of this distribution is taken for the new pixel value. This filter tends to sharpen edges; however, this happens mainly in images where individual pixels are heavily corrupted or missing ('salt-and-pepper noise'), a type of noise not characteristic for electron tomography.

Advanced noise filters should be designed such that they smooth the noise while still preserving edges. Therefore, a more elaborate analysis of the pixel neighborhood is required. Proximity relationships have to be combined with an adequate investigation of pixel values around the currently processed pixel. In the ideal case, for each pixel, noise is discriminated from the signal on the basis of the fact that the signal contains features spread over neighboring pixels whereas noise does not. Clearly, any approach can only be an approximation to this ideal.

In the context of electron tomography, three methods have thus far been proposed and realized for advanced noise filtering: wavelet filtering, non-linear anisotropic diffusion and the bilateral filter. All three are non-linear operations, but each of them is based on different approaches for the discrimination between signal and noise. In the following paragraphs, these methods will be presented and discussed, especially with respect to their specific ways of signal detection.

3.1. Wavelet Filtering

Frequently, the extraction of information from measured data is facilitated by an appropriate transformation of the data. Fourier transformation is the ideal technique to extract periodic signals. The data are decomposed into waves, and those with wavelengths corresponding to the periods of the components of the signal have strong amplitudes. The extraction of non-periodic signals from the background by Fourier transform is less efficient mainly because waves are infinitely extended, in contrast to periodic signals that are bounded and localized. This problem has motivated the development of the WT. With this transformation, the data are tested for similarity with a wavelet basis function modified for different scales and locations. Mathematically, the operation is realized by a convolution of the data $I(x)$ with a set of functions created from a basis function $h(x)$ by translations ξ and dilations a:

$$w(\zeta;a) = \frac{1}{\sqrt{a}} \int I(x) h\left(\frac{x-\xi}{a} \right) dx \tag{4}$$

The function $I(x)$ can be reconstructed exactly from the coefficients $w(\xi;a)$, provided that $h(x)$ and its Fourier transform fulfill some general conditions that have been discussed elsewhere (Daubechies, 1992). It is important that large coefficients indicate high similarity between data and the wavelet function for specified positions and scales, or, in other words, they indicate the existence of a structured and localized signal, while noise results in small coefficients. Therefore, WT and a suitable suppression of the small coefficients followed by the inverse WT offer a way for noise reduction.

It has been shown that relatively simple basis functions can be used to detect signals fairly independently of their specific structure (Daubechies, 1992). For practical realizations, it is also important that the highly redundant set of coefficients $w(\xi;a)$ can be reduced to a discrete subset by constructing an orthonormal set of functions from the basis function $h(x)$:

$$h_{mk}(x) = h(2^m x - k\Delta x) \; (m = 0,1,2, \ldots ; k = \pm 1, \pm 2, \ldots) \tag{5}$$

The continuous scale parameter a is now replaced by a 'pyramid' of dilations discretely scaled by powers of 2 ($a = 2^{-m}$) and the translation ξ is sampled by multiples of a basic distance Δx. The transformation is not only invertible but also complete in the sense that the number of coefficients

obtained by the WT is the same as the number of samples representing the original data. This type of WT in conjunction with appropriate filtering of coefficients has been applied successfully in other fields, e.g., in data compression and in the interpretation of electrocardiograms. For noise reduction, a special coefficient filtering, the 'soft thresholding' or 'coefficient shrinking', was proposed by Donoho in 1995:

$$y = \begin{cases} \text{sign}(x)(|x| - \lambda) & \text{for} \quad |x| \geq \lambda \\ 0 & \text{for} \quad |x| < \lambda \end{cases} \tag{6}$$

where x and y denote the values of WT coefficients before and after filtering, respectively. This type of filtering is superior to hard thresholding ($y = x$ for $|x| \geq \lambda$ for and $y = 0$ elsewhere) with respect to error minimization and smoothness of the filtered signal.

Multidimensional applications can be achieved by combining the 1D functions as a tensor product, as shown by Mallat (1989) and Daubechies (1992). Unfortunately, with this type of transformation, the magnitude of the coefficients depends on the orientation and position of the structural feature. Yu *et al*. (1996) proposed a brute-force approach to overcome this anisotropy of signal recognition, using the following strategy:

Step 1: from the input image, create a set of images by rotation and translation.
Step 2: subject each member of the set to WT, apply coefficient thresholding and compute inverse WT.
Step 3: invert rotation and translation of each member.
Step 4: average the set, which results in the output image.

This procedure has been realized for 2D images and applied to the slices of electron tomograms by Stoschek and Hegerl (1997). The results obtained with this pseudo-3D denoising method are evidently superior to those obtained with conventional filters. However, the method is very demanding in terms of computational power and, as discussed in the original article, may occasionally create artifacts. It may also happen that linear structures extending perpendicularly to the slices are not correctly recognized. Nevertheless, Fig. 2 shows a successful application of this approach to an electron tomographic reconstruction of a vesicle with actin filaments embedded in vitreous ice. It also demonstrates that rotation and translation invariance are essential steps in this method.

Recently, a 3D wavelet-based filter was proposed (Moss *et al*., 2005) and used to visualize structures in electron tomograms. Instead of transforming the data by the discrete WT given in equation (5), the data are multiply convoluted with a 1D basis function in varying directions but with a fixed scale parameter a. No global noise filtering can be achieved with this filter; however, with an appropriately chosen scale parameter, structures of corresponding size are highlighted against the environment.

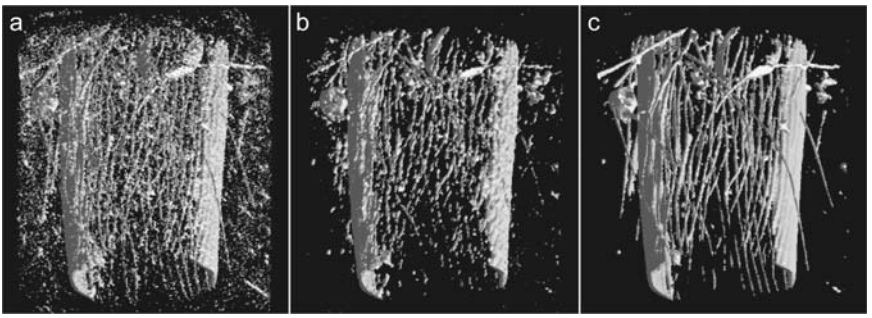


FIGURE 2. Pseudo-3D wavelet denoising applied to an electron tomographic reconstruction of a lipid vesicle with actin filaments embedded in vitreous ice. Isosurface representation of (a) original data, (b) map denoised using orthogonal WT and (c) map denoised using translation- and rotation-invariant WT. For (b) and (c), Symmlet 8 wavelet kernel and soft thresholding was applied. (From Stoschek *et al.* (1997), reproduced with permission of IEEE Computer Society Press).

3.2. Non-linear Anisotropic Diffusion (NAD)

Diffusion is a process well known in physics. Its effect is to equilibrate an inhomogeneous distribution of heat or matter steadily in time, e.g. the concentration of salt in water. Analogously, an image can be considered as an inhomogeneous distribution of gray values and, by simulating the diffusion process in the computer, the gray values can be modified toward a more homogenous distribution. Obviously, the image will be smoothed and, thereby, noise will be removed. With linear diffusion, the result is identical to Gaussian filtering, as will be shown below. The basic idea of *nonlinear* anisotropic diffusion is to control the magnitude and direction of the diffusion flux dependent on the underlying image structure. The diffusion flux across edges should be low in order to keep the edges preserved, while areas free of edges should be smoothed by a large diffusion flux.

Mathematically, the diffusion process can be modeled by a partial differential equation. Let $I(\mathbf{x},t)$ be the gray values of an image evolving in time t from the starting image $I_0(\mathbf{x}) = I(\mathbf{x},0)$, with pixels at position $\mathbf{x}$ in a space of arbitrary dimensions. The gradient ∇I causes a flux $\mathbf{j}$,

$$\mathbf{j} = -D\nabla I(\mathbf{x},t) \tag{7}$$

which aims to compensate for this gradient according to the choice of the diffusivity D. This relationship is known as the first Fick's law. The diffusion equation describes the change of intensity or gray value at position $\mathbf{x}$ by balancing the in- and outgoing flux:

$$\frac{\partial}{\partial t} I(\mathbf{x},t) = -\mathrm{div}\mathbf{j} = \mathrm{div}(D\nabla I) \tag{8}$$

Usually D is considered to be a constant parameter c, such that the equation becomes linear in all derivatives of I:

$$I_t = c\Delta \mathrm{I} \tag{9}$$

The solution can be found as a classical example in textbooks on partial differential equations: the intensity distribution $I(\mathbf{x},t)$ at time t results from the starting image $I_0(\mathbf{x})$ by a convolution of the latter with a Gaussian function using a variance of $2ct$. In other words: *linear isotropic diffusion is equivalent to conventional Gaussian filtering*. The diffusion flux has the direction of the gradient and, consequently, edges are smoothed due to the large gradient in the direction perpendicular to the edge.

A fundamental step towards better signal preservation was made by Perona and Malik in 1990 by substituting the function

$$g\left(|\nabla I|^2\right) = \frac{1}{1+|\nabla I|^2/\kappa^2} \tag{10}$$

for the diffusivity D. This makes equation (8) non-linear and the magnitude of the local flux is modified by a diffusivity that depends on the underlying image structure. In particular, the flux across edges is reduced due to the smaller diffusivity produced by the large gradient. Generally, the process results in intraregional smoothing between well-preserved edges. Unfortunately, the now reduced flux across edges does not remove the noise in the neighborhood of edges. In order to remove this remaining noise, a flux parallel to edges or perpendicular to the image gradient would be needed.

For a general control of magnitude and direction of the diffusion flux, e.g. to enable additional flux parallel to edges, an appropriately chosen tensor $\mathbf{D}$ has to be inserted for the diffusivity in equation (7), with the consequence that the diffusion process (8) becomes both non-linear and anisotropic. Weickert (1998) proposed to base this tensor on the so-called structure tensor $\mathbf{J}_\sigma$, which is the dyadic product of the gradient, ∇I. ∇I^{T}, convoluted with a Gaussian of variance σ. The structure tensor $\mathbf{J}_\sigma$ can be calculated for each voxel of the map. Its eigenvalues μ_i and eigenvectors $\mathbf{v}_i$ characterize the local structural features of an image within a neighborhood of size σ. Depending on these features, the local diffusion flux can now be decomposed into components with directions corresponding to those of the eigenvectors $\mathbf{v}_i$ with freely choosable magnitudes λ_i. For 3D images (i = 1,2,3) such as electron tomograms, the diffusivity can then be written as:

$$\mathbf{D} = [\mathbf{v}_1\mathbf{v}_2\mathbf{v}_3]\begin{bmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{bmatrix}[\mathbf{v}_1\mathbf{v}_2\mathbf{v}_3]^{\mathrm{T}} \tag{11}$$

There are various ways to define the parameters λ_i as functions of the local structure represented by the eigenvalues μ_i. It is generally known that eigenvectors and eigenvalues indicate the direction and magnitude of

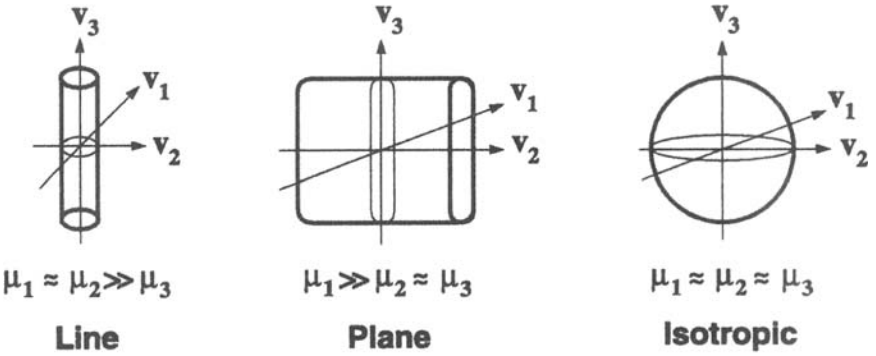


FIGURE 3. Classification of voxels according to the eigenvalues and eigenvectors of the structure tensor. (From Fernandez and Li (2003), reproduced with permission of Elsevier).

extreme variations, $\mathbf{v}_1$ and μ_1 being the largest, and $\mathbf{v}_3$ and μ_3 the smallest (see Fig. 3). This property can be used to classify voxels: for instance, the condition $\mu_1 \approx \mu_2 >> \mu_3$ assigns the voxel to a line parallel to $\mathbf{v}_3$, and $\mu_1 >> \mu_2 \approx \mu_3$ to a plane perpendicular to $\mathbf{v}_1$. In the first case, it makes sense to smooth mainly along $\mathbf{v}_3$, and in the second along $\mathbf{v}_2$ and $\mathbf{v}_3$. For 2D images, Weickert proposed two basic concepts, which he called 'edge-enhancing diffusion' (EED) and 'coherence-enhancing diffusion' (CED). With EED, the flux along the first eigenvector is a monotonically decreasing function of the first eigenvalue, while the strength along the other eigenvector is kept constant. The aim is to reduce the flux across edges and to smooth preferentially in regions bounded by these edges. The filter can be regarded as an anisotropic regularization of the Perona–Malik model. With CED, the flux is made small and constant along the first eigenvector, while the flux along the second eigenvector increases with the difference between the largest and smallest eigenvalues. This filter amplifies curvilinear structures and is an efficient tool to improve images of fingerprints.

For the denoising of electron tomograms, Frangakis and Hegerl (2001) proposed to combine the advantages of EED and CED in a 3D hybrid approach. The basic idea is to use the difference $\mu_1 - \mu_3$ as a switch between EED and CED. If this difference is small, the noise level is still high, and EED is applied to reduce the noise between edges. On the other hand, a large difference indicates the existence of a local curvilinear structure, to be enhanced by CED. The setting of λ_i is as follows:

For the EED part, the unsmoothed version $\mathbf{J}_0$ of the structure tensor is used, which has the eigenvalues $\mu_1 = |\nabla I|^2$, $\mu_1 = \mu_3 = 0$, and the eigenvector $\mathbf{v}_1 = \nabla I$. Then

$$\lambda_1 = \lambda_1 = g(\mu_1), \lambda_3 = 1 \tag{12}$$

with a suitable function g, e.g. as in equation (10) or, alternatively, with

$$g\left(|\nabla I|^2\right) = 1 - \exp\left(\frac{-C_m}{\left(|\nabla I|^2 / \kappa^2\right)^m}\right)$$

Following Weickert (1998), a choice of $m = 4$ and $C_m = 3.315$ gives visually good results.

CED is based on the smoothed structure tensor $\mathbf{J}_\sigma$ and on the choice

$$\lambda_1 = \lambda_1 = \alpha, \quad \lambda_3 = \alpha + (1-\alpha)\exp\left(\frac{-C}{(\mu_1 - \mu_3)^2}\right) \tag{13}$$

With $\alpha = 1$, CED is identical to a Gaussian filter. Figure 4 shows an example where this hybrid approach is applied to the tomographic reconstruction of a chromatin fragment.

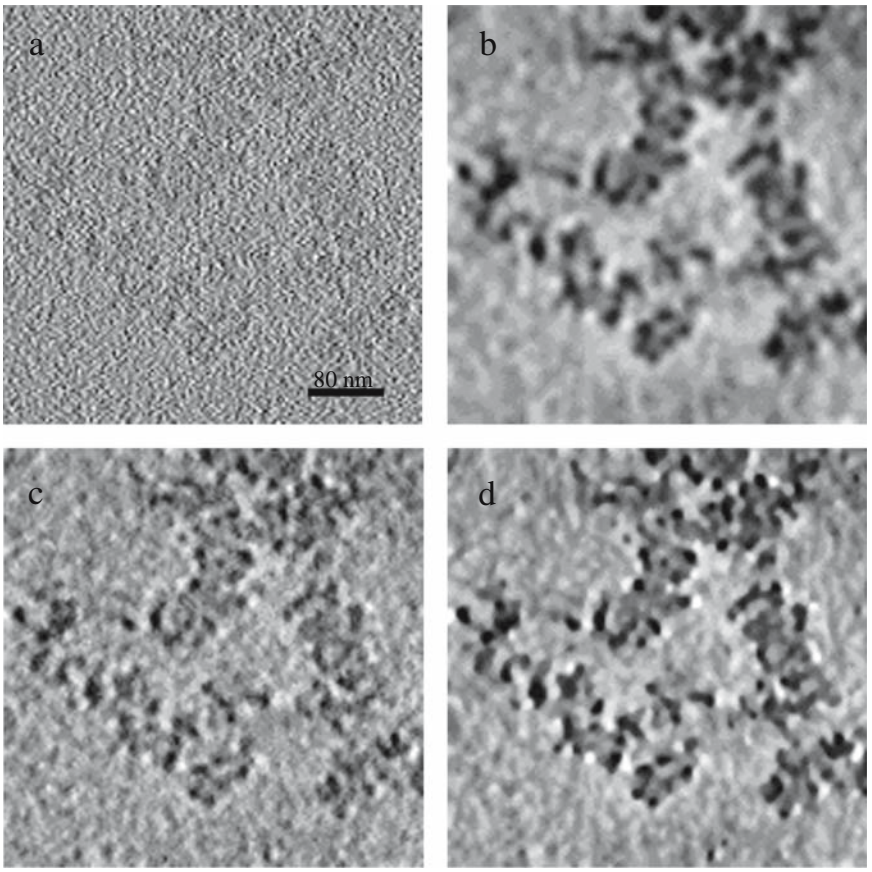


FIGURE 4. Horizontal slice through an electron tomographic reconstruction of a chromatin fragment. (a) Original (b–d) denoised using (b) Gaussian filtering, (c) median filtering and (d) the hybrid approach of NAD. (From Frangakis and Hegerl (2001), reproduced with permission of Elsevier).

Fernandez and Li (2003) used this hybrid strategy with an extended exploitation of the information given by the structure tensor. First they define parameters

$$P_1 = \frac{\mu_1 - \mu_2}{\mu_1}, \quad P_2 = \frac{\mu_2 - \mu_3}{\mu_1}, \quad P_3 = \frac{\mu_3}{\mu_1}$$

and then classify voxels according to the scheme (see Fig. 3):

$$\begin{aligned} &P_1 > P_2 \text{ and } P_1 > P_3 \Rightarrow \text{plane-like} \\ &P_2 > P_1 \text{ and } P_2 > P_3 \Rightarrow \text{line-like} \\ &P_3 > P_1 \text{ and } P_3 > P_2 \Rightarrow \text{plane-like} \end{aligned}$$

The switch between EED and CED as well as the setting of λ_i for EED are used in the same manner as described above. In the case of CED, the setting is more differentiated with respect to the classification of the voxel. For voxels that are part of a curvilinear structure, the setting of equation (13) is used, but for voxels of a plane-like structure, the magnitudes are chosen as follows:

$$\lambda_1 = \alpha, \quad \lambda_2 = \alpha + (1-\alpha)\exp\left(\frac{-C_2}{(\mu_1 - \mu_2)^2}\right), \quad \lambda_3 = \alpha + (1-\alpha)\exp\left(\frac{-C_3}{(\mu_1 - \mu_3)^2}\right) \tag{14}$$

Frangakis and Hegerl (2002) argued against an extensive use of this structural parameter because of a frequently observed strong noise sensitivity of the second eigenvalue μ_2. One should also be aware of the fact that the eigenvectors may be affected by the orientation of the missing wedge: according to test calculations (Fernandez and Li, 2003), there is a tendency of the first and third eigenvector to become perpendicular and parallel to the electron beam, respectively.

The implementation of NAD follows the technique for iterative solution of partial differential equations. Denoting D_{ik} with $D_{ik} = D_{ki}$ for the elements of the diffusivity tensor in equation (11), a discretized version of the following equation has to be solved:

$$\begin{aligned} \partial_t I = {} & \partial_x(D_{11}\partial_x I) + \partial_x(D_{12}\partial_y I) + \partial_x(D_{13}\partial_z I) \\ & + \partial_y(D_{21}\partial_x I) + \partial_y(D_{22}\partial_y I) + \partial_y(D_{23}\partial_z I) \\ & + \partial_z(D_{31}\partial_x I) + \partial_z(D_{32}\partial_y I) + \partial_z(D_{33}\partial_z I) \end{aligned}$$

3.3. The Bilateral Filter

In contrast to conventional neighborhood-averaging filters, the bilateral filter combines the influence of neighboring pixels to the filter output with respect to both the spatial domain and the intensity range (Tomasi and Manduchi, 1998). The common way to accomplish a neighborhood-averaging filter is to replace each pixel with the average of itself and its appropriately weighted neighbors. The equation

$$\hat{I}_i = \frac{\sum_j w_{ij} I_j}{\sum_j w_{ij}} \tag{15}$$

is a multidimensional representation of such a filter where the indices i or j run over all N pixels in a lexicographic manner ($i,j = 1, \ldots, N$). I_i and $\hat{I}_i$ denote the respective densities of pixel i in the unfiltered and filtered images, and w_{ij} is the weight of pixel j with respect to pixel i. The weights decrease with increasing distance between pixel i and j. Therefore, in practice, there is a maximum distance (due to the disappearance of weights beyond this distance) which limits the sum over j in equation (15), so that the filter can be realized as a kernel operation. The denominator is optional; its purpose is to preserve the local mean. One of the most popular examples of this filter type is the Gaussian filter with weights defined as

$$w_{ij} = \exp\left(-\frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2\sigma^2} \right) \tag{16}$$

where $\mathbf{x}_i$ denotes the position of pixel i. The user has to set the parameter σ, which determines the range of smoothing.

The weights of the bilateral filter are obtained by multiplying the Gaussian weights (16) with a further factor such that the underlying image structure can contribute. This factor is of similar construction, but the distance term in the exponent of equation (16) is replaced by the squared difference of density values of the corresponding pixel pair:

$$w_{ij} = \exp\left(-\frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2\sigma_D^2} \right) \exp\left(-\frac{(I_i - I_j)^2}{2\sigma_I^2} \right) \tag{17}$$

Two parameters must now to be set, σ_D for the smoothing distance as before, and an additional σ_I for an appropriate intensity range. With a large σ_I, the filter is reduced to a normal Gaussian filter. To understand the effect of this non-linear filter, let us assume that pixel i is part of a line which is characterized by a set of pixels having, ideally, the same density values. Neighboring pixels j that are also part of this line have similar densities and, therefore, the second factor in the weight is close to 1. Pixels with the same distance from pixel i, but away from the line, contribute with a smaller weight according to the larger difference $I_i - I_j$. Consequently, mainly pixels along the line contribute to pixel i, hence the line is better preserved when compared with Gaussian filtering.

It is worth mentioning that the weights w_{ij} defined in equation (17)—sometimes called 'affinity weights'—are also used in a method for image segmentation described in Chapter 12 of this volume. In this context, the weights are arranged as a matrix, the so-called affinity matrix. By analyzing eigenvectors of this matrix, a powerful foreground versus background

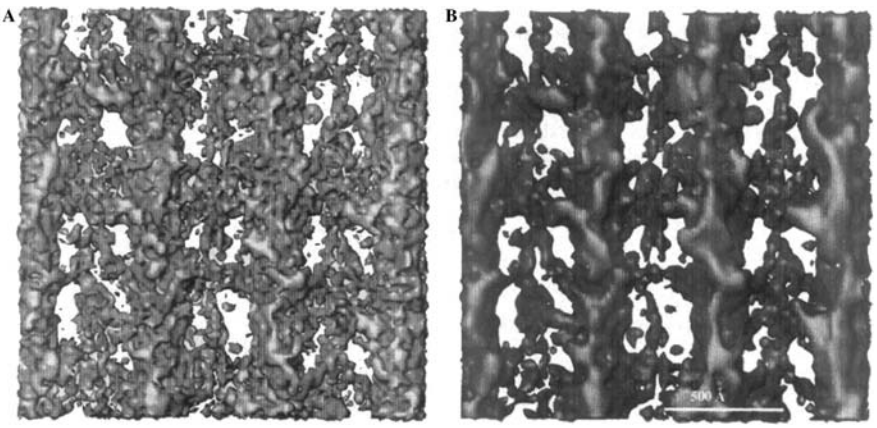


FIGURE 5. Bilateral filtering of an electron tomogram. (A) Surface view of the original density map of rigor flight muscle thin section. The map density variance is 48. (B) The bilateral filtered density map (A) with $\sigma_D = 1$ pixel and $\sigma_D = 200$ (by courtesy of Jiang *et al.*). (From Jiang *et al.* (2003), reproduced with permission of Elsevier).

separation can be achieved, e.g. the separation of a cell from the surrounding ice matrix.

The bilateral filter is easy to realize; however, some effort is required to make its application fast. The filter can be implemented as a normal kernel operation equivalent to the Gaussian filter but, in contrast to the latter, the operation is slow because the kernel has to be recalculated for each pixel. The non-linear behavior of the filter also prevents replacement of the kernel operation by the fast and elegant combination of fast Fourier transform combined with a simple multiplication in Fourier space as used for the Gaussian filter in the case of a larger kernel. Fortunately, however, the operation is well suited for parallel processing. Figure 5 shows an application of the bilateral filter to the tomographic reconstruction of a flight muscle (from Jiang *et al.*, 2003).

3.4. Constrained or Regularized Reconstruction

The denoising techniques described so far are post-processing methods, which means that they are applied to the final tomogram as obtained for instance by weighted back-projection (see Chapter 8 of this volume) or by the algebraic reconstruction technique (ART) (see Chapter 7 of this volume). There also exist some methods where noise reduction is an intrinsic part of the 3D reconstruction. Such methods subject the data to constraints derived from *a priori* information or from additional measurements. Typical examples for *a priori* information are positivity—density values are always positive (appropriate correction of contrast transfer function required)—or reconstructions in conjunction with additional principles

such as energy minimization or entropy maximization (Barth *et al*., 1988; Skoglund *et al*., 1996). Constraints from experimental data can be realized, for instance, when something is known about the shape of the specimen, which might be its thickness or surface relief. The corresponding algorithms are usually formulated as an optimization task to be solved iteratively (see Chapter 2 in this book). The effect of noise filtering is obvious in the case of minimum energy or maximum entropy because these constraints have the effect of smoothing the data. The problem is again to find the right balance between smoothness and fidelity to the measured data. Unfortunately, nothing is known about studies concerned specifically with noise reduction properties of such methods.

4. EFFICIENCY, OR 'WHAT IS GOOD DENOISING'?

The aim of all the methods we are considering is to reduce the noise as much as possible with full preservation of the structural information. However, what does this mean in practice, and how can the result be evaluated? The answer obviously depends on what is intended with denoising. Is it intended for improving the visualization of a map, or as a preprocessing step for subsequent segmentation? For the latter application, more smoothing might be useful, while less smoothing is preferable for the former. It is also important to consider the size of the structural features under scrutiny. The parameters controlling the noise filtration have to be chosen appropriately. Test calculations help to understand how the methods work and to learn how the parameters should be set. Examples can be found in the references cited in Sections 3.1–3.3. In the following treatment, test calculations are presented, which compare anisotropic diffusion with the bilateral filter, and finally some quality criteria are considered.

Before going into details of test calculations and quality criteria, it might be useful to compare the methods with respect to their ease of implementation and handling.

1. Wavelet filtering: a program for discrete WT is required for the implementation of wavelet filtering. Such transformations are available in modern program packages for image processing. An additional but not too complicated routine for rotating the complete area of an image is needed to realize rotation invariance. For the application, only one threshold parameter has to be set, at least for the simplest version of wavelet filtering. As already mentioned, the procedure is not yet realized as a real 3D filter. It is also very demanding in terms of computational power, but parallel processing of steps 1, 2 and 3 is straightforward.

2. Non-linear anisotropic diffusion: NAD is usually realized as an iterative numerical solution of a partial differential equation. This requires some programming; however, some versions are freely available (Frangakis and Hegerl, 2001; Fernandez and Li, 2003; see also IMOD

http://bio3d.colorado.edu/imod/doc/program_listingprogram_listing; and SPIDER http://www.wadsworth.org/spider_doc). Of the three denoising methods discussed, NAD is certainly the most flexible one, but it is difficult to handle in its hybrid form (see Section 3.2) due to the large number of parameters that must be set. When using the simpler edge-enhancing method, only two parameters must be defined.

3. Bilateral filter: this filter is conceptually the simplest method, and it can be realized as a kernel operation corresponding to a Gaussian filter in real space. Two parameters have to be set.

It should be mentioned that even though the three methods lend themselves to parallel computing to different degrees, parallel computing is nevertheless always possible for all three methods by dividing the volume into appropriately overlapping subvolumes (i.e. taking account of kernel width as necessary) and performing separate processing of each.

4.1. Test Calculations

The great advantage of test calculations with phantoms is that they offer the possibility of comparing the noise-filtered map quantitatively with the original structure model or with the corresponding noise-free tomographic reconstruction. This allows one to study the effects of differently chosen control parameters and to optimize the filtering with respect to special quality criteria. The tests should be performed by a simulation of the full tomographic process in order to include the effects of the 'missing wedge'. Therefore, from the phantom, projection images are produced, noise is added to each, and all projections are merged in a 3D reconstruction as usual. The resulting map can then be subjected to the different denoising methods. To make the simulation as realistic as possible, the spectra of the added noise were modeled as shown in Fig. 1.

Figure 6 summarizes the results of test calculations where edge-enhancing NAD is compared with the bilateral filter. The comparison is fair in that both methods are mainly controlled by two parameters: edge-enhanced NAD with the diffusion strength λ and the number of iteration steps, and the bilateral filter with the variances σ_I and σ_D for intensity and distance. Additionally, a conventional Gaussian filter was applied for comparison. Using data from the Protein Data Bank, models of proteasomes, thermosomes and actin filaments were calculated and randomly distributed in the test volume. The linear extension of a voxel is 0.5 nm, as is the case for each pixel in the projection images. From this volume, 53 projection images were calculated, corresponding to a tilt series with a tilt increment of 2.5° within the angular range from –65° to +65°. Simulating bright field imaging with a mild band-pass and a contrast of 5%, noise was added, corresponding to a dose of 20 or 50 electrons per pixel and per projection. The root mean square deviation (RMSD) of signal and noise was around 1 (2.4) and 4.5 (7), respectively (RMSD values correspond to a dose of 20 elec-

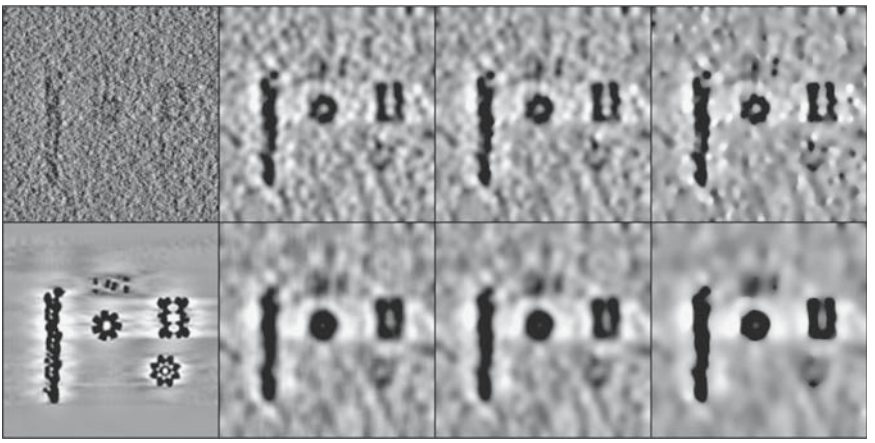


FIGURE 6. Test calculation with synthetic data simulating an electron dose of 20 electrons/pixel. All images show the same slice through a volume, each represented with the intensity range mean $\pm$ 3 $\times$ variance. (a) Noisy reconstruction, (e) noise-free reconstruction. Weak denoising with (b) Gaussian filter, (c) bilateral filter and (d) NAD. Strong denoising with (f) Gaussian filter, (g) bilateral filter and (h) NAD.

trons/pixel; those in parentheses to 50 electrons/pixel). The tomographic reconstruction from these projection images is free from alignment errors but shows typical reconstruction artifacts, mainly due to the missing wedge. Taking the RMSD of the noise-free reconstruction for the signal, the values of 1.1 (2.8) and 8.5 (13.5) characterize signal and noise, respectively

The aim of the test calculations was to find the optimum similarity between the denoised map and the noise-free reconstruction by varying the parameters mentioned above (λ, iteration steps, σ_D and σ_I). Using the cross-correlation coefficient (CCC) as a measure for similarity, two criteria for optimization were applied: (i) the maximum of CCC and (ii) again the maximum of CCC but with the additional condition that the variance of the denoised map remains larger than that of the noise-free map. The results in Fig. 6 show that criterion (i) leads to very strong denoising, with over-smoothing of the data. However, the resulting maps are quite useful since they can be applied—after further processing with appropriate morphological operators (Russ, 1995)—to segment the macromolecular complexes. All maps satisfying criterion (i) have a variance smaller than that of the noise-free reconstruction, which indicates loss of signal. This was the reason for the use of criterion (ii), which sets a lower limit to the variance of the denoised map. The resulting images show weaker denoising, with more residual noise, but also with more residual details of the structures of interest.

Regarding the different filter techniques, NAD shows the sharpest edges, while Gaussian filtering shows the smoothest. In Fig. 7, the Fourier

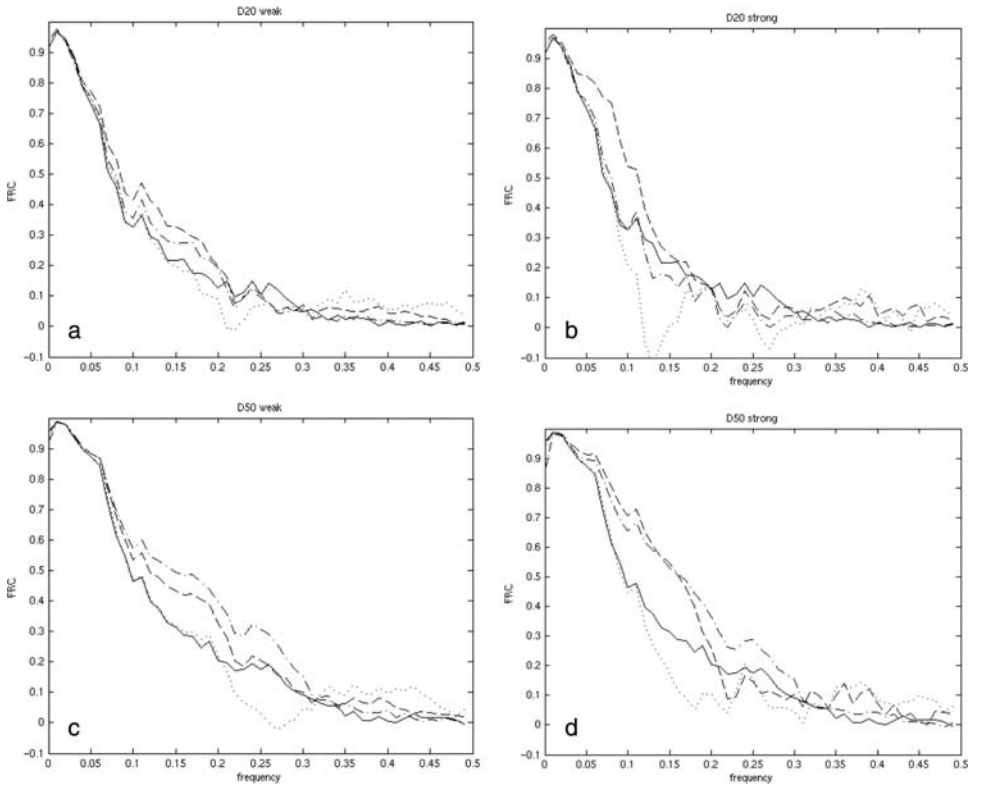


FIGURE 7. FSC between noise-free reconstruction and noisy maps: unfiltered (solid line); Gaussian-filtered (dotted line); bilateral-filtered (dash-dotted line); and NAD (dashed line). (a) 20 electrons/pixel, weak filtering; (b) 20 electrons/pixel, strong filtering; (c) 50 electrons/pixel, weak filtering; (d) 50 electrons/pixel, strong filtering.

shell correlation (FSC) between noise-free reconstruction and denoised maps is displayed to demonstrate the effects of different filter techniques in frequency space. In these examples, NAD gives better results in the low-dose case of 20 electrons/pixel, while the bilateral filter is better when applied to the less noisy tilt series with 50 electrons/pixel. Strong denoising according to criterion (i) shifts the gain in information to lower spatial frequencies.

In a study presented by Frangakis and Hegerl (2001), NAD was applied to redundant data such that the result could be compared with that obtained by averaging. A total of 468 3D maps of individual macromolecules (VAT, a ATPase homolog) obtained by tomographic reconstruction were aligned and averaged. One average image was calculated from the original maps, and another one after processing each individual map with NAD. The FSC curve of the latter was slightly better, whereas the corresponding average appeared more blurred. This seems to be a contradiction;

however, one has to consider that the FSC function measures only the similarity between two subsets of the maps. The conclusion to be drawn is that denoising improves the similarity between the particles but reduces the magnitude of Fourier coefficients in the high-frequency range. Appropriate enhancing of the high-frequency amplitudes of the denoised particle images makes the two averages indistinguishable.

4.2. Quality Criteria

Though important for understanding the theory and for optimizing parameter settings, the results of test calculations should not be overestimated in terms of their implication for real applications. The characters of signal and noise might be different from those used in the simulations due to an oversimplification of the imaging model. As discussed above, it is even difficult to define the meaning of 'good denoising' in test calculations. In applications to real data, the true signal is unknown, and the evaluation of the denoising result has to be based on only two maps, the original and the noise-filtered maps. The most obvious approach to a quality criterion is to compare the variances of both maps. All denoising methods have in common that they reduce the variance, and the problem is to find out by how much they should be allowed to do so. The following rationale appears to be the most reasonable: given the SNR of the original map, the variance should not be reduced below the fraction of variance that corresponds to the signal:

$$\frac{\mathrm{Var}\{\text{denoised image}\}}{\mathrm{Var}\{\text{original image}\}} \geq \frac{SNR}{SNR+1} \tag{18}$$

Though usually unknown, it should be possible to make an estimation of the expected SNR, at least for typical and frequently used experimental set-ups. Occasionally, the map includes a region free from structural features, containing pure noise. The variance taken from this region in denoised and original maps is a direct indicator for noise reduction. However, one should not forget that such regions might also contain spurious signals or clutter from the reconstruction process, e.g. back-projection rays, which contribute to the signal variance.

It is certainly one advantage of NAD that the iterative nature of the procedure also makes it possible to consider intermediate states of the map. For instance, the relative variance mentioned above can be monitored as a function of the iteration steps or, in other words, as a function of the time parameter. Weickert (1999) proposed to stop the iteration when the limit given in equation (18) is reached. Another termination criterion, the 'decorrelation criterion', was proposed by Mrazek and Navara (2003). It is based on the temporal development of a special cross-correlation coeffi-

cient to be calculated from the map or image at time t and at time $t = 0$, and thus does not require an estimation of the SNR. Using the notation of Section 3.2, the exact definition of this coefficient can be written as follows:

$$ccc(I(\mathbf{x},0)-I(\mathbf{x},t),I(\mathbf{x},t))=\frac{\mathrm{Cov}\{I(\mathbf{x},0)-I(\mathbf{x},t),I(\mathbf{x},t)\}}{\sqrt{\mathrm{Var}\{I(\mathbf{x},0)-I(\mathbf{x},t)\}\mathrm{Var}\{I(\mathbf{x},t)\}}} \quad (19)$$

Assuming ideal denoising at a certain time t, the noise-filtered map should be identical to the signal and, after subtracting this map from the original map, pure noise should be obtained. Thus, at time t, no correlation between the difference map and the denoised map could occur. In practice, one could expect that the decorrelation coefficient reaches a minimum, and the time corresponding to this minimum would define the end of the iteration. Unfortunately, a minimum can be observed only occasionally. For a detailed discussion of this criterion, the reader is referred to the literature (Fernandez and Li, 2003; Mrazek and Navara, 2003).

5. CONCLUSION

So far, three basically different methods for advanced noise filtering have been realized in the field of electron tomography. They differ with respect to the way in which the signal is recognized and separated from the noise. WT, analysis of the structure tensor and the calculation of affinity weights are the means, respectively. The last two of these methods, realized as NAD and the bilateral filter, are widely used, as demonstrated by publications concerned directly with applications of electron tomography. NAD is more flexible and can be designed individually for data representing specific structural features. The bilateral filter may be favored due to its simplicity. Also the wavelet-based method offers great power for signal recognition; however, alternative mathematical approaches are needed for real 3D realizations. In the future, other approaches may emerge, or the existing ones may be combined for greater efficiency.

There is also a great need for practical and reliable criteria for measuring the quality of the noise filtering process. Such criteria may be different depending on the final aim of denoising, e.g. the desired smoothness of the data. In any case, they must exploit more than a simple image property such as the variance, but this will depend on future developments.

ACKNOWLEDGEMENTS

The authors would like to thank Joachim Frank for a critical reading of the manuscript.

REFERENCES

Barth, M., Bryan, R. K., Hegerl, R. and Baumeister, W. (1988). Estimation of missing cone data in three-dimensional electron microscopy. *Scanning Microsc. Suppl.* **2**:277–284.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM Publications, Philadelphia.

Donoho, D. L. (1995). De-noising by soft thresholding. *IEEE Trans. Inform. Theory* **41**:613–627.

Fernández, J. J. and Li, S. (2003). An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms. *J. Struct. Biol*. **144**:152–161.

Frangakis, A. and Hegerl, R. (2001). Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. *J. Struct. Biol*. **135**:239–250.

Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **31**:303–319.

Fujiyoshi, Y. (1998). The structural study of membrane proteins by electron crystallography. *Adv. Biophys.* **35**:25–80.

Hegerl, R. and Hoppe, W. (1976). Influence of electron noise on three-dimensional image reconstruction. *Z. Naturforsch*. **31a**:1717–1721.

Jiang, W., Baker, M. L., Wu, Q., Bajaj, C. and Chiu, W. (2003). Applications of a bilateral denoising filter in biological electron microscopy. *J. Struct. Biol*. **144**:114–122.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.* **2**:674–693.

Moss, W. C., Haase, S., Lyle, J. M., Agard, D. A. and Sedat, J. W. (2005). A novel 3D wavelet-based filter for visualizing features in noisy biological data. *J. Microsc*. **219**:43–49.

Mrazek, P. and Navara, M. (2003). Selection of optimal stopping time for nonlinear diffusion filtering. *Int. J. Comput. Vis*. **52**:189–203.

Perona, P. and Malik, J. (1990). Scale-space and edge detection using anistropic diffusion. *IEEE Trans. Pattern Anal. Machine Intell.* **12**:629–639.

Russ, J. C. (1995). *The Image Processing Handbook*. CRC Press, Boca Raton, Florida.

Skoglund, U., Öfverstedt, L. G., Burnett, R. M. and Bricogne, G. (1996). Maximum-entropy three-dimensional reconstruction with deconvolution of the contrast transfer function: a test with application with adenovirus. *J. Struct. Biol.* **117**:173–188.

Stoschek, A. and Hegerl, R. (1997). Denoising of electron tomographic reconstructions using multiscale transformations. *J. Struct. Biol*. **120**:257–265.

Stoschek, A., Yu, T. P. Y., and Hegerl, R. (1997) Denoising of electron tomographic reconstructions biological specimens using multidimensional multiscale transforms. Proc. of ICA SSP97, Munich, IEEE Computer Society Press, Vol. 4, pp. 2793–2796.

Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. In *Proceedings of the IEEE International Conference on Computer Vision*. Bombay, pp. 59–66.

Weickert, J. (1998). *Anisotropic Diffusion in Image Processing*. Teubner, Stuttgart.

Weickert, J. (1999). Coherence-enhancing diffusion of color images. *Image Vis. Comput*. **17**:201–202.

Yu, T. P. Y., Stoschek, A. and Donoho, D. L. (1996). Translation- and direction-invariant denoising of 2D and 3D images: experience and algorithms. *Proc. SPIE* **2825**:608–619.

12

Segmentation of Three-dimensional Electron Tomographic Images

Achilleas S. Frangakis and Reiner Hegerl

Achilleas S. Frangakis • EMBL, European Molecular Biology Laboratory, Meyerhofstr.1, D-69117 Heidelberg, Germany. E-mail: frangak@embl.de

Reiner Hegerl • Max-Planck-Institut für Biochemie, Am Klopferspitz 18, D-82152 Martinsried, Germany. E-mail: hegerl@biochem.mpg.de

1. INTRODUCTION

The intuitive understanding of the process of segmentation is that of a compartmentalization of the image into coherent regions and the extraction of independent objects. Perhaps the most sophisticated segmentation mechanism is human vision, which is capable of interpreting a large variety of groups, associating them into classes and compartments, as well as finding relationships among them. Computer-based image segmentation algorithms typically perform only a single task, which is coupled to a specific application. Humans use a large variety of different criteria to segment images, e.g. similarity, proximity, continuity and symmetry. In electron tomography, the observer usually searches for a known shape or multiply occurring shapes to guide his segmentation. The separation criteria used are the gray value and the contrast between the feature and the environment. In a general sense, the aim is to group pixels or voxels into subsets which correspond to meaningful regions or objects. When regarding pictures by eye, one has an intuitive sense for the boundaries of meaningful objects and regions. When using the computer, however, it is difficult to find quantitative criteria which define meaningful areas on the basis of pixel properties such as contours, brightness, color, texture, etc.

Segmentation algorithms incorporate edge detection and shape recovery, which are essential for many visualization tools. Independently of whether the technique is interactive, semi-automatic or fully automated, the separation of 2D or 3D images into several more coherent regions, the properties of which—as defined by intensity or texture—are roughly uniform, becomes indispensable as the image complexity grows. In electron microscopy, the feature of interest usually corresponds to pixels possessing different intensities compared with the background, at least locally, or those grouped into certain patterns. The goal is to locate these regions within the image and then perform segmentation.

The need for segmentation increases dramatically when the images go from two dimensions to three. Apart from the increased image complexity, which is intrinsic to an increase of the dimensionality, the comprehension of an image is much more difficult, due to the limited visualization capabilities. The perception of a certain object is significantly hampered because it might be covered, shadowed or obscured by noise and other surrounding objects. Through segmentation, the field of view can be cleared and simplified to allow for its interpretation. In this way, the segmentation can be exploited for the design of masks, allowing the interpretation to focus on different features of an object.

Image segmentation techniques can be classified into two broad families: contour-based and region-based approaches. Contour-based approaches, e.g. snakes, usually start with a first stage of edge detection, followed by a linking process with the aim of detecting a boundary. Region-based approaches, e.g. segmentation with eigenvectors, to be discussed later, try to find partitions of the image pixels into sets with similar image properties.

The major difference between these techniques is the decision criterion: local versus global. Contour-based techniques make local decisions (e.g. merge/split, stop/evolve) that may be proven premature. Region-based techniques make the decision by defining a global objective function. The advantage in the latter case is that decisions are made when information from the whole image is taken into account simultaneously, at the expense of higher computational costs.

The mathematical theory of the techniques described below is exhaustively presented in various scientific papers and books. Therefore, we will not aim to give a detailed mathematical description, but rather a short overview, which provides the basic literature of each technique and its applications in electron tomography. Since segmentation is a subjective process, the properties of each technique will be discussed, but not compared, since the algorithms perform differently for different types of data. Nevertheless, the examples presented for segmented tomograms will give a flavor of the performance and function of each algorithm, and will give an idea of 'what an algorithm can do' instead of 'what an algorithm cannot do'. It is important to notice that the algorithms presented can be easily combined, and that the result might surpass the performance that each algorithm is able to achieve on its own. Specific choices for tuning the algorithms toward better application to electron tomographic data sets will also be mentioned.

2. SPECIFIC DEMANDS OF ELECTRON TOMOGRAPHIC DATA ON SEGMENTATION ALGORITHMS

Electron tomographic images represent a tough challenge for most of the current image and signal processing applications in computer science. The reasons are mainly 3-fold: (i) the low signal-to-noise ratio (SNR); (ii) the missing information (missing wedge) in Fourier space; and (iii) the fact that the transfer function of the microscope is virtually unknown, especially for thick specimens. Although each of these issues alone would not be prohibitive by itself, the combination produces a very difficult task.

Cryoelectron tomograms, in particular, suffer from a very low SNR due to the low electron dose that must be used. The use of higher doses would increase the SNR, but would destroy or falsify high-frequency information due to alterations of the specimens. Even though the noise sources are multiple and hard to describe, the noise can be considered to be white or band-pass-filtered white noise (see Chapter 11 of this volume). Furthermore, the missing wedge problem is inherent in the fact that cryoelectron tomography is not capable of recording the complete 3D information, due to the limited tilt range of the tilt device (compustage). The effect is nicely demonstrated in Fourier space, where a wedge-shaped region does not contain any information (see Fig. 1 in Chapter 1 of this volume). Effectively, for a normal tilt range of ±60°, one-third of the data are missing, which results in a blurring of the object features in the direction of the beam. The blurring

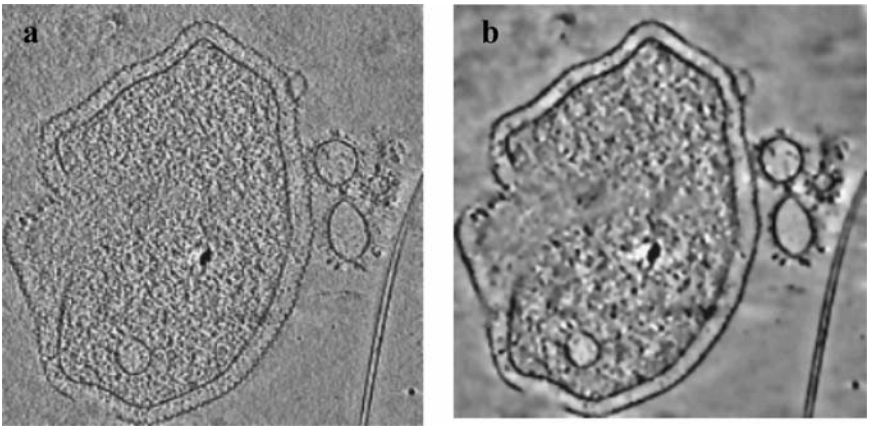


FIGURE 1. (a) A slice through a 3D image of a *Pyrodictium abyssi* cell (Baumeister *et al.*, 1999) and (b) a slice denoised by non-linear anisotropic diffusion. (From Frangakis *et al.* (2002), reproduced with permission by Elsevier).

amounts to a virtual elimination of the gradient (by gradient we mean the image gradient defined as the first derivative of the image intensity), which is the most important feature-delimiting criterion, and is used by all segmentation techniques.

3. DENOISING AND EDGE-ENHANCING TECHNIQUES

Denoising techniques are important tools for the processing of cryo-electron tomographic reconstructions, which are both complex and noisy. In the examples presented below, the segmentation techniques will be applied to 2D and 3D images that have been denoised by non-linear anisotropic diffusion (Frangakis and Hegerl, 2001; see also Chapter 11 of this volume). In non-linear anisotropic diffusion, the shape, eccentricity and direction of a Gaussian filter are modified anisotropically and adjusted locally as a function of the magnitude and direction of the local gradient and the gray flow of the image, such that edges are preserved while noise is smoothed. The segmented tomograms given as examples were all denoised by non-linear anisotropic diffusion (Fig. 1). Segmentation can, in principle, also work on non-denoised tomograms, but it is simpler to tune the parameters on data with less noise.

4. VISUALIZATION OF THE SEGMENTATION RESULTS

Most segmentation techniques produce a binary result, in the form of a mask, conveying the decision on where an object is located. The direct visual-

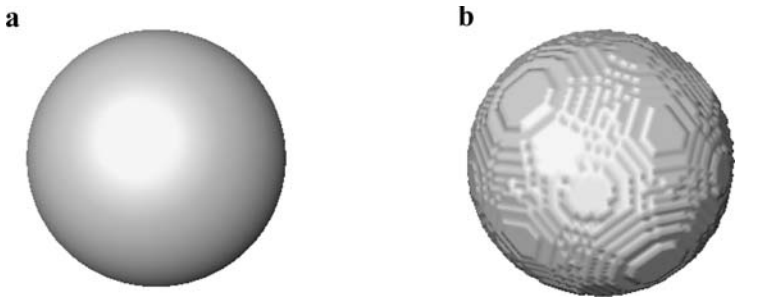


FIGURE 2. Visualization of segmentation results. For demonstration, consider a spherical mask, representing the result of segmenting a Gaussian blob. (a) Proper visualization of the segmentation is obtained by multiplying the mask by the original density distribution—no interference of the mask with the object is noticeable (e.g. in the form of discontinuities, breaks, etc.) (b) The actual mask used for the segmentation. Visualization of the segmentation result itself, which in the general case is a binary image, produces a bad, discontinuous visualization.

ization of this mask by surface rendering produces facets, with the corresponding normal vectors being perpendicular to each other, producing a non-smooth, Lego-like impression. To convey the result of a segmentation, not the mask but rather the product of the mask with the original 3D image should be visualized, which is normally free of such artefacts as the original intensities in the image are preserved. For demonstration purposes, a Gaussian function has been segmented with a sphere and an isosurface has been generated from the mask (Fig. 2*b*), as well as from the segmented object (Fig. 2*a*).

5. SEGMENTATION TECHNIQUES

5.1. Thresholding

The simplest and probably the most frequently used segmentation method is 'thresholding'. Thresholding is an operation whereby pixels or voxels are selected according to their intensity above a user-defined parameter and assigned to the foreground (white, high density, '1') or the background (black, low density, '0'). The resulting binary I_{bin} is produced in the following way:

$$I_{bin}(\mathbf{x}) = \begin{cases} 1 \text{ if } I(\mathbf{x}) \geq thr \\ 0 \text{ if } I(\mathbf{x}) < thr \end{cases} \quad (1)$$

where $I(\mathbf{x})$ is the image value at the position $\mathbf{x}$ and *thr* is the user-defined threshold value. This binary image is then used for further processing. Thresholding is implicitly used for many visualization techniques, e.g. isosurface rendering. Thresholding techniques can be applied with a large

variety of options, e.g. combination with the presentation of the density histogram, or two-value thresholding.

The choice of the threshold is typically user defined. For tomographic data, there is no objective criterion on how to define this parameter. In the special case of single-particle reconstruction of purified proteins, where the resolution of the data set is significantly better and the object under investigation has a known molecular weight, a threshold can be set by calculation of the total volume occupied by the protein, considering the density of the protein. A modification of the threshold changes the number of selected voxels and the corresponding occupied volume. In tomographic data, this way of determining the threshold is usually not possible as no volume information is available; therefore, threshold parameters must be set subjectively.

5.2. Properties and Operations of Binary Data and Masks

Segmentation of electron tomograms by means of thresholding or other techniques can result in a binary image with thousands to millions of small islands (an island being a limited region with white pixels separated from the background), most of them representing noise. Each island can be assigned to a different number, which in computational jargon is simply called a 'label'. Several different measurements exist that may be used to quantify the labels and choose which one represents the object of interest. These measurements include, but are not limited to: occupied volume, length, curvature, principal axis and moments of inertia of each label. These measurements can be used to separate labels representing regions with interesting objects from those representing noise (e.g., a cell would occupy a larger volume than shot-noise). An example of the simplicity and effectiveness of this application is visualized in Figs 3–6. A threshold was applied to a tomogram of a *Pyrodictium* cell (Fig. 1b) and a binary image was produced (Fig. 3a). The binary image was labeled and then the properties of each label were analyzed.

There is a very small number of labels occupying a very large volume, as shown in the histogram (Fig. 4). If these labels are separated from the rest and visualized separately, then it turns out that they represent pieces important for the segmentation process; i.e. the inner and outer boundary of the cell (Fig. 5). The same process can be applied on a *Dictyostelium* cell recorded by Ohad Medalia (Medalia *et al*., 2002). In this case, all the pixels occupying a volume smaller than a certain number are simply excluded from the visualization. The effects of this procedure are depicted in Fig. 6.

5.3. Binary Morphological Operations

Dependent on the procedure for generating a mask (manual or automated segmentation, thresholding, etc.), the masks tend to have holes, rough or disconnected boundaries, and other defects, which affect both the

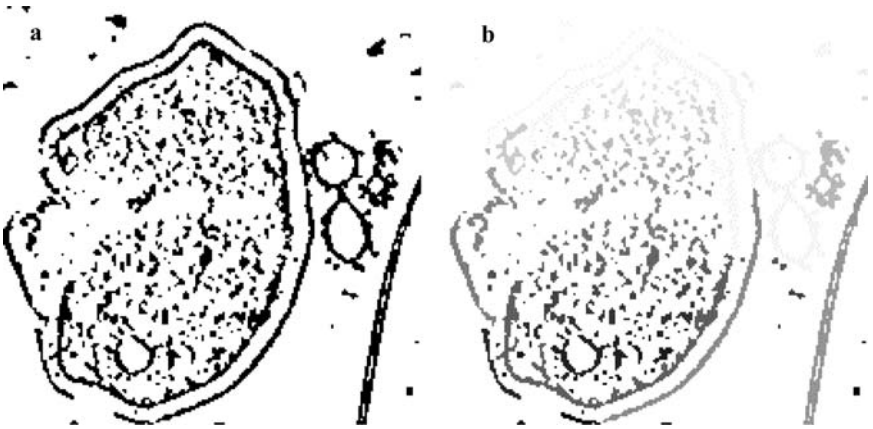


FIGURE 3. (a) A binary image created by thresholding of the denoised 3D image of *Pyrodictium*. Note that the contrast has been inverted for visualization purposes. (b) Labeling of the binary image. The gray value of the labels corresponds linearly to their label number: islands on the top of the image have a low label number, therefore they appear bright. Conversely, islands at the bottom have a higher label number, therefore they appear darker. (Note that the binary images are presented in inverted contrast, i.e. the pixels with value 1 are black and the pixels with value 0 white, for simplification of the black and white print.)

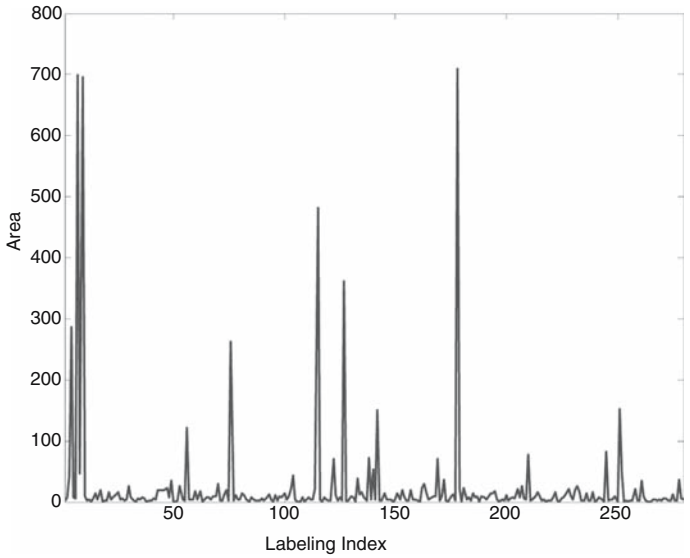


FIGURE 4. Quantization of the volume occupied by the different labels. The *y*-axis is the volume occupied, with respect to the label number plotted on the *x*-axis.

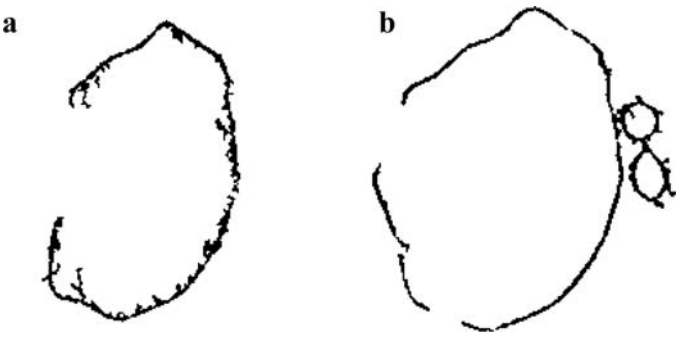


FIGURE 5. The labeled islands occupying the largest volume in Fig. 3. They prove to represent the inner membrane (a) and the outer membrane (b) of the *Pyrodictium* cell.

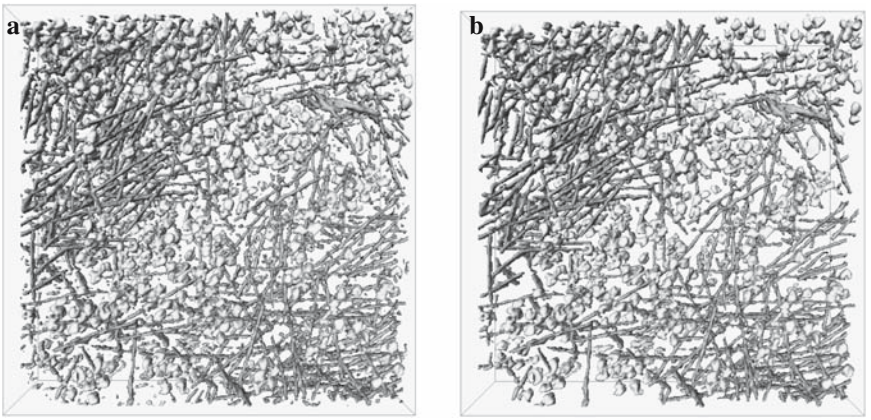


FIGURE 6. Isosurface visualization of a part of a *Dictyostelium* cell (Medalia *et al.*, 2002) (a) before and (b) after removing all labels that had <200 voxels.

segmentation results and the visualization. Morphological operations on binary images are useful for improving the quality of the masks (Weickert, 1999). Morphological operations are local modifications of the multi-dimensional image I based on the shape of a so-called structuring element S of the same dimensionality. The structural elements can be of arbitrary shape but they are typically disks (spheres, in 3D), squares (cubes) or ellipses (ellipsoids). The aim of these operations is to transform the signal into simpler ones by removing irrelevant information. Some of the most common operations are dilation and erosion. The mathematical description for a binary image of such an operation is the following:

$$\text{Dilation: } (I \oplus S)(x) := \min\{I(x - y), y \in S\} \text{ and} \tag{2}$$

$$\text{Erosion: } (I \otimes S)(x) := \min\{I(x + y), y \in S\} \tag{3}$$

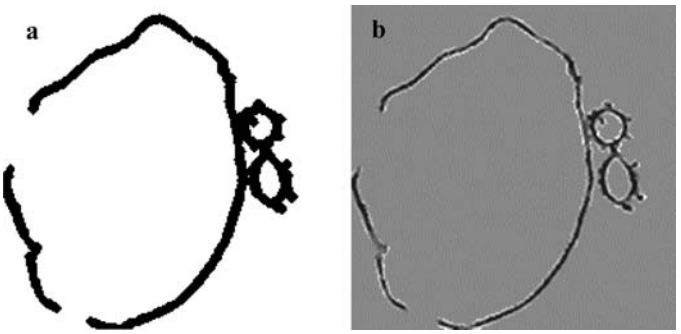


FIGURE 7. (a) Dilation of the boundaries from Fig. 5b, which represent the outer membrane of the *Pyrodictium* cell. (b) The mask is multiplied by the original object, highlighting the intensities of the cell membrane.

where $\oplus$ and $\otimes$ are symbols indicating the application of these operations on an image, given a structural element. If a pixel is 'switched on', i.e. has value 1, then its surroundings gets the value of the structuring element. A dilation operation would expand the boundaries of the mask, while an erosion operation would shrink them (Fig. 7). With an iterative and alternating application of these two elementary operations, an 'opening' or 'closing' of disconnected boundaries can be achieved, holes can be filled and various defects in the masks can be corrected in a determined way. The opening of boundaries can therefore be described in the following way:

$$\text{Opening: } (I \circ S)\ (x) := \min((I \otimes S) \oplus S)(x) \tag{4}$$

which is effectively an erosion followed by a dilation. The closing of the boundaries is the inverse operation:

$$\text{Closing: } (I \bullet S)\ (x) := \min((I \oplus S) \otimes S)(x) \tag{5}$$

Again, symbols have been introduced for a short notation of these operations. Erosion and dilation operations can also be defined for gray-level images, and be easily implemented by level sets (Sethian, 1996).

5.4. Manual Segmentation

Manual segmentation is the most subjective way of segmenting objects; however, due to the extreme complexity of electron tomographic images, it is still the most popular (see Marsh *et al.* (2001) for an extraordinary example of manual segmentation performed on highly complex 3D images). The simplest way is to use a brush, and paint the regions of interest, thereby creating a mask for the segmentation (Perkins *et al.*, 1997). Similar tools have been also developed for the direct segmentation in 3D

(Li *et al*., 1997). However, this manual tracing is both inaccurate and laborious. Solutions can be provided, for example by intelligent scissors algorithms, which allow objects within digital images to be extracted quickly and accurately using the mouse (Ballard and Brown, 1982). When the user defines a position in proximity to an object edge, the boundary snaps at the edge and the user can precisely wrap the boundary around the object of interest. The line segments are automatically fitted to the closest edge. With further mouse clicks, fixed points at the contour to be segmented can be defined. These types of algorithms find the shortest path in a cost matrix, which is computed from the image's gradient image.

5.5. Watershed Transform

The watershed transform is a well-established image processing technique for gray-scale images (Russ, 2002). In this method, the gray values of the pixels or voxels are considered as a topographical relief, where the brightness value of each voxel corresponds to a physical elevation. This relief is now considered as being filled with water. The positions where the water arriving from two different minima merge are called the watersheds (Fig. 8). The procedure results in partitioning the image into regions, the borders of which separate the minima from each other.

A common problem of the watershed algorithms is oversegmentation. Due to noise, the complexity and especially the band-pass characteristics of the data, the volume is partitioned into numerous small regions, which even outnumber the islands created by thresholding (Fig. 9). By low-pass filter-

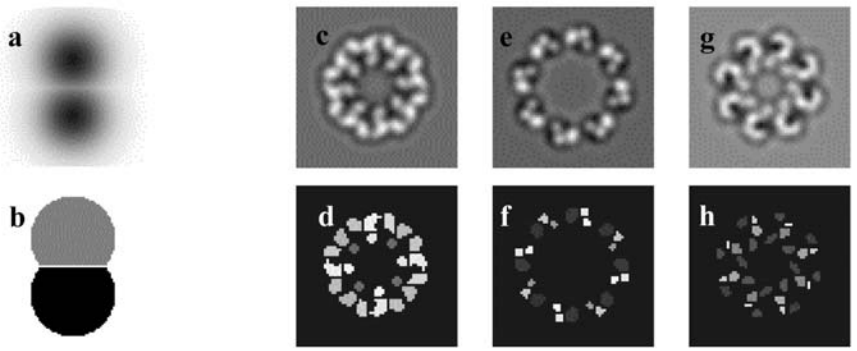


FIGURE 8. Watershed segmentation applied to synthetic data. (a and b) Synthetic data set of two Gaussian blobs touching each other. The watershed algorithm fills each region with water and, when the regions touch, then a barrier is set in between them. (b) Two distinct objects have been produced by segmentation, corresponding to the position of each blob. (c–h) Application of the watershed transform on real data, exemplified by a 3D image of the *Thermosome*. (c, e and g) Three consecutive slices through the macromolecule. (d, f and h) The corresponding results of the segmentation.

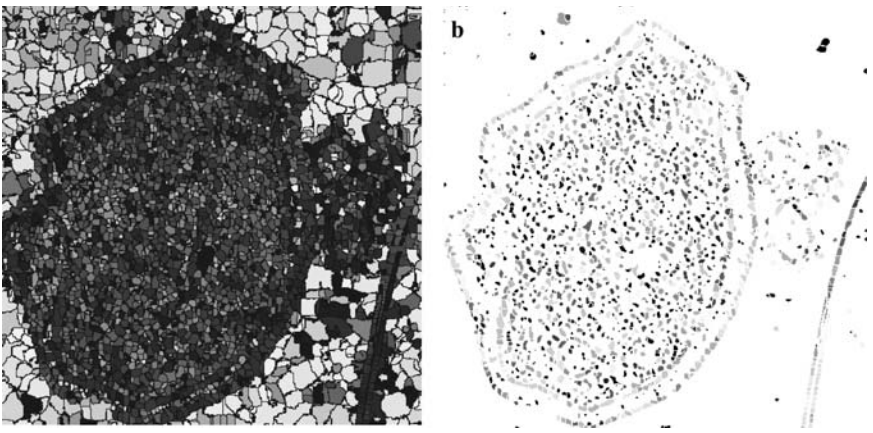


FIGURE 9. Watershed segmentation of a slice from *Pyrodictium* (Fig. 1b). On the left is the watershed result. Each watershed island has been assigned to a different label. On the right, the watershed result has been multiplied by the binary data from Fig. 3a, showing the separation of the different subunits.

ing of the data (e.g. by application of a Gaussian filter), it is possible to reduce the number of unimportant details and thus to reduce the number of regions. A different way to achieve this is by the introduction of a finite step size as an additional parameter (Volkmann, 2002). Depending on the choice of that parameter, the position of the watershed can vary, so that local, insignificant minima do not contribute to the segmentation result.

The watershed transform has real advantages when a few smooth objects need to be separated from each other. This is the case when subunits of macromolecules need to be segmented. For tomographic data, a pre-segmentation is necessary, in order to avoid oversegmentation. The watershed transform can subsequently be used to refine this segmentation, by removing any noise blobs that were included initially.

5.6. Snakes

The original idea of Kass and co-workers (Kass *et al.*, 1988) for segmenting an object was to describe the evolution of a curve $\gamma(t)$ in time t, called the 'snake', which would shrink like an elastic band and, dependent on the object properties, would balance to a stable state indicating the object boundaries. The aim of this elastic band approach is that it smoothly describes the contours (i.e. the edges) of the object and bridges gaps and openings that occur due to noise or incompleteness of the data, and might be several pixels wide. This very intuitive way of segmentation had some initial limitations in the discretization scheme, i.e. how to describe the curve when it needs to break into two parts. When two separated objects need to

be segmented, then the algorithm should be able to deal with a discrete implementation. These discretization issues need to be described when a curve is expanding from the inside out. Fortunately, these issues can generally be solved by a technique for propagating interfaces, called level sets, which was conceived by Osher and Sethian (1988). The level sets are so named because the snake $\gamma(t)$ is now considered as a $(N-1)$-dimensional curve, i.e. the zero-level set, of a higher (N)-dimensional manifold ϕ, which propagates over the time t:

$$\gamma(t) \rightarrow \phi(\mathbf{x}(t),t) = 0 \tag{6}$$

This embedding allows the snake every freedom in topological changes and movement including the application of several regularization properties. Additionally, this formulation overcomes the numerical instability problem posed by a direct discretization of the equations of a propagating front in general, and the snake in particular (Osher and Sethian, 1988). When the snake moves as a function of time toward the direction of the normal vector of its boundary with the speed F, the equation describing this movement equations is:

$$\phi_t + F \cdot |\nabla \phi| = 0, \text{ given } \phi(x, t=0) \tag{7}$$

where $F = g \cdot (\pm 1 - \varepsilon \cdot \kappa)$ is the so-called speed function (Fig. 10). The speed function is defined at every position of the image as a function of external forces $g = \dfrac{1}{1+|\nabla G_\sigma * I|}$ which depends on the gradient of image I low-passed filtered by a Gaussian filter with variance σ and internal forces depending on the curvature κ and a balance parameter ε (Malladi *et al.*, 1996). Depending on the sign of the first summand of the speed function F, the snake can shrink (+) or expand (–). The values of the weighting function are very close to one for small gradients and approximately zero otherwise. The curvature term plays a regularizing role, as it smoothes out the areas of high curvature, i.e. it eliminates small noise features and covers openings of the objects, creating more coherent regions. Since the stopping criterion depends on the strength of the gradient, and the gradient is essentially zero in the direction of the beam, due to the missing wedge, it might be advantageous to modify the force pushing the snake toward the object. Thereby the force component in the beam direction can be set to zero, in order to avoid a collapse of the snake due to the absence of a useful boundary.

The user must make the effort to tune the internal regularization parameters, i.e. the curvature as well as the external parameters, which are image dependent, such as the image gradient, etc. The parameter setting is usually chosen in such a way that small and insignificant features can be ignored, so that the snake will move toward, and smoothly lean against, the object boundary.

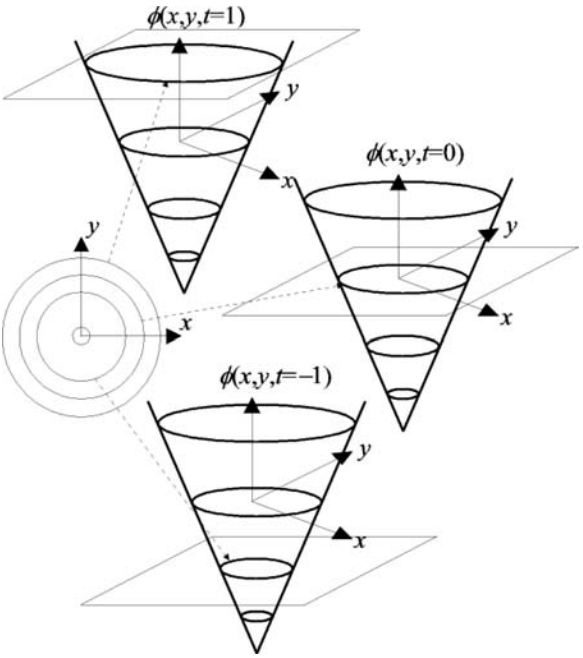


FIGURE 10. Demonstration of the expansion or collapse of a snake (zero-level set) as a function of the normal vector and the iteration time. The position of the zero-level set is indicated by the square. As the iteration time t evolves ($t = -1, 0, 1, \ldots$), the square moves higher up and the curve expands. In this case, no external forces are considered. Through the description of the 3D function, ϕ the snake always stays mathematically describable, even if the curve should break.

5.7. Initialization of the Segmentation Curve for Electron Microscopy Data

The segmentation is sensitive to the initial snake conditions. An arbitrarily chosen snake initialization will probably not converge to the desired solution, since it might get caught in high-contrast, large-scale features. Therefore, it will be advantageous for the user to indicate a rough initialization curve and then let the snake evolve under the pre-set parameters. For a fully automated segmentation, Bajaj and co-workers (Bajaj *et al.*, 2003) suggested a procedure for generating an initialization curve that is based on the calculation of seed points where the snakes get initialized and from where they will expand. In order to guarantee a reliable stopping criterion, they initialized 'dual contours', which move antagonistically to the segmentation contours from different positions in the 3D image. Ultimately, both contours will approach each other closely and balance at the boundaries of the objects, providing a nice segmentation. An example for a segmentation based on this technique is presented in Fig. 11.

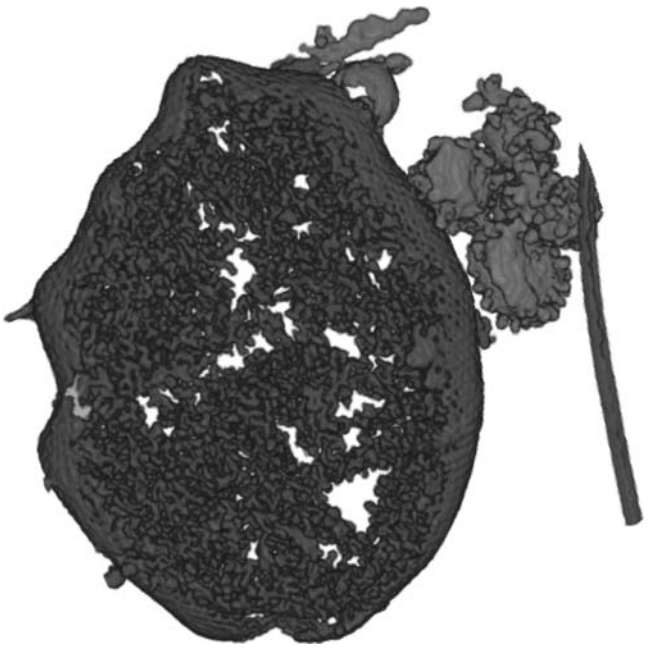


FIGURE 11. Segmentation of a slice from *Pyrodictium* using level sets according to the procedure described by Bajaj *et al.* (2003). (See color plate)

5.8. Manual Segmentation with Snakes

The snakes can also be used for manual segmentation. They interact with the image in a way similar to the intelligent scissors by interfering with the edges. The difference from the automated segmentation is that now the snakes are moved and stopped manually by the user. A minimization of the energy functional is not necessary, since the user decides at which position the evolution will stop. Parameters such as edge strength and curvature of the snake can be applied in order to assist the user to place the snake faster and more accurately in the desired position. In this way, different tolerance values can be applied in order to stop the contour from moving at certain points. The connectivity of the snake is preserved at all times.

5.9. Segmentation with Eigenvectors

The segmentation with eigenvectors is based on the minimization of a global cost function, which exploits the information given by all pixels of an image as a whole. Its goal is to subdivide the image into two parts, e.g. foreground and background, thereby extracting the 'main' feature in the image (Malik *et al.*, 2001). Theoretically, the segmentation result should constitute a binary image. In practice, however, gray levels are assigned to each pixel, with a distribution of intensity values, such that a meaningful area can

easily be extracted by simple thresholding. Segmentation into more areas is achieved by a hierarchically descending application of the procedure to the segmented regions.

The procedure starts by assigning a similarity value $w_{i,j}$ to every pair i,j of pixels in the image as a bilinear form. These values form the affinity matrix **W** between all pixels **x** with gray values I_i. The similarity measurements can be a function of intensity, gradient, texture, energy and distance. For multidimensional electron microscopic images (in 2D and 3D), the affinity matrix, as for instance used in Frangakis and Hegerl (2002), is a combination of the similarity in gray value and the Euclidean distance:

$$w_{i,j} = \exp\left(-(I_i - I_j)^2 / a_I\right)\exp\left(-(\mathbf{x}_i - \mathbf{x}_j)^2 / a_d\right) \tag{8}$$

where a_I is a variable adjusting for the importance of the gray value and a_d for the importance of the Euclidean distance. Large values of $w_{i,j}$ indicate high overall similarity between the two pixels.

The aim is now to separate the total set of pixels or voxels into two disjoint subsets A and B with maximum dissimilarity ($V = A \cup B$ and $A \cap B = \varnothing$). The dissimilarity thereby is measured by the sum of the elements $w_{i,j}$ of the affinity matrix, which connect an element of A with an element of B. Shi and Malik (1997) introduced for that purpose the *normalized cut* criterion, as a possible way to quantify this notion:

$$Ncut(A,B) = \frac{cut(A,B)}{asso(A,V)} + \frac{cut(B,A)}{asso(B,V)} \tag{9}$$

with $asso(A,V) = \sum_{i\in A, j\in V} w_{ij}$ and $cut(A,B) = \sum_{i\in A, j\in B} w_{ij}$. A good segmentation would minimize $Ncut(A,B)$. Therefore, the solution that minimizes the normalized cut criterion is the optimal separation of the pixels into two subsets and, correspondingly, of the image into two regions. As shown by Shi and Malik, the solution of this criterion can be formulated as an eigenvalue–eigenvector problem (Shi and Malik, 1997). Following this formulation, the connectivity pattern that minimizes this criterion can be approximated by the second smallest eigenvector of the affinity matrix **W**. The vector elements of this 'indicator' vector assign a number to each pixel, such that the entries cluster around two different values, which are interpreted as two different labels identifying the two pixel subsets.

5.10. Segmentation Procedure

The inspection of the segmentation results confirms the feasibility of foreground/background separation, for which this technique has been designed. The indicator vector containing the segmentation information possesses bright gray values at the place where the cell lies, and darker values everywhere else (Figs 12–14).

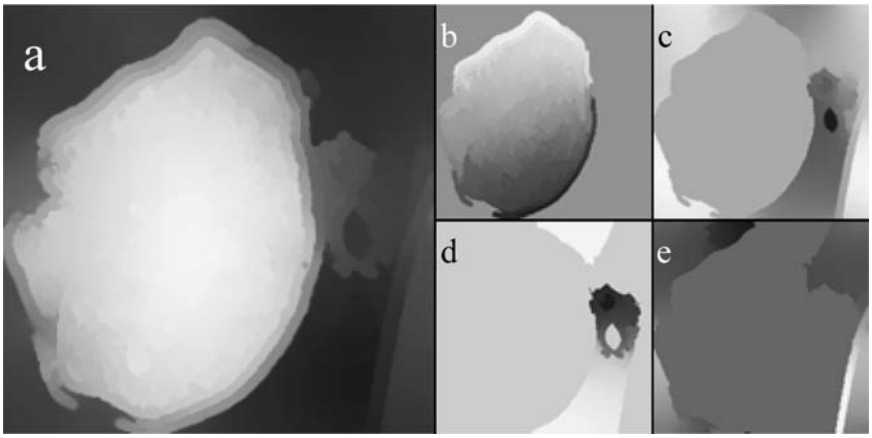


FIGURE 12. Results of eigenvector-based segmentation for a slice from a tomogram of the *Pyrodictium abyssi* cell; (a) the second eigenvector representing the segmentation result. (b–e) Hierarchical segmentation: the high-intensity values are extracted from (a) with thresholding. These high-intensity values are subjected to a further segmentation producing (b) and (c). In the next application step of the algorithm, the segmentation of two further features is achieved (d and e). (From Frangakis *et al.* (2002), reproduced with permission of Elsevier).

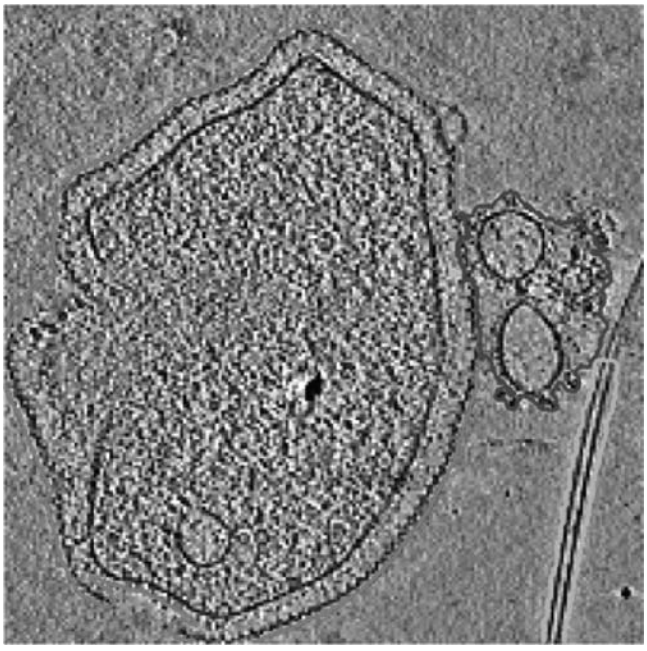


FIGURE 13. Segmentation of a slice from the tomogram of *Pyrodictium* using the eigenvector method. The original slice is overlaid with contour plots of the segmentation. The three main features on the image are seen to be separated from one another. (From Frangakis *et al.* (2002), reproduced with permission of Elsevier). (See color plate)

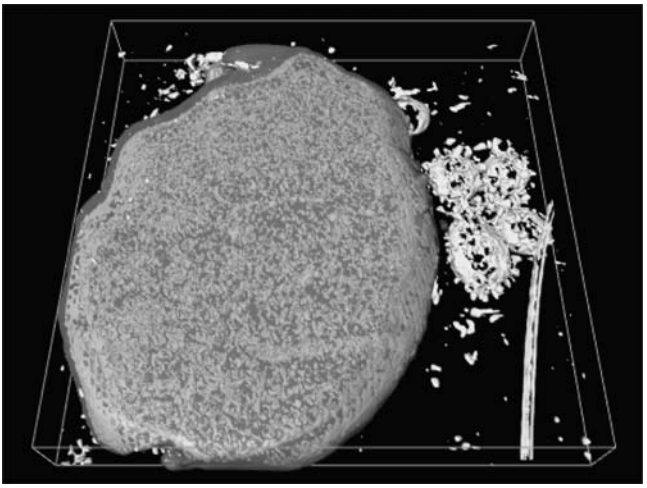


FIGURE 14. Eigenvector-based segmentation presented in Fig. 13, visualized in three dimensions. (From Frangakis *et al.* (2002), reproduced with permission by Elsevier). (See color plate)

6. CONCLUSIONS

Segmentation is a largely subjective operation, which increasingly becomes a significant aid for the visualization of 3D images. There are several segmentation algorithms in the literature. Their performance depends in most of the cases on the application, and there is no single algorithm (at least none known to the authors) that performs equally well with various types of data. Therefore, only those algorithms that are gaining importance in electron tomography or have been specifically developed for electron tomography were presented and exemplified in this chapter. Perhaps in computer science books, some even better performing algorithms might exist, and the existing ones might be further improved in the future. Nevertheless, we hope the examples shown have indicated that by applying the algorithms presented above, good segmentation results can be obtained and a large variety of data types can be analyzed. It should be noted finally that the algorithms are not mutually exclusive but, on the contrary, they can be mixed and matched for best results. For all of the algorithms discussed, fast implementations exist which can be integrated in many visualization packages currently in use.

ACKNOWLEDGEMENTS

The authors would like to thank Anne-Marie Glynn, Daniel Castano and Anja Seybert for a critical reading of the manuscript, and Ohad Medalia

and Rudo Grimm for providing the data of the *Dictyostelium* and the *Pyrodictium* cells, respectively. The authors would like to thank Joachim Frank for a critical reading of the manuscript.

REFERENCES

Bajaj, C., Yu, Z. and Auer, M. (2003). Volumetric feature extraction and visualization of tomographic molecular imaging. *J. Struct. Biol.* **144**:132–143.

Ballard, D. H. and Brown, C. M. (1982). *Computer Vision*. Prentice Hall, Englewood Cliffs, New Jersey.

Baumeister, W., Grimm, R. and Walz, J. (1999). Electron tomography of molecules and cells. *Trends Cell Biol.* **9**:81–85.

Frangakis, A. S. and Hegerl, R. (2001). Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. *J. Struct. Biol.* 135:239–250.

Frangakis, A. S. and Hegerl, R. (2002). Segmentation of two- and three-dimensional data from electron microscopy using eigenvector analysis. *J. Struct. Biol.* **138**:105–113.

Kass, M., Witkin, A. and Terzopoulos, D. (1988). Snakes: active contour models. *Int. J. Comput. Vis.* **1**:362–385.

Li, Y., Leith, A. and Frank, J. (1997). Tinkerbell—a tool for interactive segmentation of 3D data. *J. Struct. Biol.* **120**:266–275.

Malik, J., Belognie, S., Leung, T. and Shi, J. (2001). Contour and texture analysis for image segmentation. *Int. J. Comput. Vis.* **43**:7–27.

Malladi, R., Sethian, J. A. and Vemuri, B. C. (1996). A fast level set based algorithm for topology-independent shape modeling. *J. Math. Imag. Vis.* **6**:269–289.

Marsh, B. J., Mastronarde, D. N., Buttle, K. F., Howell, K. E. and McIntosh, J. R. (2001). Organellar relationships in the Golgi region of the pancreatic beta cell line, HIT-T15, visualized by high resolution electron tomography. *Proc. Natl Acad. Sci. USA* **98**:2399–2406.

Medalia, O., Weber, I., Frangakis, A. S., Nicastro, D., Gerisch, G. and Baumeister, W. (2002). Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* **298**:1209–1213.

Osher, S. and Sethian, J. A. (1988). Fronts propagating with curvature dependent speed: algorithms based on Hamilton–Jacobi formulation. *J. Comput. Phys.* **79**:12–49.

Perkins, G. A., Renken, C. W., Song, J. Y., Frey, T. G., Young, S. J., Lamont, S., Martone, M. E., Lindsey, S. and Ellisman, M. H. (1997). Electron tomography of large, multicomponent biological structures. *J. Struct. Biol.* **120**:219–227.

Russ, J. C. (2002). *The Image Processing Handbook*. CRC Press, Boca Raton, Florida.

Sethian, J. A. (1996). *Level Set Methods Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge University Press.

Shi, J. and Malik, J. (1997). Normalized cuts and image segmentation. Paper presented at: Proceedings of the IEEE Conferences on Computer Vision and Pattern Recognition.

Volkmann, N. (2002). A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. *J. Struct. Biol.* **138**:123–129.

Weickert, J. (1999). *Anisotropic Diffusion in Image Processing*, B. G. Teubner Stuttgart.

13

Segmentation of Cell Components Using Prior Knowledge

Ming Jiang, Qiang Ji, Xun Wang and Bruce F. McEwen

1. INTRODUCTION

Electron tomography is a method for determining 3D structure by electron microscopy, using multiple tilt views of the specimen (Lucic *et al.*, 2005; McEwen and Marko 2001; McIntosh *et al.*, 2005). Since electron

Ming Jiang, Qiang Ji and Xun Wang • Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

Bruce McEwen • Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, PO Box 509, Albany, NY 12201-0509, USA

tomography does not employ averaging or require the presence of symmetry, it can be used in biological applications to image single copies of subcellular components *in situ*. When specimen preparation is optimized by use of rapid freezing, and imaged either directly in the frozen-hydrated state, or after freeze substitution and plastic embedding, electron tomography provides a relatively high-resolution view of biological structure in a native, or near-native, cellular context.

To realize the full potential of this powerful breakthrough, investigators must be able to identify and segment components of interest from the complex and densely packed cellular environment characteristic of well-preserved biological specimens. This is particularly challenging because, in addition to the crowded environment they portray, images of biological specimens have inherently low contrast and a low signal-to-noise ratio (SNR). As a result, segmentation of cellular components has been dominated by manual procedures that rely on the expert knowledge of biologists to recognize specific structures. However, such manual procedures are time-consuming, subjective and ill suited to handling the data throughput required to make statistical correlations among data sets recorded under differing functional conditions.

Many semi-automatic and fully automatic segmentation methods have been developed to overcome the limitations of manual segmentation, and some of these have been adapted for electron tomographic volumes. In this chapter, we first review segmentation work in the field of electron tomography, and then discuss the rationale for using prior knowledge to improve the segmentation of a 3D reconstruction portraying a densely packed cellular environment with a high amount of noise present (low SNR). This is followed by a detailed description of our implementation of prior knowledge-based segmentation, including practical applications, experimental results and evaluations.

2. REVIEW OF VOLUME SEGMENTATION IN ELECTRON TOMOGRAPHY

Quantitative analysis of cell components in an electron tomographic volume has become an important tool at the frontier of structural biological research. This analysis often involves segmenting cell components, measuring their dimensions, locating critical points and determining spatial relationships among the components (e.g. Harlow *et al*., 2001; Marsh *et al*., 1998; Perkins *et al*., 1997; Renken *et al*., 2002; Scorrano *et al*., 2002). In a number of applications, segmentation was achieved primarily by stacking the manually traced contours from the individual slices to construct the 3D model (He *et al*., 2003; McEwen and Marko 1999; Tregear *et al*., 2004). Several software packages can be used to contour electron tomographic data sets, including: SYNU (Perkins *et al*., 1997), IMOD (Kremer *et al*., 1996), SPIDER/STERECON (Frank *et al*. 1996; Marko and Leith, 1996) and NIH Image (Rasband and Bright, 1995).

Manual tracing quickly becomes tedious and prone to operator error for complex structures such as the Golgi complex (Marsh *et al.*, 2001). It is not computationally practical when a large number of data sets are involved. To address this issue, semi-automated segmentation techniques have been developed that insert models for common objects, such as vesicles, and propagate contours traced manually in selected 2D slices across all other slices in the volume (Harlow *et al.*, 2001; Hessler et al., 1992; Hessler *et al.*, 1996; Marko and Leith 1996; Marsh *et al.*, 2001; Segui-Simarro *et al.*, 2004; Sosinsky *et al.*, 2005). Most semi-automated methods provide manual editing to correct problems with over- and undersegmentation. An alternative approach is to employ 3D tracing using a stereo viewer (Li *et al.*, 1997).

Although semi-automated procedures have clearly improved the efficiency and accuracy of tracing methods, the improvement is insufficient for segmenting the large throughput of volumes that is produced by modern methods for electron tomography. To address this issue, researchers have experimented with various semi-automated or fully automated segmentation techniques, particularly those based upon active contour methods. In the geodesic active contour method developed by Bartesaghi and coworkers (2004), a point inside the selected object must be provided for the segmentation to converge to the desired object boundaries. Bajaj and colleagues (2003) developed a boundary segmentation method based on a gradient vector diffusion and fast marching technique (Sethian 1996), a simplified and faster variant of the traditional level set method (Sethian 1999). Babu and colleagues (2004) and Jacob and colleagues (2002) applied parametric active contours to the segmentation of chromosomes and 3D DNA filaments, respectively. The common drawback of the above active contouring methods is that they require high contrast image, fine parameter tuning and appropriate initialization, in order to converge correctly.

Two other segmentation methods, graph cut (Frangakis and Hegerl, 2002) and watershed (Volkmann, 2002), have recently been extended for the segmentation of electron tomographic data. It is not clear how well they will perform for smaller structures in the image. Graph cut, watershed and the active contour methods are reviewed in more depth in Chapter 12 of this volume.

3. IMPROVED SEGMENTATION USING PRIOR KNOWLEDGE

The above segmentation methods depend mainly on raw image intensities and gradient information, which tend to be poorly defined in electron tomographic reconstructions. As a result, such data-driven methods often converge to the wrong solution when there are large image contrast variations and outliers in the tomographic volume. Frequently, however, the general shape, location and gray-level appearance of cellular objects such as the cytoskeletal filaments and membranes are well known and can be

incorporated as constraints on the model parameters or into the model-fitting procedure. While use of prior knowledge is relatively novel for segmentation of electron tomographic volumes, the concept has been widely adopted in segmentation for other modalities. In this section, we classify model-based segmentation techniques into four categories based on the way prior knowledge is incorporated. The categories are: rigid modeling, constrained deformable modeling, statistic shape modeling and contour grouping techniques.

The rigid model-based methods include template matching and Hough transform. In template matching (Böhm *et al*., 2000; Frangakis *et al*., 2002; Frank and Wagenknecht, 1984), the template is generally a structure that has been determined at high resolution, usually by an imaging modality other than electron microscopy (most often X-ray crystallography). The template is then used to detect the object in the image by cross-correlation. Despite its current limitations such as low computation efficiency, with further development template matching promises to become an effective method for locating and segmenting small macromolecular assemblies that have a relatively uniform shape.

The Hough transform is used in rigid model-based segmentation to detect parameterized straight lines or curves in images, by mapping image points into manifolds in the parameter space and then finding the peaks in the parameter space. It can be combined with edge detection to improve the efficiency and robustness of segmentation (Yu and Bajaj, 2004; Zhu *et al*., 2003). The generalized Hough transform (Ballard, 1981) was developed to detect arbitrary known shapes. The drawbacks of both the original and the generalized Hough transforms are that they cannot account for object shape variability, and they are usually time-consuming.

Constrained deformable modeling achieves improved robustness by using prior knowledge to guide deformation. One way to achieve this is to incorporate size and shape constraints explicitly into the energy function of the active contour (Ray *et al*., 2002). This makes it possible to detect obscured objects. A more general application of this technique is the use of the deformable template (Jain *et al*., 1996), in which the object shape is described by a prototype template and a set of probabilistic deformations on the template. A Bayesian scheme is employed to find a match between the deformable template and objects in the image.

The third model-based segmentation approach, statistical shape modeling methods, is represented by the active shape model (ASM) technique developed by Cootes and colleagues (1995). In ASM, the variations of object shape are estimated by applying principal components analysis to the training samples and utilizing the result to regulate the segmentation result. The ASM has been used successfully for many tasks in general medical imaging (Behiels *et al*., 1999; Duta and Sonka 1998; Smyth *et al*., 1996), as well for as segmentation of nerve capillary structures from electron microcope images (Rogers *et al*., 2000). A similar statistical model was

proposed by Staib (1990). In their work, concise parameterized models of contours or surfaces were developed using Fourier descriptors to allow the systematic incorporation of prior shape information in a general and flexible way (Staib and Duncan, 1992; Worring *et al*., 1996). Prior probability distributions on the parameters are used to introduce a global shape model with a bias toward an expected range of shapes. A drawback of this Fourier descriptor-based model is that it has difficulty in characterizing sharp corners along the object boundary.

Contour grouping techniques (Elder and Zucker 1996; Mahamud *et al*., 2003; Wang *et al*., 2003) are based on high-level perceptual cues such as closure, proximity and continuity, instead of specific structural knowledge about the object shape. The grouping usually consists of searching for a weighted shortest closed path in a graph constructed from the edge map. One problem with grouping methods is that it is difficult for them to incorporate prior knowledge other than these high-level cues. Also, the construction of the graph is a non-trivial task due to gaps and junctions in the edge map. Recently, Elder and colleagues (2003) improved contour grouping by combining it with models of object shape and appearance. However, the performance of their method has not yet been validated for medical imagery.

Although incorporation of prior knowledge into segmentation has been well studied, application of the above techniques to segmentation of cell components from electron tomographic volumes requires further refinements, due to the unique challenges imposed by these data sets. First, the segmentation should allow sufficient flexibility to reflect biological shape variability exhibited by the objects in the volume. Overemphasis of prior knowledge, i.e. rigid size and shape constraints, tends to produce similar segmentations for dissimilar objects, thereby making comparisons among the objects meaningless. Secondly, although the goal of segmentation is often to extract fine features within cell components, a direct detection of such fine features in the whole tomographic volume involves searching too great a space and is excessively sensitive to irrelevant features at similar scales. Therefore, it is much more efficient and effective to design multiple model-based techniques that decompose the prior knowledge at different levels. For example, it is often appropriate first to locate large-scale objects spatially related to the fine features and then to segment the fine features in the localized context.

In the rest of this chapter, we demonstrate the use of a model-based segmentation framework that fully utilizes prior geometric knowledge, with two example applications: the extraction of kinetochore microtubules along with the associated plus-ends, and the segmentation of membranes. For the first application, we have developed an automated approach with a coarse-to-fine scale scheme consisting of coarse feature enhancement, object localization, fine feature enhancement and object segmentation. For the second application, we have designed a deformable contouring method with constrained optimization to integrate boundary and region information. In

both cases, prior knowledge, including object geometrical properties and gray-level appearance, is systematically incorporated into the segmentation procedures to achieve robust and accurate segmentation.

4. EXTRACTION OF KINETOCHORE MICROTUBULES AND PLUS-ENDS

In this section, we demonstrate our prior knowledge-based methodology with the segmentation of kinetochore microtubules and their morphologically variable plus-ends (Howard and Hyman, 2003; Maiato *et al.*, 2004; McEwen *et al.*, 2002; O'Toole *et al.*, 2003). Automated extraction of the fine features from the tomographic volume is challenging, because these structures are in close contact with the cellular matrix and are densely surrounded by proteins of similar appearance, as shown in Fig. 1. In this section, we present a model-based approach to the segmentation of electron tomographic volumes as an alternative to the more time-consuming and subjective manual operations.

In order to build a framework for model-based segmentation that fully exploits prior knowledge about microtubules and their plus-ends, we considered the following known features: (i) microtubules are tubular objects on a large scale; (ii) the microtubule wall constitutes a surface feature in the localized volume; (iii) the variability of microtubule cross-sectional contours can be modeled statistically; (iv) microtubule plus-ends display variable amounts of curling that can be modeled as curved surface patches or

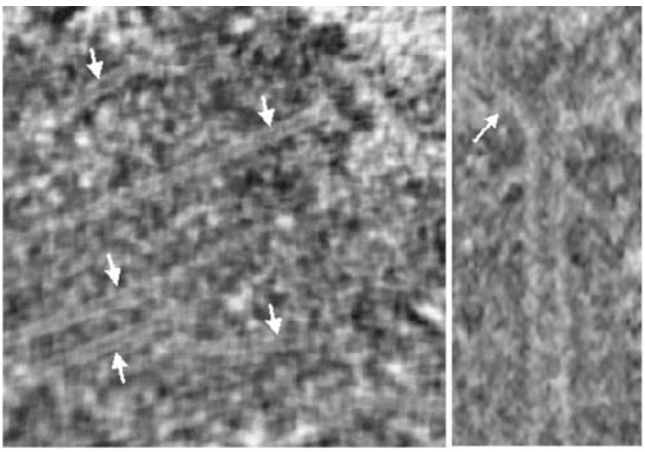


FIGURE 1. Left: a slice in the tomographic volume of a metaphase PtK cell kinetochore with associated microtubules. The microtubules (arrows) are embedded in a dense background of the cellular cytoplasm, or in the kinetochore plate. Right: a close-up view of a microtubule, with the plus-end indicated by an arrow. (Left panel: from Jiang *et al.* (2004b), © 2004 IEEE).

filamentous sheets that are connected to the body of the microtubule; and (v) the curvature of microtubule plus-ends does not change sign because the plus-end is either straight (zero curvature) or curling away from the microtubule wall. Our automated approach then extracts the microtubules, including their plus-ends, with a coarse-to-fine scale scheme. First, coarse tubular features are enhanced using an anisotropic invariant wavelet transform and a geometrical tube-enhancing filter. Next, microtubules are localized with a thinning algorithm applied to the enhanced tubular features. Then, fine features are enhanced using a geometrical surface-enhancing filter to accentuate the microtubule boundary features. In the segmentation phase, the main body of the microtubule is segmented with a modified ASM. Starting from the end of the tubular portion of the microtubule, the plus-ends are segmented with a probabilistic tracing method. The following subsections describe the approaches and present examples of experimental results.

4.1. Coarse Feature Enhancement

The low SNR and the presence of irrelevant structures with similar intensities nearby necessitate a pre-processing step to enhance the microtubules. For this purpose, we use a model-based feature enhancement, which combines an anisotropic invariant wavelet transform with a geometrical tube-enhancing filter. The wavelet transform enhances the microtubule globally in the transform domain, while the tube-enhancing filter locally accentuates tubular features. The output of this coarse feature enhancement is used to localize the microtubules.

4.1.1 Anisotropic Invariant Wavelet Filtering

In the tomographic volume, the microtubule features are so weak locally that any general averaging technique used for noise reduction will readily smooth out the microtubules completely. Based upon the tubular shape of the microtubules, we use a wavelet transform with anisotropic basis for enhancement.

Given a 1D orthogonal wavelet basis:

$$\psi_{j,k}(\boldsymbol{t}) = 2^{-(j/2)}\psi(2^{-j}\boldsymbol{t} - \boldsymbol{k}), \quad \boldsymbol{j},\boldsymbol{k},\boldsymbol{t} \in Z \tag{1}$$

where ψ denotes the mother wavelet, Z is the integer set, $\boldsymbol{j}$ denotes scale and $\boldsymbol{k}$ denotes the translation. The anisotropic basis functions for higher dimension can be constructed by the tensor product of the 1D wavelet basis functions. For example, the basis functions for 2D wavelet transform can be obtained as:

$$\psi_{j_1,k_1;j_2,k_2}(\boldsymbol{t}_x,\boldsymbol{t}_y) = \psi_{j_1,k_1}(\boldsymbol{t}_x)\cdot\psi_{j_2,k_2}(\boldsymbol{t}_y) \tag{2}$$

where $\boldsymbol{j}_i,\boldsymbol{k}_i \in \boldsymbol{Z},(\boldsymbol{t}_x,\boldsymbol{t}_y) \in \boldsymbol{Z}^2$.

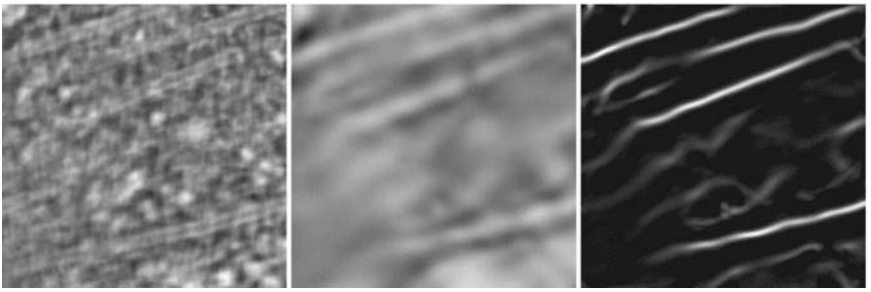


FIGURE 2. Wavelet and tubular enhancement. Left: the original image. Middle: enhanced by application of the wavelet transform to the left hand image. The white tubular features correspond to microtubules. Right: enhanced further by application of the tube-enhancing filter to the middle image. (From Jiang *et al.* (2006a), © 2006 IEEE).

With a different scale parameter for every dimension, the anisotropic basis is superior to an isotropic basis, in that it can transform anisotropic image features more effectively into coefficients with large magnitude. This allows us to set a large threshold on the inverse transform to remove not only noise but also isotropic features, while preserving anisotropic features. An inherent problem of the conventional wavelet is its lack of shift invariance, which will result in artifacts on inverse transformation; these are especially serious at image discontinuities such as edges. To overcome this problem, we have made the wavelet transform shift-invariant by averaging over shifts, and orientation-invariant by averaging over rotation (Yu *et al.*, 1996). Figure 2 shows the enhancement of the tomographic image with the wavelet transform. Since microtubules are elongated structures, their global features are well preserved for further processing. Symmlet 8 basis was used in this and subsequent experiments. In practice, we apply 2D wavelet transform to the 3D volume data in a slice-by-slice manner. This approach is much more computation efficient than true 3D transform. It is also sufficient to capture the microtubule features, which are highly anisotropic in the 2D plane.

4.1.2. Geometrical Tube-enhancing Filter

After wavelet filtering, the microtubules are still embedded in the low-contrast volume. In this situation, it is difficult to use intensity-based methods to enhance the desired features further. Therefore, we enhance the microtubules with a 3D geometrical tube-enhancing filter by exploiting the local geometric properties of microtubules as tubular structures.

A filter based on the Hessian matrix was previously proposed to produce maximum response for voxels located on tubular structures (Frangi *et al.*, 1999). The problem with this filter is that gradient information is totally ignored, and the output typically decays smoothly towards the

object boundary. In our application, a filter with sharp response is more useful in distinguishing the tubular structures from other object morphologies in low-contrast images. To address this issue, we propose a tube-enhancing filter constructed with the eigenvalues of the Weingarten matrix (Monga *et al.*, 1992).

Consider the hypersurface $\boldsymbol{I}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ associated with the 3D volume intensity $\boldsymbol{S}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ described by:

$$\boldsymbol{S}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \rightarrow (\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\ \boldsymbol{I}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})) \tag{3}$$

At each point of this hypersurface, there are three principal directions corresponding to the three extreme values of the curvatures of the hypersurface. To compute the principal directions and the curvatures, we use the Weingarten matrix expressed as

$$\mathbf{W} = \mathrm{F}_2 \cdot \mathrm{F}_1^{-1} \tag{4}$$

with first fundamental form:

$$\mathbf{F}_1 = \begin{pmatrix} 1+I_x^2 & I_x I_y & I_x I_z \\ I_x I_y & 1+I_y^2 & I_y I_z \\ I_x I_z & I_y I_z & 1+I_z^2 \end{pmatrix} \tag{5}$$

and second fundamental form:

$$\mathbf{F}_2 = -\frac{1}{\sqrt{1+I_x^2+I_y^2+I_z^2}} \begin{pmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{xy} & I_{yy} & I_{yz} \\ I_{xz} & I_{yz} & I_{zz} \end{pmatrix} \tag{6}$$

where subscripts denote image derivatives that can be obtained by convolving the volume with derivative Gaussian kernels.

Let λ_1, λ_2 and λ_3 be the eigenvalues of the matrix W with magnitudes in increasing order, and let $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ be the corresponding eigenvectors. The eigenvalues are the curvatures of the hypersurface, with the corresponding eigenvectors representing the principal directions. The relationships among these curvatures provide insight into the geometric properties of the local image structure. For tubular structures in 3D image, it can be assumed that λ_1 is small, while both λ_2 and λ_3 have large magnitude and negative signs. Based on this, we use the following geometrical filter to produce maximum response to voxels on the center line of a tubular object:

$$\boldsymbol{L} = \begin{cases} 0 & \textbf{\textit{if}} \quad \lambda_2 > 0 \quad \textbf{\textit{or}} \quad \lambda_3 > 0 \\ \left(1-\exp\left(-\dfrac{\boldsymbol{K}_a^2}{2\alpha^2}\right)\right)\exp\left(-\dfrac{\boldsymbol{K}_b^2}{2\beta^2}\right)\left(1-\exp\left(-\dfrac{\boldsymbol{h}^2}{2\gamma^2}\right)\right) & \textbf{\textit{else}} \end{cases} \tag{7}$$

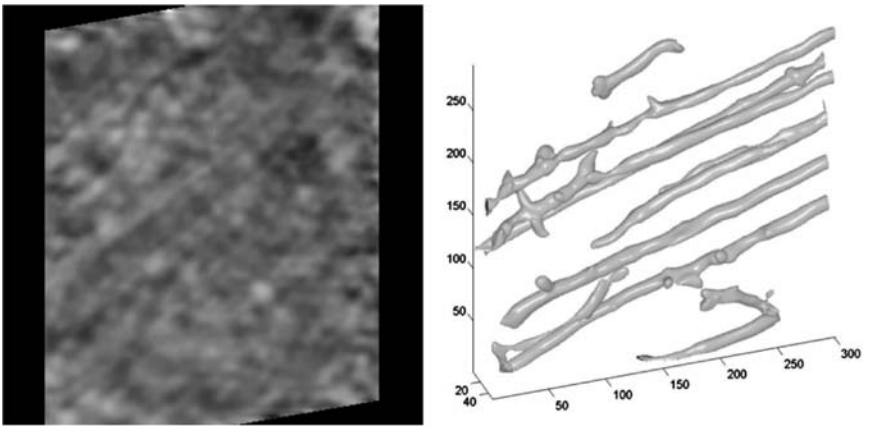


FIGURE 3. Volume enhancement. Left: the original volume. Right: the enhanced volume. The microtubules in the original volume are barely visible, due to the surrounding materials in the volume. (From Jiang *et al.* (2006a), © 2006 IEEE).

where the local shape measure $\boldsymbol{K}_a = |\lambda_2|/|\lambda_3|$ indicates cross-sectional circularity, while $\boldsymbol{K}_b = |\lambda_1|/\sqrt{|\lambda_2\lambda_3|}$ indicates elongation, and $\boldsymbol{h} = \|\boldsymbol{W}\|_{\mathrm{F}} = \sqrt{\sum_i \lambda_i^2}$ is used to suppress weak background features. α, β and γ are thresholds selected experimentally. Both α and β are set to 0.5, and γ is set to half of the maximum **h** in the image. By calculating the output value of this filter at each voxel, we can selectively enhance tubular structures and attenuate other structures. Since the microtubules have stable diameters in our data sets, we use fixed scales for the derivative Gaussian kernels. However, adaptation to various microtubule diameters (due to magnification) could be achieved in a straightforward way by convolving with a set of kernel scales and choosing the maximum response across the scales.

The response of our filter is sharp near the center line of tubular objects, and it falls off quickly away from the center line. Our filter is also insensitive to contrast variations. This is advantageous, because the microtubules typically show large contrast variations along their length. As a result, our filter is effective at enhancing tubular features (Fig. 2, right hand panel). In Fig. 3, we show a surface-rendered view of an enhanced $300 \times 300 \times 60$ volume. Initially the microtubules are barely visible in the raw volume on the left because they are buried in noise and surrounded by numerous cellular materials. The right hand panel of Fig. 3 shows that the microtubule locations become obvious after the enhancement.

4.2. Microtubule Localization

Since most of the fine microtubule boundary features can be distinguished meaningfully only in a constrained volume, localization is indis-

pensable for the subsequent fine feature enhancement and segmentation steps. We localize the microtubules by extracting their center lines in the enhanced volume with a thinning method that exploits the properties of tubular structures. Our method is similar to the ridge-based method (Aylward and Bullitt, 2002), but it requires neither seeding points nor specified tracing directions. The center lines are extracted in the enhanced volume as follows:

1. Each foreground voxel is evaluated against its eight neighbors in the local normal plane defined in 3D by the eigenvectors $\mathbf{v}_2$ and $\mathbf{v}_3$, perpendicular to the principal direction defined by $\mathbf{v}_1$. If the value of the voxel is a local maximum in the normal plane, then the voxel is recorded as a candidate center line voxel. Since the enhanced microtubule has a sharp intensity ridge on the center line, this local maximum indicates the candidate center line voxel reliably.
2. After all voxels have been evaluated, the center lines are formed by connecting the neighboring local maximum voxels. Morphological restrictions are used to trim off minor branches and to remove isolated voxels. In the end, only center lines with length above a specified threshold are kept as reliable references for localization of the microtubules. The extracted center lines for the enhanced volume in Fig. 3 are illustrated in Fig. 4, where they are used to localize subvolumes containing each microtubule. Note that the short enhanced features in the right hand panel of Fig. 3 have been eliminated from the center line plot. The right hand panel of Fig. 4 shows the center lines as viewed from an oblique direction.

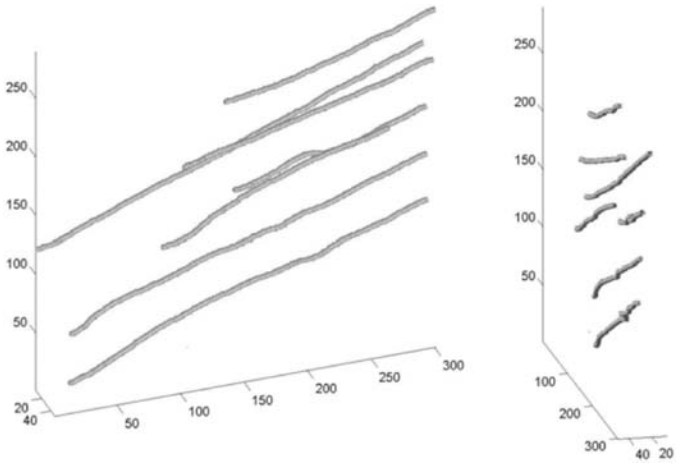


FIGURE 4. Extracted microtubule center lines from the enhanced volume shown in the right hand panel of Fig. 3. Left: the same orientation as in Fig. 3. Right: seen in an oblique view. Shorter enhanced features in Fig. 3 are removed by an adjustable minimum length constraint that filters out irrelevant structures.

4.3. Fine Feature Enhancement

The volume enhanced in the previous steps is not directly usable for segmentation, because most of the fine features have been smoothed out, as the price of enhancement for localization at a large scale. We enhance the localized volume at fine scale with a surface-enhancing filter because the boundaries of microtubule walls and the plus-end filaments can be characterized as surface patches or filamentous sheets. The surface-enhancing filter is also constructed with the eigenvalues of the Weingarten matrix. The idea is that for surface features, one eigenvalue, λ_3 , has a larger magnitude corresponding to the intensity change along the surface normal, while the other two eigenvalues λ_1 and λ_2 have smaller magnitudes corresponding to the tangent directions of the local surface. While it is possible to distinguish between a curved surface and a flat surface by further comparing the two smaller eigenvalues, it is not desirable since we do not want to produce bias to either one of them. Therefore, we choose to use only the largest and the smallest eigenvalues to construct the geometrical surface-enhancing filter as:

$$\boldsymbol{F} = \begin{cases} 0 & \textbf{\textit{if}} \quad \lambda_3 > 0 \\ \exp\left(-\dfrac{\boldsymbol{K}_c^2}{2\alpha^2}\right)\left(1-\exp\left(-\dfrac{\boldsymbol{h}^2}{2\beta^2}\right)\right) & \textbf{\textit{else}} \end{cases} \tag{8}$$

where $\boldsymbol{K}_c = |\lambda_1|/|\lambda_3|$, and α, β and $\boldsymbol{h}$, have the same meaning as in the tube-enhancing filter.

Figure 5 demonstrates that the surface-enhancing filter enhances microtubule boundaries significantly, by exploiting full 3D information. These enhanced microtubule wall features provide a boundary strength model that will be used in subsequent microtubule segmentation. It can also be seen from Fig. 5 that the enhanced microtubule cross-sectional contours

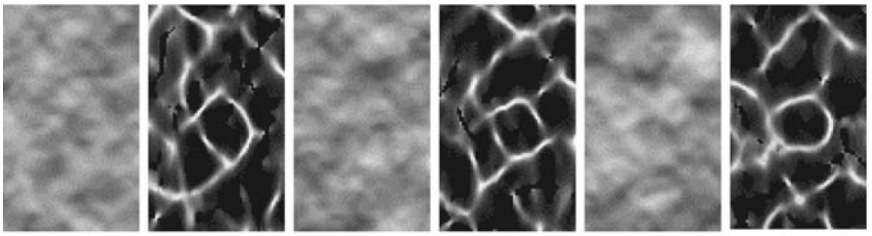


FIGURE 5. Surface enhancement of individual microtubules viewed in cross-section. From left to right, each original cross-sectional slice is paired with the corresponding enhanced version, showing microtubule boundaries in the center of the field. While the microtubule boundaries are fuzzy in the 2D cross-sectional slices, the surface-enhancing filter estimates the boundaries in 3D to produce distinct cross-sections. (From Jiang *et al.* (2005), © 2005 IEEE).

have circular or elliptical shape, which can be detected with the Hough transform to initialize the ASM segmentation described in the next section.

4.4. Segmentation of Microtubule Main Body with the Active Shape Model Technique

The segmentation of the microtubule body is performed within the localized volume enhanced with the surface-enhancing filter. Even so, automatic extraction of the microtubule boundaries remains challenging, due to missing features on the microtubule boundaries and overlap by surrounding materials. Therefore, we have adapted ASM for segmentation. With prior shape constraints obtained from training samples, ASM segmentation is robust against confusing image features. In this section, we first introduce the statistical modeling of microtubule cross-sectional contours using a point distribution model. In order to characterize the microtubule boundary reliably, we replace the gray-level appearance model with a boundary strength model obtained from surface-enhancing filtering. We also improve the shape searching by relating the boundary strength to the weight matrix of the ASM fitting criterion. Finally, we incorporate Kalman filtering, to impose a smoothness constraint along the longitudinal direction.

4.4.1. Determining Microtubule Contours with a Point Distribution Model

We model the microtubule cross-sectional contours with a point distribution model that examines the statistical distribution of the boundary points across a given training set (Cootes *et al.*, 1995). The model is built by applying principal component analysis to a set of manually annotated microtubule contours.

Let $\mathbf{x}_i = (x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots, x_{in}, y_{in})$ be a vector describing the n points of the sample shape i in the training set. The mean shape is calculated as:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{9}$$

The covariance matrix is calculated as:

$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{10}$$

The modes of variations of the points of the shape are described by $\mathbf{p}_k$, the unit eigenvector of $\boldsymbol{C}$, such that:

$$\boldsymbol{C}\mathbf{p}_k = \mathbf{e}_k\mathbf{p}_k \tag{11}$$

where $\mathbf{e}_k$ is the kth eigenvalue and $\mathbf{e}_k \geq \mathbf{e}_{k+1}$.

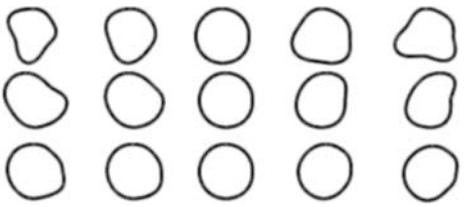


FIGURE 6. The effect of individually varying the first three modes of the microtubule point distribution model. Top row, first mode; middle row, second mode; bottom row, third mode.

Since the majority of the shape variation can be explained by a small number of modes, any shape in the training set can be approximated using the mean shape and a weighted sum of these deviations obtained from the first m modes:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Pb} \tag{12}$$

where $\mathbf{P}$ is a $2\boldsymbol{n} \times \boldsymbol{m}$ matrix whose columns are the unit eigenvectors $\mathbf{p}_k$, and $\mathbf{b}$ is a vector of weights for the eigenvectors. This allows the generation of new shape examples by varying the parameters in $\mathbf{b}$ within suitable limits, so that the new shape is similar to those in the training set. Figure 6 shows an example of plausible microtubule contours generated by individually varying the first three modes of the statistic shape model. It can be seen that these modes describe the roundness of the microtubule contour, and how much distortion is allowed.

4.4.2. Boundary Strength Model and ASM Searching with Adaptive Weights

In ASM, a gray-level appearance model is usually trained to describe image structure around each boundary point, and is then fitted to a shape model of the object in the image. However, for our problem, such a model is not reliable, because the microtubule cross-sectional boundaries are often blurred and connected to background features. Therefore, we replace the gray-level appearance model with a boundary strength model based on the surface-enhancing filtering result, which reliably characterizes the microtubule boundary by exploiting 3D image information.

The ASM searching seeks to find new instances of the modeled object in images, which involves finding the alignment and shape parameters that cause the model to coincide with structures of interest in the image. In our case, to find the microtubule boundary points in the image, we make the algorithm search along the normal lines, in the cross-sectional slices of the model, for the points with maximum boundary strength. Constraints are imposed by the point distribution model.

The shape model **x** is fitted to the image object **X** by minimizing the fitting criterion:

$$\mathbf{E}(\mathbf{X},\mathbf{x};\Omega) = (\mathbf{X} - T(\mathbf{x},\Omega))^T \mathbf{W}(\mathbf{X} - T(\mathbf{x};\Omega)) \tag{13}$$

where T denotes the spatial transformation, including translation, rotation and scaling, Ω denotes the transformation parameters, and **W** is the diagonal weight matrix that we define as:

$$w_i = \frac{F_i}{\sum_{i=1}^{N} F_i} \tag{14}$$

The transformation adjusts the weights adaptively based on the feature prominence at the boundary points. This has practical significance in our problem, because the microtubule boundaries are subject to defects due to missing data caused by the limited angular range over which electron tomographic tilt series can be collected.

4.4.3. Modeling Microtubules in 3D with Kalman Filtering

Kalman filtering is used to improve the ASM operation by imposing longitudinal smoothness along the microtubules, assuming small shape variations among neighboring slices. Across neighboring slices, the Kalman filter recursively estimates the shape parameter using state prediction and state updating with the measurement obtained from the ASM searching. The ASM modified in this way is robust against missing data and outliers present in the kinetochore tomographic volume. Details about the integration of Kalman filtering and ASM can be found in our previous work (Jiang *et al.*, 2004a).

The effect of Kalman filtering is demonstrated in Fig. 7. The results obtained by ASM with Kalman filtering (lower row of Fig. 7) exhibit smooth transitions across the slices, which is more plausible than the unfiltered or manually trace models (see below). Therefore, incorporation of Kalman filtering into the segmentation process reduces sensitivity to missing boundaries or clutter in individual slices, because it utilizes information from neighboring slices to suppress attraction from confusing image features in the current slice.

4.5. Extraction of Microtubule Plus-ends

The heavily cluttered cellular environment makes it difficult to use data-driven deformable models for microtubule plus-end segmentation. It is also not straightforward to model plus-ends using statistical methods, because the plus-ends have various lengths and show large shape variations. On the other hand, the fact that plus-ends are filamentous features

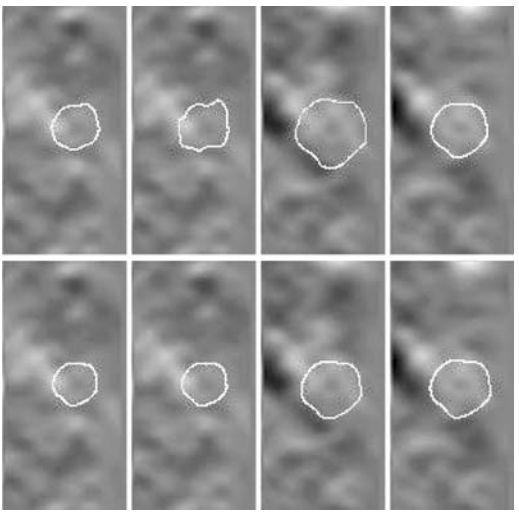


FIGURE 7. Cross-sectional traces demonstrating the effect of Kalman filtering on ASM segmentation. Top row: consecutive cross-sectional slices of a microtubule segmented using ASM processing without Kalman filtering. Bottom row: the corresponding cross-sectional slices of a microtubule segmented using ASM processing with Kalman filtering. Kalman filtering makes segmentation more robust and more consistent between successive cross-sectional slices. (From Jiang *et al.* (2004a), © 2004 IEEE).

connected with the microtubule body suggests the usage of tracing tools for the segmentation task. To this end, we have designed a probabilistic plus-end tracing method augmented with the prior knowledge that protofilaments are either straight lines or curves without inflections. Such curvature restrictions can serve as a useful constraint for tracing. In the following sections, we first describe radial slicing of the localized volume and the initialization of tracing, and we then formulate the probabilistic framework of the tracing.

4.5.1. Radial Slicing of the Localized Volume and Initialization of Tracing

When microtubules transit from the assembly to the disassembly state, the 13 or so parallel protofilaments that make up the microtubule wall adopt a curve conformation (Howard and Hyman, 2003; Fig. 8 left hand panel). As a result, the microtubules appear to be frayed at the ends. Current evidence strongly suggests that the curvature at the ends of protofilaments is planar and variable along the length dimension (Janosi *et al.*, 1998). In the kinetochore volume, the plus-ends tend to have less regular conformations due to the interference from other kinetochore materials. However, planar shapes still characterize the overall curvature of the plus-

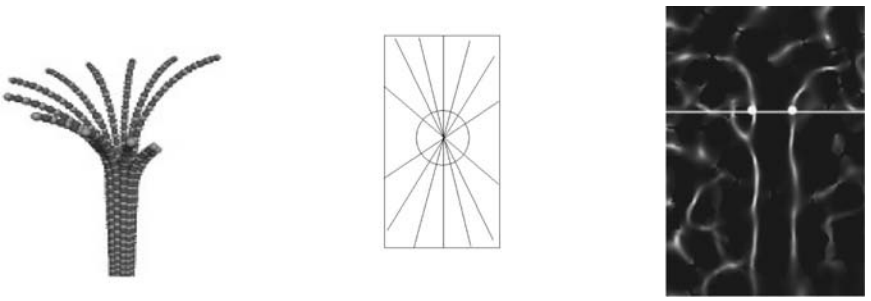


FIGURE 8. Illustration of radial slicing of microtubules. Left: idealized array of protofilaments at the plus-end of a disassembling microtubule. Middle: cross-sectional view of the radial slicing process. The circle represents the microtubule wall, and the cutting planes are perpendicular to the plane of the page. Right: example of a radial slice from an enhanced volume. The starting points for subsequent tracing are indicated by the bright spots on the microtubule wall. (From Jiang *et al.* (2006b), © 2006 IEEE).

ends of protofilaments. It is therefore sufficient to approximate microtubule plus-ends by tracing the curvature in 2D radial slices passing through the microtubule axis. Compared with tracing in 3D, tracing in such 2D planes leads to robust performance, because the adverse influence from other structures is significantly decreased.

The radial slicing is performed around the microtubule, as shown in the middle panel of Fig. 8. To be conservative, we limit the radial slicing within $(-40, 40)^\circ$, to avoid data distorted by the missing angular range. Since there are 13–15 protofilaments for each microtubule, there will be six or seven protofilaments within $(-40, 40)^\circ$. Therefore, the angles -40, -30, -20, $-10, 0, 10, 20, 30$ and 40° are sufficient to sample the protofilaments. We have tested with finer angular slicing, but this did not improve the final plus-end classification.

The center of the radial slicing is chosen as the centroid of the microtubule contours in a region of the cylindrical portion of the microtubule located near to the plus-end, as determined in the prior segmentation steps described above. In each radial slice, the tracing is selected to begin slightly before the end of the cylindrical portion of the microtubule, as illustrated in the right hand panel of Fig. 8. The path for tracing is selected to be along the microtubule edge and towards the plus-end.

4.5.2. The Probabilistic Tracing Framework

The probabilistic tracing method is designed in the framework of particle filtering. Particle filtering techniques have proven to be adept at tracking in the presence of complicated likelihood functions and non-linear dynamics, and it has recently been applied successfully to interactive contour extraction and road detection (Perez *et al.*, 2001).

We represent the plus-end protofilament under tracing as an ordered sequence $x_{0:n} = \{x_0 \cdots x_n\} \in \Lambda^{n+1}$, where x_is are random points in the image plane. The tracing is accomplished by growing this sequence based on prior dynamics and a data model $p(y|x_{0:n})$, where y is the observed image data.

Taking second-order prior dynamics as an example, the *a priori* density on Λ^{n+1} is:

$$p(x_{0:n}) = p(x_{0:1}) \prod_{i=2}^{n} p(x_i | x_{i-2:i-1}) \tag{15}$$

Assuming measurement on $x_{0:n}$ is an independent process, then:

$$p(y|x_{0:n}) = \prod_{u \in \Omega} p_{off}(y(u)) \prod_{u \in x_{0:n}} \frac{p_{on}(y(u))}{p_{off}(y(u))} \tag{16}$$

where Ω denotes the measurement locations in the image plane, p_{on} the likelihood of when u lies on the curve, and p_{off} the likelihood of when u does not lie on the curve.

Using Bayesian rules, we derive the posterior density on Λ^{n+1} up to a multiplicative factor independent of $x_{0:n}$:

$$p(x_{0:n}|y) \propto \prod_{i=2}^{n} p(x_i | x_{i-2:i-1}) \prod_{i=0}^{n} l(y(x_i)) \tag{17}$$

where $l = p_{on}/p_{off}$ denotes the point-wise likelihood ratio, which in this application is determined as the value of the image generated by equation (8), with the corresponding radial slice as shown in Fig. 8, right hand panel.

The tracing relies on the computation of the posterior density based on the following recursion derived from (17):

$$p(x_{0:n+1}|y) \propto p(x_{0:n}|y) p(x_{n+1}|x_{n-1:n}) l(y(x_{n+1})) \tag{18}$$

Although we have analytical expressions for the prior dynamics and the likelihood, the recursion in (18) cannot be computed analytically since there is no closed-form expression for the posterior distributions. Instead, the posterior distribution can be propagated using sequential Monte-Carlo methods (Doucet *et al.*, 2000). In this framework, the posteriors are approximated by a finite set of samples:

$$p(x_{0:n}|y) \approx \sum_{m=1}^{M} w_n^m \delta(x_{0:n} - x_{0:n}^m) \tag{19}$$

where $\delta(\cdot)$ denotes the Dirac delta measure and w_n^m denotes the importance weight attached to sample $x_{0:n}^m$.

At each iteration, the weights are chosen using *importance sampling*. An easy way to draw the samples is to use the prior density as the proposal function. The weights thus are given by the likelihood (Doucet *et al.*, 2000):

$$w_{n+1}^m \propto l\left(y(\tilde{x}_{n+1}^m)\right) \tag{20}$$

where $\tilde{x}_{n+1}^m$ is the position predicted based on prior dynamics as in equation (21) below.

An overview of the probabilistic tracing is as follows:

1. Initialization with starting points
2. Importance sampling step. For each sample $m = 1$ to M, do the following:
   (i) Predict the next position $\tilde{x}_{n+1}^m$ using prior dynamics.
   (ii) Evaluate the weight w_{n+1}^m based on likelihood at $\tilde{x}_{n+1}^m$ as in equation (20)
3. Selection step. Select the most probable next point using the MAP (maximum a posterior) estimate and the normalized importance weights $w_{n+1}^m = w_{n+1}^m \Big/ \sum_m w_{n+1}^m$. Go to step 2.

In our application, we choose the prior dynamics to be specified by the direction change at a given point x_n. Given a normal distribution $p(\theta_n) = N(\theta_n; \bar{\theta}_n, \sigma_\theta^2)$, the next point is predicted as:

$$x_{n+1} = x_n + \mathbf{R}(\theta_n)(x_n - x_{n-1}) \tag{21}$$

where $\mathbf{R}(\theta_n)$ is a rotation matrix. To adapt the prior dynamics to plus-end tracing, a parabola model is fitted locally to a number of traced points preceding the current point x_n. The parabola fitting does not allow inflection points, which makes it suitable to model plus-end protofilaments. A point x_e is then extrapolated along the parabola and $\bar{\theta}_n$ is calculated as the angle between $(x_e - x_n)$ and $(x_n - x_{n-1})$. Compared with tracing along the tangent direction ($\bar{\theta}_n = 0$), tracing along a parabola model is able to take into account the curving direction of the existing traced segments. For straight protofilaments, the parabola fitting will degenerate into line fitting and the tracing will degenerate into tracing along the tangent direction.

The tracing proceeds recursively until the stopping criteria have been met: either the maximum tracing step has been reached, or the value of the image data is lower than the pre-set threshold value.

Figure 9 shows that the tracing successfully captures the plus-end conformation changes in different radial slices in the same volume. This is quite difficult for a human operator because of the tedium of manually performing a tracing process that requires precise localization of the microtubule, careful rotation of the volume and examination of many slices.

4.6. Results and Evaluation

In this section, we present sample segmentation results. After localization of microtubules in the larger volume (Figs 2–4), the microtuble body is segmented in a localized volume using the improved ASM method. Figure 10 compares the automatic and the manual results as seen in a surface

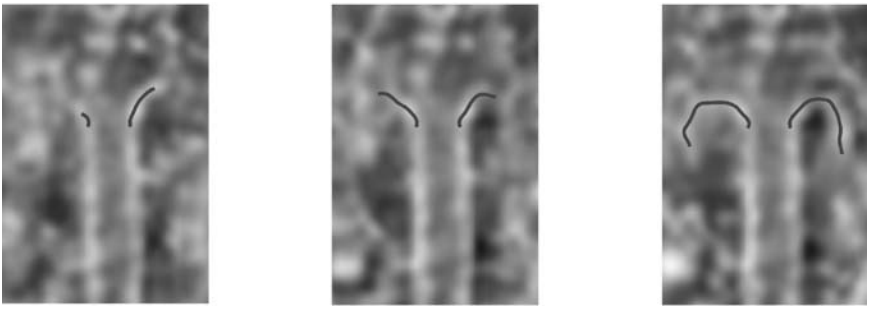


FIGURE 9. Tracing in different radial slices of the same volume captures protofilaments with differing shapes. (From Jiang *et al.* (2006b), © 2006 IEEE).

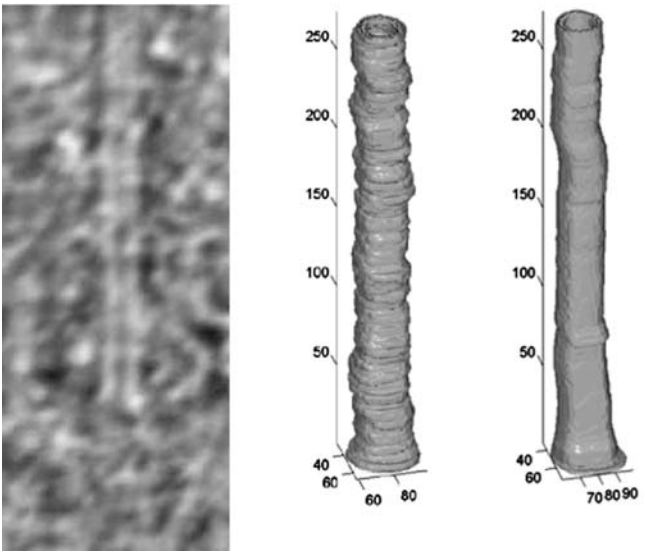


FIGURE 10. Segmentation of a microtubule body. Left: a slice from the original tomographic volume. Middle: surface representation of the results of manual segmentation of the same microtubule. Right: surface representation of the results from automated segmentation. (From Jiang *et al.* (2004a), © 2004 IEEE).

representation. The automated method clearly produces much smoother morphologies. The plus-ends are then obtained using the tracing method with the seeding voxels provided by the segmented microtubule body. Figure 11 shows the segmented microtubules with their plus-ends.

We quantitatively evaluated the microtubule body segmentation, by measuring the overlap ratio between the areas enclosed by each automatically segmented cross-sectional contour and the corresponding manual

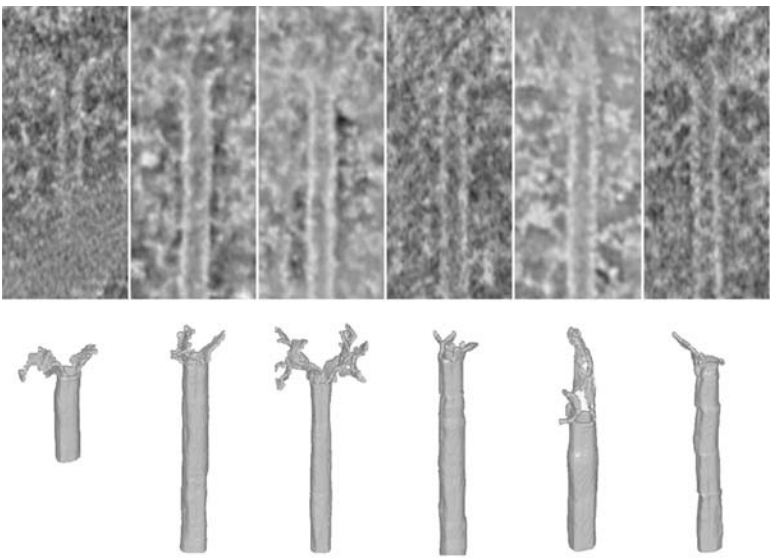


FIGURE 11. Segmentations of microtubules displaying various plus-end conformations. Top row: central slices of individual microtubules as they appear in the original volume. Bottom row: surface views of the corresponding segmented volumes. (From Jiang *et al.* (2005), © 2005 IEEE).

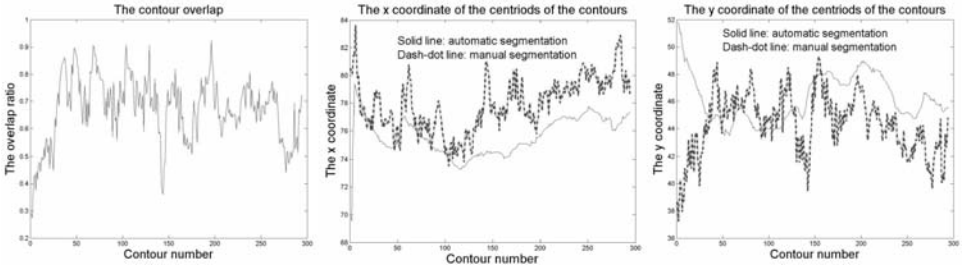


FIGURE 12. Evaluation of the segmentation results, by comparison with manual segmentation. Left: the contour overlap ratio (see text). Middle: plot of the *x* coordinate of each cross-sectional contour as a function of contour number. Right: plot of the *y* coordinate of each cross-sectional contour as a function of contour number. In the middle and right hand plots, the solid line depicts automated segmentation, and the dash–dot curve depicts manual segmentation. (From Jiang *et al.* (2004a), © 2004 IEEE).

result. The contour overlap ratio in Fig. 12 indicates that the automatic results and manual results roughly agree with each other, for most slices. The automatic segmentation results also show smooth changes of the centroid position across the slices, which means that shape variations among neighboring slices are small. In contrast, the manual results have relatively large and irregular changes of the centroid position across the slices, mainly

due to the subjectivity and inconsistency of the human operator, and the lack of tools to incorporate information from neighboring slices. We therefore believe that in general the automatic segmentation is more accurate, consistent and reliable.

Visual inspection of the plus-end segmentation indicates that the automatic results are able to characterize the underlying plus-end structure. Results such as those in Fig. 11 provide valuable information for biological study. In addition, the automatic results often reveal meaningful fine features not obvious to human operators.

Our method is computationally efficient. For example, on a 2.0 GHz PC with Pentium 4 processor and 512M RAM, the automated approach requires about 8 min to extract each microtubule and the associated plus-end, while it takes hours to trace contours manually on microtubule cross-sections.

5. SEGMENTATION OF MEMBRANE STRUCTURES

Membranes are vital because they form a boundary to contain cells while maintaining specific types of communication with the outside world. Membranes also serve to compartmentalize important processes and events within cells. Finally, membranes are vital for the exchange of material between the cell and the outside world (uptake and excretion). For this reason, membrane structures are of great interest to cell and structural biologists, and are frequently segmented from electron tomographic reconstructions. For an idealized membrane, one possible tool for this segmentation task is the use of deformable contours as an energy minimization problem based on image gradient. However, difficulties arise when deformable contours are applied to the membranes in an electron tomographic volume: they are often attracted to local energy minima due to the crowded cellular environment. Application of a region-based deformable contour method is one way to avoid the problems of boundary-based methods, by considering boundary extraction as the minimization of region-based energy functions (Chan and Vese, 2001; Samson *et al*., 2000). However, region-based deformable contouring methods tend to overlook boundary continuity, which may result in inaccurate segmentation due to lack of image gradient information.

To overcome these problems, we have adapted a deformable contour method for membrane segmentation, by integrating both boundary and region information into one segmentation framework (Wang *et al*., 2004). Prior knowledge about the region-based image characteristics, such as texture and homogeneity within a given object in the image, and heterogeneity across object boundaries, is incorporated as an extra constraint to the existing boundary-based deformable contour formulation. The contour energy minimization problem is then formulated as the search for an energy minimum contour with its interior satisfying certain region-based criteria.

In the following subsections, we first introduce the basics of a deformable contour method that incorporates constrained optimization, and we then apply the method to the segmentation of membrane structures.

5.1. Deformable Contour with Constrained Optimization

A deformable contour method performs segmentation by deforming a closed contour toward a salient object boundary. An energy function is associated with the candidate contour to encode some general characteristics of the object boundary. Given a closed contour $C(s,t)$ enclosing region $\Omega_c(t)$ at time t, the target boundary is obtained by minimizing the energy function that can be defined as:

$$\boldsymbol{E}(\boldsymbol{C}(\boldsymbol{s},\boldsymbol{t})) = \frac{1}{\oint_c \boldsymbol{ds}} \left\{ \oint_c \frac{1}{1+|\nabla \boldsymbol{G} * \boldsymbol{I}(\boldsymbol{x}(\boldsymbol{s}),\boldsymbol{y}(\boldsymbol{s}))|^p} \boldsymbol{ds} \right\} \tag{22}$$

where p = 1 or 2, $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega_c(\boldsymbol{t})$, $\nabla \boldsymbol{G} * \boldsymbol{I}$ is used to obtain a smoothed image derivative with derivative Gaussian kernel, and s is the contour arc length. Although the term in the inner integral is not always stated in this form, it is traditionally based on image gradient only. This often leads to incorrect convergence when a boundary feature is not prominent. We tackle this problem by imposing the above energy minimization with a region-based constraint defined as:

$$\boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y}) \geq \boldsymbol{T}_V \tag{23}$$

where $\boldsymbol{T}_V$ is a threshold value, and $\boldsymbol{D}(\boldsymbol{x}, \mathbf{y})$ is the region constraint function defined by:

$$\boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{B}(\boldsymbol{x}, \boldsymbol{y}) \tag{24}$$

with smoothness based on the point gradient as:

$$\boldsymbol{A}(\boldsymbol{x},\boldsymbol{y}) = 1 \Big/ \left(1 + |\nabla G * I(\boldsymbol{x},\boldsymbol{y})|^2\right) \tag{25}$$

and smoothness based on deviation from averaged region intensity as:

$$B(\boldsymbol{x},\boldsymbol{y}) = \exp(|I(\boldsymbol{x},\boldsymbol{y}) - I_0| / \sigma) \tag{26}$$

where $\boldsymbol{I}_0$ is the average intensity over the enclosed region, and σ is a threshold value selected based on the intensity distribution. While $\boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})$ can be used to model a general class of region-based image information, it is considered here as a special case for homogeneity measurement of smoothness at an interior point. A large value of the gradient and a large deviation of the point gray value from $\boldsymbol{I}_0$ produce a small value of $\boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})$

It can be seen that the smoothness constraint is based on image information at both the local level and the regional level. The local minima of the energy minimization can be removed effectively by combing such region-based constraints with the deformable contour.

We solve the constrained optimization problem with an evolutionary strategy to deform $C(s, t)$ until an optimum is reached. An analog to biological evolution, evolutionary optimization is a recursive process in which a population of parents and mutants recombine to generate a large population of offspring (Buck *et al.*, 1997). These offspring are then evaluated according to a fitness function, and a best subset of offspring is selected to replace the existing parents. In our case, a population of contour individuals evolves under the regional smoothness constraint with stochastic perturbations. Contour energy serves as the fitness function and is used to evaluate and select the contour individuals. Based on this optimization strategy, the segmentation algorithm can be described as:

1. Enter the Gaussian filtered image, initial interior location, estimated average intensity, threshold values, number of multiple candidate contours to generate and number of evolution iterations.
2. Initialize the multiple candidate contours and set the iteration number to zero.
3. For each of the candidate contours: (i) use level set or fast marching to solve the converged contour based on equation $\partial \boldsymbol{C}(\boldsymbol{s}, \boldsymbol{t})/\partial \boldsymbol{t} = (\boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{T}_v - \boldsymbol{t}\boldsymbol{N}_G(0,1))\vec{\boldsymbol{n}}$, where $\vec{\boldsymbol{n}}$ denotes the outward unit normal of the curve $\boldsymbol{C}$, t is time, $\boldsymbol{N}_G$ is a Gaussian perturbation and τ is a parameter modulating the perturbation; and (ii) recalculate the average intensity based on the current enclosed region, and calculate the value of the energy defined in equation (22).
4. If the iteration number is not reached, select 50% of the contours with the lowest energy level, duplicate each of them, and go to step 3. Otherwise, stop the iteration.
5. The contour with the minimum energy level is the solution contour.

5.2. Experimental Results

The constrained deformable contour method is applied to the segmentation of membrane structures from electron tomographic volumes. Figure 13 demonstrates application of the constrained deformable contour method on a low SNR image. Although the membrane features of the cell components are corrupted by noise and blurred by other cellular components, the constrained contour method is able to recover the membranes successfully, by using the region-based intensity and texture information. Figure 14 shows that the constrained deformable contour method can adapt to various object shapes and sizes, from large round objects to small elongated objects.

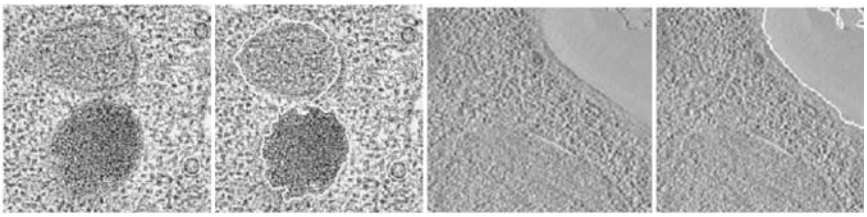


FIGURE 13. Examples of membrane segmentation despite poor boundary features. Each original image is paired with the segmented version.

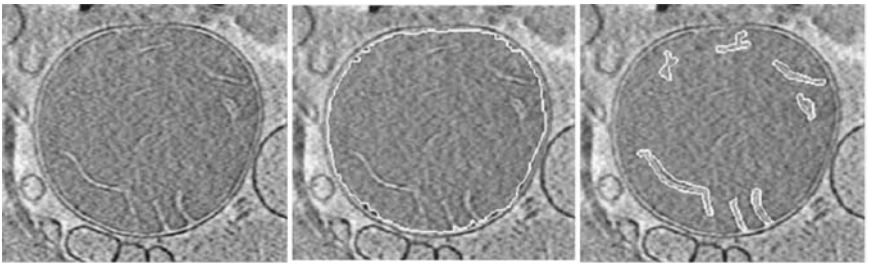


FIGURE 14. Segmentation of mitochondrial membranes. Left: original electron microscopic image. Middle: the same with overlay of segmented inner membrane. Right: the same as the middle image with overlay of segmented mitochondrial cristae.

6. SUMMARY AND CONCLUSIONS

Although electron tomography has opened up new possibilities in the 3D imaging of cellular structures, with an accuracy that was previously unachievable, interpretation of the tomographic volume remains a challenging task, due to the quality of the acquired data and the dense packing and overlap of cell materials in the 3D reconstruction. One important task in electron tomography is therefore to extract the cell components from the volume data, as a prerequisite for both qualitative and quantitative biological studies.

To overcome the low efficiency and subjectivity of manual segmentation, we have described how the utilization of prior knowledge can enable robust and efficient automated segmentation in the presence of noise and clutter. In particular, we have emphasized two aspects concerning the importance of appropriate modeling when using prior knowledge. First, it is often necessary to decompose the prior knowledge into multiple levels, and implement these in segmentation techniques. Secondly, the prior knowledge should be incorporated in such a way that the segmentation remains sufficiently flexible so as to reflect object shape variability faithfully.

We have demonstrated this methodology with two example applications. We first presented the segmentation of microtubules and their associated plus-ends from tomographic volumes of kinetochores. A hybrid approach was designed to perform automated image enhancement and segmentation. In image enhancement, a transform domain technique and a spatial domain technique were combined to exploit the tubular geometrical property of microtubules at the global level and at the local level. In segmentation, the microtubule body was extracted by modeling its cross-sectional contour statistically and imposing longitudinal smoothness with Kalman filtering. Starting from the end of the microtubule body, the plus-ends are extracted based on the spatial connectivity and gray-level similarity between plus-ends and microtubule in the probabilistic framework of particle filtering.

We then presented the segmentation of membranes using a deformable contour with constrained optimization. In this method, we introduced the regional information as an extra constraint, into the existing boundary-based deformable contour. The incorporation of such an object-level gray-level appearance helped to overcome the local minima during the contour energy minimization, and thereby improved the segmentation performance. In both example applications, the accompanying experimental results indicated that a prior knowledge-based method is a promising approach for the segmentation of cell components from electron tomographic volumes, for which, although the data tend to be noisy, object models are often available.

ACKNOWLEDGEMENTS

The authors thank Adriana Verschoor, Christian Renken and Jeffery Bentley for helpful comments regarding the manuscript. We gratefully acknowledge support for model-based segmentation from the National Institute of Health grants GM06627003 (to B.F.M.) and RR01219 (to Joachim Frank, in support of the National Resource for Visualization of Biological Complexity at the Wadsworth Center). Parts of the study also made use of the Wadsworth Core Facility for Electron Microscopy.

REFERENCES

Aylward, S. and Bullitt, E. (2002). Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction. *IEEE Trans. Med. Imag.* **21**:61–75.

Babu, S., Liao, P. C., Shin, M. C. and Tsap, L. V. (2004). Towards recovery of 3d chromosome structure, *IEEE Workshop on Articulated and Nonrigid Motion ANM2004 (in conjunction with CVPR'04)*. Proceedings on a CD-ROM, Washington, DC, June 2004.

Bajaj, C., Yu, Z. and Auer, M. (2003). Volumetric feature extraction and visualization of tomographic molecular imaging. *J. Struct. Biol.* **144**:132–143.

Ballard, D. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recogn.* **13**:111–122.

Bartesaghi, A., Sapiro, G., Lee, S., Lefman, J., Wahl, S., Subramaniam, S. and Orenstein, J. (2004). A new approach for 3D segmentation of cellular tomograms obtained using three-dimensional electron microscopy. In *Proceedings of the 2nd IEEE International Symposium on Biomedical Imaging*, pp. 5–8.

Behiels, G., Vandermeulen, D., Maes, F., Suetens, P. and Dewaele, P. (1999). Active shape model-based segmentation of digital X-ray images. In *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, 1679, MICCAI '99, pp. 128–137.

Böhm, J., Frangakis, A. S., Hegerl, R., Nickell, S., Typke, D. and Baumeister, W. (2000). Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proc. Natl Acad. Sci. USA* **97**:14245–14250.

Buck, T., Hammel, U. and Schwefel, H. (1997). Evolutionary computation: comments on the history and current state. *IEEE Trans. Evol. Comput.* **1**:3–17.

Chan, T. and Vese, L. (2001). Active contours without edges. *IEEE Trans. Image Processing* **10**:266–277.

Cootes, T. F., Taylor, C. J. and Cooper, D. H. (1995). Active shape models—their training and application. *Comput. Vis. Image Understand.* **61**:38–59.

Doucet, A., de Freitas, N. and Gordon, N. (2000). *Sequential Monte Carlo Methods in Practice*. Springer, Berlin.

Duta, N. and Sonka, M. (1998). Segmentation and interpretation of MR brain images: an improved active shape model. *IEEE Trans. Med. Imag.* **17**:1049–1067.

Elder, J. H., Krupnik, A. and Johnston, L. A. (2003). Contour grouping with prior models. *IEEE Trans. Pattern Anal. Machine Intell.* **25**:661–674.

Elder, J. H. and Zucker, S. W. (1996). Computing contour closure. *Proceeding of the 4th European Conference on Computer Vision 1*. Cambridge, UK, pp. 399–412.

Frangakis, A. S., Böhm, J., Forster, F., Nickell, S., Nicastro, D., Typke, D., Hegerl, R. and Baumeister, W. (2002). Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proc. Natl Acad. Sci. USA* **99**:14153–14148.

Frangakis, A. S. and Hegerl, R. (2002). Segmentation of two and three dimensional data from electron microscopy using eigenvector analysis. *J. Struct. Biol.* **138**:105–113.

Frangi, A., Niessen, W., Hoogeveen, R., van Walsum, T and Viergever, M. (1999). Model-based quantitation of 3D magnetic resonance angiographic images. *IEEE Trans. Med. Imag.* **18**:946–956.

Frank, J. and Wagenknecht, T. (1984). Automatic selection of molecular images from electron micrographs. *Ultramicroscopy* **12**:169–176.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M. and Leith, A. (1996). SPIDER and WEB: processing and visualization of images ion 3D electron microscopy and related fields. *J. Struct. Biol.* **116**:190–199.

Harlow, M. L., Ress, D., Stoschek, A., Marshall, M. and McMahan, U. J. (2001). The architecture of active zone material at the frog's neuromuscular junction. *Nature* **409**: 479–484.

He, W., Cowin, P. and Stokes, D. (2003). Untangling desmosome knots with electron tomography. *Science* **302**:109–113.

Hessler, D., Young, S. J., Carragher, B. O., Martone, M. E., Lamont, S., Whittaker, M., Milligan, R. A., Masliah, E., Henshaw, J. E. and Ellisman, M.H. (1992). Programs for visualization in three-dimensional microscopy. *Neuroimage* **1**:55–68.

Hessler, D., Young, S. J. and Ellisman M. H. (1996). A flexible environment for the visualization of three-dimensional biological structures. *J. Struct. Biol.* **116**:113–119.

Howard, J. and Hyman, A. A. (2003) Dynamics and mechanics of the microtubule plus end. *Nature.* **422**:753–758.

Jacob, M., Blu, T. and Unse, M. (2002). 3-D reconstruction of DNA filaments from stereo cryo-electron micrographs. In *Proceedings of 2002 IEEE International Symposium on Biomedical Imaging*, pp. 597–600.

Jain, A. K., Zhong, Y. and Lakshmanan, S. (1996). Object matching using deformable templates. *IEEE Trans. Pattern Anal. Machine Intell.* **18**:267–278.

Janosi, I. M., Chretien, D. and Flyvbjerg, H. (1998). Modeling elastic properties of microtubule tips and walls. *Eur. Biophys.* **27**:501–513 .

Jiang, M., Ji, Q. and McEwen, B. F. (2004a). Model-based automated segmentation of kinetochore microtubule from electron tomography. In *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1656–1659.

Jiang, M., Ji, Q. and McEwen, B. F. (2004b). Enhancement of microtubules in EM tomography. Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging, Arlington, VA, April 15–18, 2004, pp. 1123–1126.

Jiang, M., Ji, Q. and McEwen, B. F. (2005). Automated extraction of microtubules and their plus-ends. Seventh IEEE Workshop on Applications of Computer Vision Proceedings, Breckenridge, Colorado, January 5–7, 2005, pp. 336–341.

Jiang, M., Ji, Q. and McEwen, B. F. (2006a). Model-based automated extraction of microtubules from electron tomography volume. IEEE Transactions on Information Technology in Biomedicine, Vol. 10, No. 3, 2006, pp. 608–617.

Jiang, M., Ji, Q. and McEwen, B. F. (2006b). Automated extraction of fine-features of kinetochore microtubules and plus-ends from electron tomography volume. IEEE Transactions on Image Processing, Vol. 15, No. 7, 2006, pp. 2035–2048.

Kremer, J. R., Mastronarde, D.N. and McIntosh, J. R. (1996). Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**:71–76.

Li, Y., Leith, A. and Frank, J. (1997). Tinkerbell—a tool for interactive segmentation of 3D data. *J. Struct. Biol.* **120**:266–275.

Lucic, V., Forster, F. and Baumeister, W. (2005). Structural studies by electron tomography. *Annu. Rev. Biochem.* **74**:833–865.

Mahamud, S., Williams, L.R., Thornber, K. K. and Xu, K. (2003). Segmentation of multiple salient closed contours from real images. *IEEE Trans. Pattern Anal. Machine Intell.* **25**:433–444.

Maiato, H., DeLuca, J., Salmon, E. D. and Earnshaw, W. C. (2004). The dynamic kinetochore microtubule interface. *J. Cell Sci.* **117**:5461–5477.

Marko, M. and Leith, A. (1996). Sterecon—three-dimensional reconstructions from stereoscopic contouring. *J. Struct. Biol.* **116**:93–98.

Marsh, B. J., Mastronarde, D. N., Buttle, K. F., Howell, K. E. and McIntosh, J. R. (2001). Organellar relationships in the Golgi region of the pancreatic beta cell line, HIT-T15, visualized by high resolution electron tomography. *Proc. Natl Acad. Sci. USA* **98**:2399–2406.

McEwen, B. F., Barnard, R. M., Portuese, T. and Hsieh, C. E. (2002). Using electron tomography to determine the interaction between mammalian kinetochores and microtubules. *Proc. Int. Congr. Electron Microsc.* **15**:179.

McEwen, B. F. and Marko, M. (1999). Three-dimension transmission electron microscopy and its application to mitosis research. *Methods Cell Biol* **61**:81–111.

McEwen, B. F. and Marko, M. (2001). The emergence of electron tomography as an important tool for investigating cellular ultrastructure. *J. Histochem. Cytochem.* **49**:553–563.

McIntosh, R., Nicastro, D. and Mastronarde, D. (2005). New views of cells in 3D: an introduction to electron tomography. *Trends Cell Biol.* **15**:43–51.

Monga, O., Benayoun, S. and Faugeras, O. (1992). Using partial derivatives of 3D images to extract typical surface features. In *Proceedings of The Third Annual Conference of AI, Simulation, and Planning in Highway Autonomy Systems*, pp. 225–236.

O'Toole, E., McDonald, K., Mantler, J., McIntosh, J.R., Hyman, A. and Muller-Reichert, T. (2003). Morphologically distinct microtubule ends in the mitotic centrosome of *Caenorhabditis elegans*. *J. Cell Biol.* **163**:451–456.

Perez, P., Blake, A. and Gangnet, M. (2001). Jetstream: probabilistic contour extraction with particles. In *Proceedings of the Eighth IEEE International Conference on Computer Vision* Vol. 2, pp. 524–531.

Perkins, G., Renken, C., Martone, M. E., Young, S. J., Ellisman, M. and Frey, T. (1997). Electron tomography of neuronal mitochondria: three-dimensional structure and organization of cristae and membrane contacts. *J. Struct. Biol.* **119**:260–272.

Rasband, W. and Bright, D. (1995). NIH Image. *Microbeam Anal.* **4**:20–33.

Ray, N., Acton, S. T. and Ley, K. (2002). Tracking leukocytes *in vivo* with shape and size constrained active contours. *IEEE Trans. Med. Imag.* **21**:1222 –1235.

Renken, C., Siragusa, G., Perkins, G., Washington, L., Nulton, J., Salamon, P. and Frey, T. (2002). A thermodynamic model describing the nature of the crista junction: a structural motif in the the mitochondrion. *J. Struct. Biol.* **138**: 137–144.

Rogers, M., Graham, J. and Malik, R. A. (2000). Exploiting weak shape constraints to segment capillary images in microangiopathy. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention*, pp. 717–726.

Samson, C., Blanc-Feraud, L., Aubert, G. and Zerubia, J. (2000). A level set model for image classification. *Int. J. Comput. Vis.* **40**:187–197.

Scorrano, L., Ashiya, M., Buttle, K., Oakes, S. A., Mannella, C. A. and Korsmeyer, S. J. (2002). A distinct pathway remodels mitochondrial cristae and mobilizes cytochrome c during apoptosis. *Dev. Cell* **2**:55–67.

Segui-Simarro, J. M., Austin, J. R., White, E. A. and Staehelin, A. (2004). Electron tomographic analysis of somatic cell plate formation in meristematic cells of Arabidopsis preserved by high-pressure freezing. *Plant Cell* **16**:836–856.

Sethian, J. A. (1996). A marching level set method for monotonically advancing fronts. *Proc. Natl Acad. Sci. USA* **93**:1591–1595.

Sethian, J. A. (1999). *Level Set Methods and Fast Marching Methods*, 2nd edn. Cambridge University Press, Cambridge.

Smyth, P. P., Taylor, J. and Adams, J. E. (1996). Automatic measurement of vertebral shape using active shape models. In *Proceedings of the British Machine Vision Conference*, pp. 705–714.

Sosinsky, G. E., Deerinck, T. J., Greco, R., Buitenhuys, C. H., Bartol, T. M. and Ellisman, M. H. (2005). Development of a model for microphysiological simulations: small nodes of Ranvier from peripheral nerves of mice reconstructed by electron tomography. *Neuroinformatics* **3**:133–162.

Staib, L. H. (1990). Parametrically Deformable Contour Models for Image Analysis, PhD Thesis, Yale University, New Haven, Connecticut.

Staib, L. H. and Duncan, J. S. (1992). Boundary finding with parametrically deformable models. *IEEE Trans. Pattern Anal. Machine Intell.* **4**:1061–1075.

Tregear, R. T., Reedy, M. C., Goldman, Y. E., Taylor, K. A., Winkler, H., Armstrong, C. F., Sasaki, H., Lucaveche, C. and Reedy, M. K. (2004). Cross-bridge number, position, and angle in target zones of cryofixed isometrically active insect flight muscle. *Biophys. J.* **86**: 3009–3019.

Volkmann, N. (2002). A novel three dimensional variant of the watershed transform for segmentation of electron display maps. *J. Struct. Biol.* **138**:123–129.

Wang, S., Kubota, T. and Siskind, J. M. (2003). Salient boundary detection using ratio contour. In *Neural Information Processing Systems Conference* (NIPS). Vancouver, Canada, pp. 1571–1578.

Wang, X., He, L. and Wee, W. G. (2004). Deformable contour method: a constrained optimization approach. *Int. J. Comput. Vis.* **59**:87–108.

Worring, M., Smeulders, A. W. M., Staib, L. H. and Duncan, J. S. (1996). Parameterized feasible boundaries in gradient vector fields. *Comput Vis. Image Understand.* **63**:135–144.

Yu, T. P. Y., Stoschek, A. and Donoho, D. L. (1996). Translation- and direction-invariant denoising of 2D and 3D images: experience and algorithms. In *Proceedings of the SPIE, Wavelet Applications in Signal and Image Processing* IV:2825, pp. 608–619.

Yu, Z. and Bajaj, C. (2004). Picking circular and rectangular particles based on geometric feature detection in electron micrographs. *J. Struct. Biol.* **145**:168–180.

Zhu, Y., Carragher, B., Mouche, F. and Potter, C. S. (2003). Automatic particle detection through efficient Hough transforms. *IEEE Trans. Med. Imaging*, **22**:1053–1062.

14

Motif Search in Electron Tomography

Achilleas S. Frangakis and Bimal K. Rath

1. INTRODUCTION

Cryoelectron tomography aims to act as an interface between two levels of 3D imaging: *in vivo* cell imaging and techniques achieving atomic resolution (e.g., X-ray crystallography). This most likely will happen through a computational motif search by mapping structures with atomic resolution

Achilleas S. Frangakis • EMBL, European Molecular Biology Laboratory, Meyerhofstr. 1, D-69117 Heidelberg, Germany

Bimal K. Rath • Wadsworth Center, Empire State Plaza, Albany, NY 12201-0509, USA

into lower-resolution tomograms of cells and organelles. There exist a large variety of pattern recognition techniques in engineering, which can perform different types of motif search. This chapter will focus on cross-correlation techniques, which aim to identify a motif within a noisy 3D image (the tomogram or the 3D reconstruction). Generally, the success of the cross-correlation approach depends on the resolution of the tomograms, the degree of corruption of the motif by noise as well as the fidelity with which the template matches the motif. For maximal detection signal, the template should have the same impulse response as the motif, which in this case is the macromolecule sought. Since the noise in the tomogram cannot be significantly decreased after data recording, the task of designing an accurate template reduces to the determination of the precise parameters of the image recording conditions, so that the searched motifs may be modeled as accurately as possible.

The range of algorithms available in electron microscopy for the computational search of different motifs is still significantly wider for 2D than for 3D images. The obvious reason is that only lately the quality of the tomograms has improved to a resolution where such a search is becoming sensible. A variety of these motif search algorithms, which were developed for high-throughput needs of single-particle cryoelectron microscopy, were compared, and their performance was assessed comprehensively (Zhu *et al*., 2004). Among those 2D pattern recognition techniques, the locally normalized cross-correlation approach proved to be the most robust and reliable technique (Roseman, 2004).

The feasibility of these cross-correlation based motif search techniques has also been demonstrated on 3D data and led to the unambiguous detection of known macromolecular structures encapsulated in vesicles (Frangakis *et al*., 2002). Furthermore, it has been demonstrated that template-matching techniques also perform reasonably well on tomograms of organelles, as shown in the localization of ryanodine receptors attached to the membrane of sarcoplasmic reticulum vesicles (Rath *et al*., 2003). However, there is still a large need and potential for improvements to develop good, quantitative detection schemes that will allow a comprehensive evaluation and cross-validation of the results.

In this chapter, a short overview of the cross-correlation techniques will be given, followed by feasibility studies with synthetic data. Some applications on real data will be presented, followed by a brief discussion on the potential and perspectives of the motif search approaches.

2. CREATION OF TEMPLATES

The generation of the templates is a difficult task, since ideally they should look as similar as possible to the motif contained in the tomogram. High-resolution structures derived from various techniques can be modi-

fied in order to be used as templates for the motif search. However, alternatively, a feature contained in the tomogram itself can be used for this purpose. This approach is the ultimate solution if no other information on the structure and identity of the macromolecule is available.

A flowchart depicting the path from an X-ray structure to a simulated 3D electron microscopic image, such that it can be used as a template, is presented in Fig. 1. X-ray structures, which provide the bulk of atomic resolution structures, are modified in a way depending on the contrast transfer function (CTF), which in turn depends on the recording parameters of the data (e.g. defocus, thickness of the samples, sampling increment, amplitude contrast ratio, etc.) and transformed into a lower resolution density map. The 3D Coulomb potential distribution is generated by interpolating the atomic *Z*-numbers into a Cartesian grid with the pixel size of the tomographic data (Stewart *et al.*, 1993). Next, the resulting density is convoluted with the estimated CTF and low-pass filtered with a (smooth) cut-off at the second zero crossing of the CTF. The resulting density map may be represented by an isosurface that overlays the X-ray structure (Fig. 1).

Both approaches, using data from X-ray structures and using subimages from the tomogram, have their individual merits. The merits of the latter are that the template will have (i) the same electron optical parameters and (ii) the same physical environment as the copies of the motif that is being searched for inside the target volume. On the other hand, the disadvantages of the tomographic subimages approach include (i) the low signal-to-noise ratio inherent in an electron tomogram and (ii) the fact that

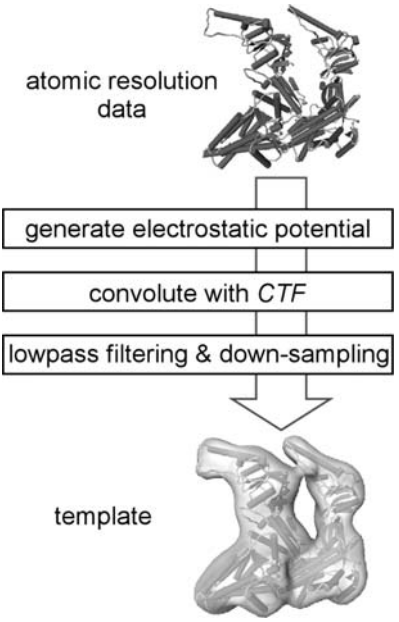


FIGURE 1. Generation of a template: high-resolution data, typically from X-ray structures, are modified to match the electron microscopic structure recorded by the instrument as closely as possible.

the missing angular region (missing wedge) of the template and the repeats of the motif inside the target do not match due to their different relative orientations to the electron beam. The merits of using X-ray structures include (i) the high signal-to-noise ratio of the template and (ii) the absence of distortion due to missing information. On the other hand, since the CTF of the microscope cannot be perfectly measured and since the defocus is locally varying due to the tilt of the specimen, the template and the target will normally have different CTFs.

3. ALGORITHMS

3.1. Cross-correlation based techniques

The cross-correlation coefficient (CCC) is a measure of similarity between two features, in this case a signal x and a template r. The numerator is the unnormalized part of the CCC, and the two terms under the square root in the denominator are the variances of the two correlated objects. The CCC between an image x and the template r, both with the same size R, expressed in one dimension is:

$$CCC = \frac{\sum_{n=1}^{R} (x_n - \bar{x}) \cdot (r_n - \bar{r})}{\sqrt{\sum_{n=1}^{R} (x_n^2) - R \cdot \bar{x}^2} \sqrt{\sum_{n=1}^{R} (r_n^2) - R \cdot \bar{r}^2}}. \quad (1)$$

Here $\bar{x}$ and $\bar{r}$ are the mean values of the image and the template, respectively. For the sake of simplicity, the equation is presented for the 1D case, since the extension to 3D is straightforward.

For the calculation of the normalized cross-correlation function $CC(k)$, the template is shifted in relation to the image and, at each position k, followed by calculation of the CCC. The formula then looks very similar to equation (1), with the difference being that the position of the template r, denoted by its index, is changing, resulting in the following formulation

$$CC(k) = \frac{\sum_{n=1}^{N} (x_n - \bar{x}) \cdot (r_{n-k} - \bar{r})}{\sqrt{\sum_{n=1}^{N} (x_n - \bar{x})^2} \sqrt{\sum_{n=1}^{R} (r_n - \bar{r})^2}} \quad (2)$$

where $k = 1, 2, \ldots, N$, and $\bar{x}$ and $\bar{r}$ are the mean values of x and r, respectively, and N is the number of data points. The calculation of the numerator for every position in the image leads to a very high computational cost, in the order of N^2 operations. However, the expression $CC(k) = \sum_{n=1}^{N} x_n \cdot r_{n-k}$ is almost identical to a convolution product, with the only difference that

the index of the template has reversed sign. Therefore, the cross-correlation function, and similarly also other expressions of this mathematical construction, can be calculated in Fourier space with a computational cost in the order of $Nlog(N)$ operations:

$$CC(k) = F^{-1}\{F\{x\} \cdot F^*\{r\}\} \tag{3}$$

with F being the Fourier and F^{-1} the inverse Fourier transformation, and * indicating the complex conjugate.

In the 3D case, there are six degrees of freedom which need to be probed in order to find the position of highest similarity between the image and the template: three translational and three rotational degrees. The peak of the cross-correlation function, which can be calculated rapidly and in a single step in Fourier space, indicates the potential translational positions producing the highest similarity. However, such an acceleration is not easily achieved for the three rotational degrees of freedom. The angles have to be determined by a grid search, a computationally very expensive process, which, however, can be easily parallelized.

Certain strategies can be applied to avoid the grid search of the angles, which results in a significant acceleration of the speed with which the match is found. One proposal is to use spherical harmonics to represent the signal, similar to the developments for search of electron density in X-ray crystallography (Pavelcik *et al*., 2002). Concrete implementation, which exploits the spherical harmonics representation and virtually reformulates the original search space with three rotational and three translational degrees of freedom into a combination of one translational and five rotational degrees of freedom, has already been achieved by Kovacs and colleagues (Kovacs *et al*., 2003). In this approach, an intensive grid search is only performed on the translational parameter, whereas the remaining five rotational parameters are done by fast, inexpensive correlation in Fourier space. While the extension of this methodology to cryoelectron tomography would require incorporating the missing wedge, this methodology has already been used for the alignment of 2D projection images for single-particle tomography (Cong *et al*., 2003).

3.2. Correlation with non-linear weighting

Cross-correlation performs nicely when the searched 3D image has a uniform local variance. In electron tomographic images, however, the variance of the features varies significantly (e.g. gold beads have a much higher variance than a protein complex). Therefore, the cross-correlation function has to be normalized locally, i.e. the CCCs need to be computed within a defined region R around each voxel k, which belongs to a large volume N (where $N >> R$). This is described by the following equation:

$$CC(k) = \frac{\sum_{n=1}^{N} (x_n - \bar{x}(k)) \cdot (r_{n-k} - \bar{r}) \cdot rect(n-k)}{\sqrt{\sum_{n=1}^{N} (x_n - \bar{x}(k))^2 \cdot rect(n-k)} \sqrt{\sum_{n=1}^{R} (r_n - \bar{r})^2}}$$
$$= \frac{\sum_{n=k+1}^{k+R} (x_n \cdot r_{n-k}) - R \cdot \bar{x}(k) \cdot \bar{r}}{\sqrt{\sum_{n=k+1}^{k+R} (x_n - \bar{x}(k))^2} \sqrt{\sum_{n=1}^{R} (r_n - \bar{r})^2}}, \quad (4)$$

where $k = 1,2,\ldots,N-R-1$; $rect(n-k) = \begin{cases} 1 \text{ for } 1 \le n-k \le R \\ 0 \text{ else} \end{cases}$; and $\bar{x}(k) = \sum_{n=k}^{k+R} (x_n)$ represents the local mean value of x, and $\bar{r}$ the mean value of r. The denominator contains the normalization terms (square root of the variance terms). The term $\sum_{n=k+1}^{k+R} (x_n - \bar{x}(k))^2 = \sigma_x$ represents the local image variance at every position k with reference to the surrounding region denoted by R. Normally, when a spherical mask is used, it can be calculated in real space for a single orientation only, because in this case the mask is rotationally invariant.

Equation (4) describes the derivation of the locally normalized cross-correlation function, for translation coordinates and for a single, fixed angle. This operation has to be performed separately for every angular orientation of the template, and the maximum CCC has to be derived. Since the tomogram already contains the missing wedge as an 'inevitable' filter, the missing wedge does not have to be considered in calculating the numerator. However, it is important that this term be considered in the variance term of the template in the *denominator*. Generally, the template, even at a resolution close to 4 nm, is not perfectly spherical, therefore different Fourier coefficients of the template are filtered out in Fourier space, dependent on the rotational angle, and, consequently, the variance of the template changes for different rotation angles needs to be updated for each set of angles.

3.3. Correlation with asymmetric masks

The use of spherical masks is advantageous because the local variance in the 3D image has to be calculated for just one orientation. However a spherical mask does not necessarily surround the template tightly, and another problem is that using such a mask implicitly assumes that the motif is surrounded by solvent, which has a relatively flat background. This 'void' volume around the template affects the normalization term significantly, potentially leading to incorrect results, especially in very crowded environments. When the motif is in close proximity to other objects, the assump-

tion of a flat background is no longer valid and, therefore, in order to achieve more accurate results, the local variance needs to be calculated for every orientation of the motif in a region under the footprint of the motif. Hence, an asymmetric mask m_R has to be used. For such a mask, the local image variance σ_x can be written as

$$\sigma_x(k) = \sum_{n=1}^{N} (x_n - \bar{x}(k))^2 \cdot m_R(k) = \sum_{n=1}^{N} x_n^2 \cdot m_R(k) - \left(\sum_{n=1}^{N} x_n \cdot m_R(k) \right)^2 \qquad (5)$$

As in the computation of the nominator of the cross-correlation function, the calculation of the summation terms in the above equation in real space represents a significant computational effort, and therefore it is useful to calculate those terms in Fourier space (Roseman, 2003).

A flowchart describing the motif search procedure is presented in Fig. 2. The templates are rotated through all possible combinations of Eulerian angles on a finite grid. This process entails a large amount of computation, since the cross-correlation function has to be calculated for each orientation of the template. Depending on the nature of the application, either a spherical or an asymmetric mask, derived from the shape of the molecule by thresholding the molecule's density map, can be used. This

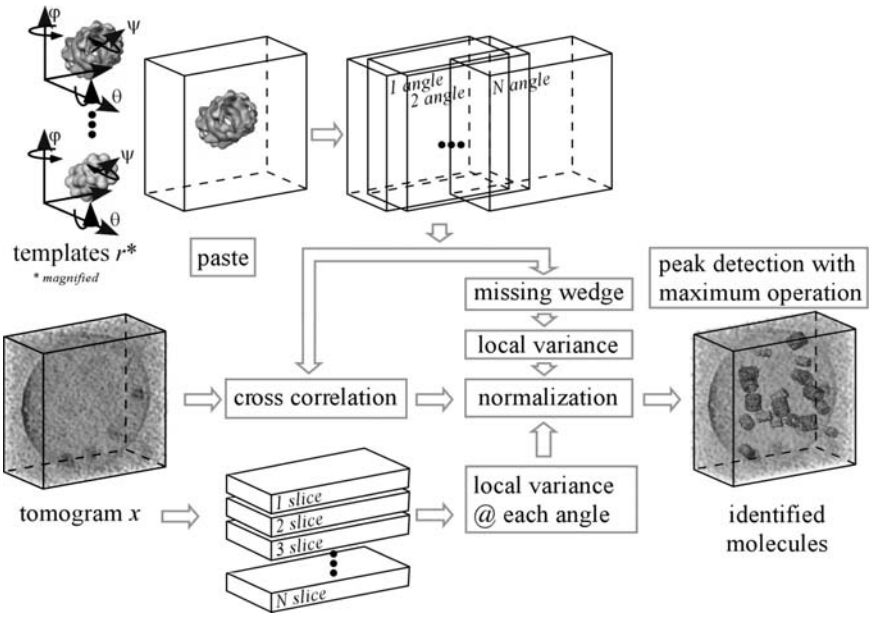


FIGURE 2. Flowchart of the motif search based on cross-correlation. On the upper left side, various templates are used to probe the 3D image (the tomogram in the lower left side) for a motif (see text). (From Frangakis *et al.* (2002), reproduced with permission of the National Academy of Sciences).

asymmetric mask has to be rotated to the same set of Eulerian angles as the template, and the variance of the 3D image under the footprint of the mask m_R at each set of Eulerian angles and at each translational position needs to be calculated. In order to perform the operation in Fourier space, and thus gain computation time, the following procedure is used

1. The corresponding binary mask is first pasted inside a blank 3D image, that has the same size as the searched 3D image. The local variance under the footprint of this mask is computed using equation (5).
2. The template is multiplied by the matching binary mask, the missing wedge is taken care of by using a Fourier filter, and the variance of the resulting image is calculated under the footprint of the mask as a function of the Eulerian angles.
3. The structural signature of the template is pasted inside a blank 3D image of the same size as the searched 3D image (such that the center of the template coincides with that of the 3D blank image). The cross-correlation is calculated, using fast Fourier transform techniques, between the 3D image that contains the motif (i.e. the tomogram) and the blank 3D image into which the structural signature of the template has been pasted.
4. The cross-correlation so derived is then normalized by the corresponding local variance of the tomogram at each position and by the variance of the template. Since the local variance of the tomogram can be calculated only when the template is fully contained inside the borders of the tomogram, the resulting locally normalized cross-correlation array must be delineated to exclude such points where the local standard deviation of the test image cannot be calculated.
5. For each orientation of the template, each position of the delineated locally normalized cross-correlation array is checked for whether it represents a maximum value compared with the delineated locally normalized cross-correlation arrays calculated for orientations already tested.
6. If this is true, the values in an output 3D image containing only the positions and values of the maxima of the cross-correlation array are updated, as well as those in a 3D image containing information about the three Eulerian angles of the corresponding positions.
7. After all possible angle combinations have been used, the locations and orientations associated with the peaks correspond to all putative locations and orientations of the motif that it assumes inside the tomogram, and at those locations the peak height may be used to assign a probability value for the occurrence of the motif.

4. APPLICATIONS

In a cellular context, the ability to identify macromolecules by their 3D intensity signature depends on the resolution of the tomograms. Various studies on both real and synthetic data have explored the feasibility of the

identification. At the current resolution of ~4 nm, only the largest complexes can be identified, and even those only with relatively large ambiguity. We will present the results of some of these studies in order to outline the current state and the potential of the technique.

One of the first feasibility studies was published by Böhm and colleagues (Böhm *et al.*, 2000), who explored the potential of cross-correlation approaches. We will go over the results of this study in some detail. Next, a study will be reviewed that demonstrated the performance of the locally normalized cross-correlation method, by showing the detection and localization of ribosome proteins (Rath *et al.*, 2003). That study showed how accurate and impressive the results can be when the resolution of the underlying 3D map is close to, or better than, 1 nm. Our account of feasibility studies concludes with the detection of very large macromolecular complexes, namely the proteasome and the thermosome. In that study, the complexes were encapsulated in phospholipid vesicles having cellular dimensions such that the recording of the tomographic data set closely resembled that for whole cells. Finally, an application on tomograms of organelles will be presented in which ryanodine receptors were tentatively located within sarcoplasmic reticulum vesicles.

4.1. Feasibility studies

Feasibility studies have the principle advantage that they can be performed under controlled conditions. For cryoelectron tomography, this is an even more significant advantage, since, in contrast with a typical motif search in a cellular tomogram, where negative controls are very difficult, cross-validation can be performed. At a resolution of 4 nm, a distinct cross-correlation peak at the position of a macromolecule, e.g. a thermosome inside a cellular tomogram, does not necessarily mean that the particular object is a thermosome since the resolution is insufficient to allow discrimination of competing structures in that size range. Furthermore, even if a thermosome has been successfully identified, the orientation found in the search is most probably quite imprecise, since high-resolution information required for angular definition is missing (Rosenthal and Henderson, 2003). Characteristic for this situation is that the detection of ribosomes in a cellular context works similarly well with templates that are spheres, cubes or particles with various other geometric shapes. This indicates that clear cross-validation methods need to be developed to enable critical evaluation of the results of detection.

4.1.1. Resolution-Dependence of the Detection Performance

In order to measure the resolution dependence of the detection performance of macromolecular complexes, the following experiment was performed: large macromolecules of similar size and shape (GroEL, ther-

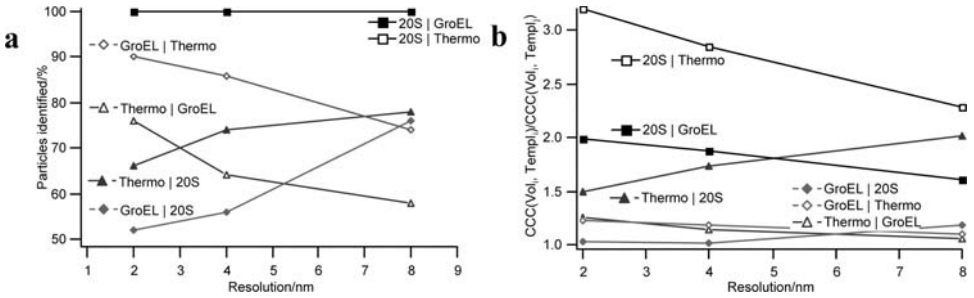


FIGURE 3. Results of identification for synthetic data at different resolutions. In (a), the percentage of particles detected correctly is shown. The detection criterion was as follows: the correlation peak representing perfect detection was obtained by correlating the 3D image of a particle with the 3D image of the 'correct' template (e.g. a thermosome particle with the thermosome template). The result was divided by the correlation peak of a particle with the 'wrong' template (e.g. a thermosome particle with a proteasome template). If the result was >1, the identification was assumed to be correct. In (b), the average of this ratio over all particles is plotted. Due to the difference in diameter, the 20S proteasome can be easily discriminated from the two other particles. To distinguish the thermosome from GroEL, a good resolution is obligatory since the low-resolution information of the two particles is basically identical (same size and shape), whereas the high-resolution data differ due to the distinct symmetries (8-fold versus 7-fold). The opposite is true for the discrimination of GroEL from the 20S proteasome: the two particles are identical in symmetry properties, but different in size, explaining why a resolution of 8 nm is sufficient for a successful identification. (It should be noted that in the plot on the left hand side, the curves for 20S GroEl and 20S Thermo are superimposed on each other.) (From Böhm *et al.* (2000), reproduced with permission by the National Academy of Sciences).

mosome and proteasome) were placed at random positions and orientations inside a 3D image and were altered by addition of noise of different types (e.g. band-pass-filtered noise) (Böhm *et al.*, 2000). Subsequently, the 3D images were searched with different templates, and the discrimination capabilities were evaluated as a function of the signal-to-noise ratio and the type of template used. The results (Fig. 3) demonstrate that certain complexes can be distinguished under certain conditions even at a resolution of ~4 nm. However, the main criterion for the discrimination proved to be the size of the macromolecule. Therefore, the discrimination of macromolecules having different structural signatures, but similar sizes, was not possible. This, however, proved feasible at a resolution of 2 nm.

4.1.2. Performance of the Locally Normalized Correlation as Compared with Globally Normalized Correlation

For finding the similarities between a template and a small region of a larger image, the locally normalized cross-correlation performs much

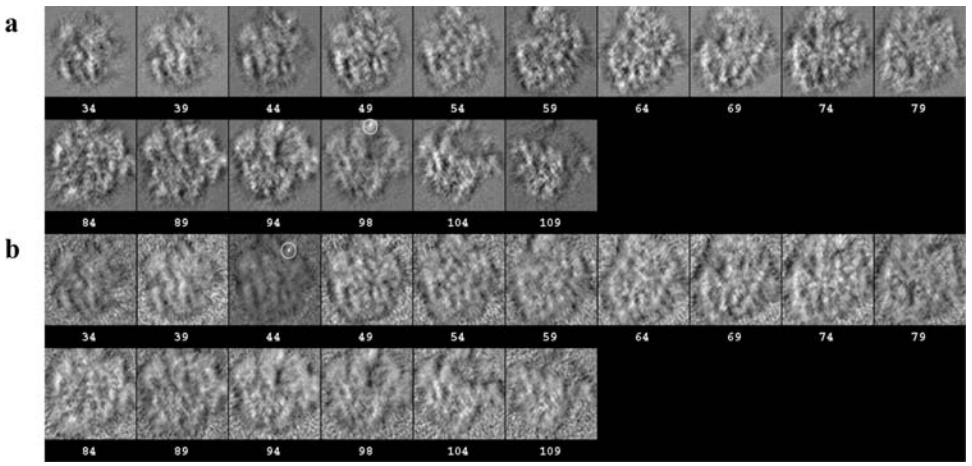


FIGURE 4. *Z*-slices of a globally (a) and locally (b) normalized cross-correlation obtained by searching for protein S2 inside a cryoelectron microscopy density map of the 70S *E. coli* ribosome. In each case, the circle indicates the location of the highest peak (see text). (From Rath *et al.* (2003), reproduced with permission by Elsevier).

better than its globally normalized counterpart. This is because the locally normalized cross-correlation makes use of an individual re-scaling of the template within the underlying subregion of the target. In Fig. 4, a number of z-slices, from a globally normalized cross-correlation array and a cross-correlation array locally normalized with an asymmetric mask, are shown. These slices were obtained while searching for protein S2 inside a 0.8 nm cryoelectron density map of the 70S *Escherichia coli* ribosome. The highest peak in the locally normalized cross-correlation corresponds to the correct position and orientation of the searched S2 motif. In contrast, however, the highest peak in the globally normalized cross-correlation array does not give the correct result. Rather, it is the second highest peak, which is smaller by 0.02 standard deviations than the highest peak, that corresponds to the correct position and orientation of the motif. In contrast, in the locally normalized cross-correlation array, the highest peak is higher by 3.3 standard deviations than the second highest peak, and hence provides highly significant and unambiguous results.

4.1.3. Detection of Macromolecules in Phantom Cells

The evaluation of results from attempts to identify molecules is difficult with cellular data, because a cross-validation is currently not possible. Furthermore, simulations with synthetic data and data from single-particle

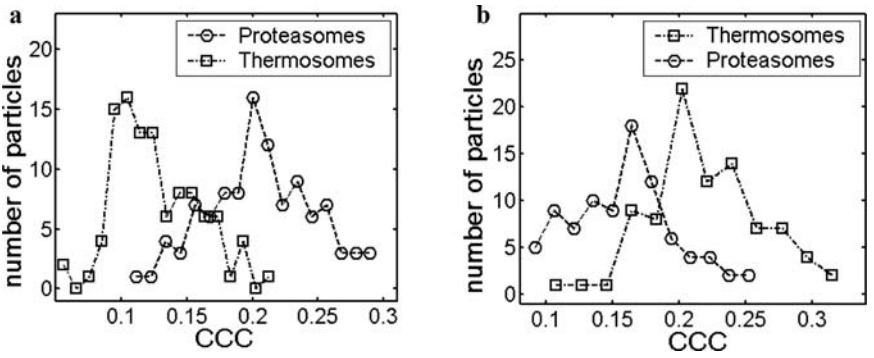


FIGURE 5. Histogram of the cross-correlation between the two X-ray structures in the phantom cell containing (a) proteasomes and (b) thermosomes. (From Frangakis *et al.* (2002), reproduced with permission by the National Academy of Sciences).

analysis resemble the real situation in tomograms only incompletely in terms of—among other properties—resolution, missing wedge effect and defocus in the tilt series used for the reconstruction. An experiment mimicking, as closely as possible, the conditions in cells and at the same time allowing the result to be cross-validated was performed by Frangakis *et al.* (2002). Phospholipid vesicles were filled with a single type of known macromolecular complex, and were visualized by electron tomography. In this way, two issues were addressed at the same time: (i) is the resolution achieved in tomograms sufficient to detect macromolecules?; and (ii) can these complexes be unambiguously identified inside the vesicles, i.e. does the analysis give a negative result when the volume is searched with the 'wrong' template?

In an attempt to address these questions, two vesicles with two different populations of macromolecules were created: in one vesicle, thermosomes and, in the other, proteasomes were encapsulated. The vesicles, independent of their content, were searched with the X-ray structures of both the thermosome and the proteasome, each suitably represented by densities. The detection results, at the positions with the highest correlation peaks, are visualized in Fig. 5. In both cases, the correct contents of a given vesicle could be identified; however, the histograms of the CCCs stemming from the two species do overlap significantly (Frangakis *et al.*, 2002). This is due partially to the very high similarity of the different species of macromolecules. A check of the identification numbers shows that the resolution is just sufficient to distinguish these two large complexes, and that the distinction is mainly based on large-scale features such as size and symmetry. In a vesicle containing a mixture of proteins, as presented in Fig. 6, the definitive evaluation of the identification results is not trivial, since in case of doubt, the unambiguous identification of a particle is not possible.

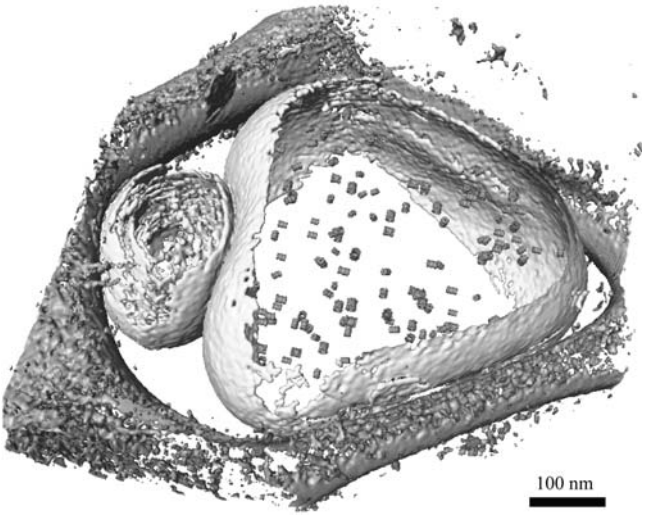


FIGURE 6. Segmented and denoised 3D image of a vesicle containing two types of macromolecules, visualized in different colors. In the isosurface visualization, two vesicles (in beige) surrounded by the carbon film (in gray) are shown. In the larger vesicle, the two different macromolecular complexes were located and positively identified, and were then replaced with their low-pass-filtered X-ray structure. Proteasomes are shown in orange and thermosomes in blue. (From Frangakis *et al.* (2002), reproduced with permission of the National Academy of Sciences). (See color plate)

4.2. Application to real data

To locate ryanodine receptor (RyR) molecules, a cryoelectron tomogram of sarcoplasmic reticulum (SR) vesicles with a resolution of ~50 Å was searched using the locally normalized cross-correlaton. The initial template was created by windowing a recognizable RyR found inside the tomogram itself. The search was performed using a spherical mask and searching exhaustively through all sets of Eulerian angles with 4° increments.

The locations associated with the top 50 peaks, as obtained at the end of a complete search, were labeled inside the tomogram and were mostly found at plausible RyR locations, i.e. along vesicle membranes. Subvolumes around the peaks were windowed and each one was rotated according to the orientation associated with the corresponding peak. Furthermore, using multivariate statistical analysis (MSA), a cluster of eight subvolumes was found in the factor-1 versus factor-2 map in which factor 1 and factor 2 together accounted for 24% of the total intervolume variance. As predicted, these eight RyRs identified were found to be associated with SR vesicles. An average of the windowed and rotated subvolumes was created. In Fig. 7, a z-slice of the labeled tomogram, x-, y- and z- slices of the average subvolume, an enlarged image of a y-slice of the initial template, and the corresponding slice from the average subvolume are shown.

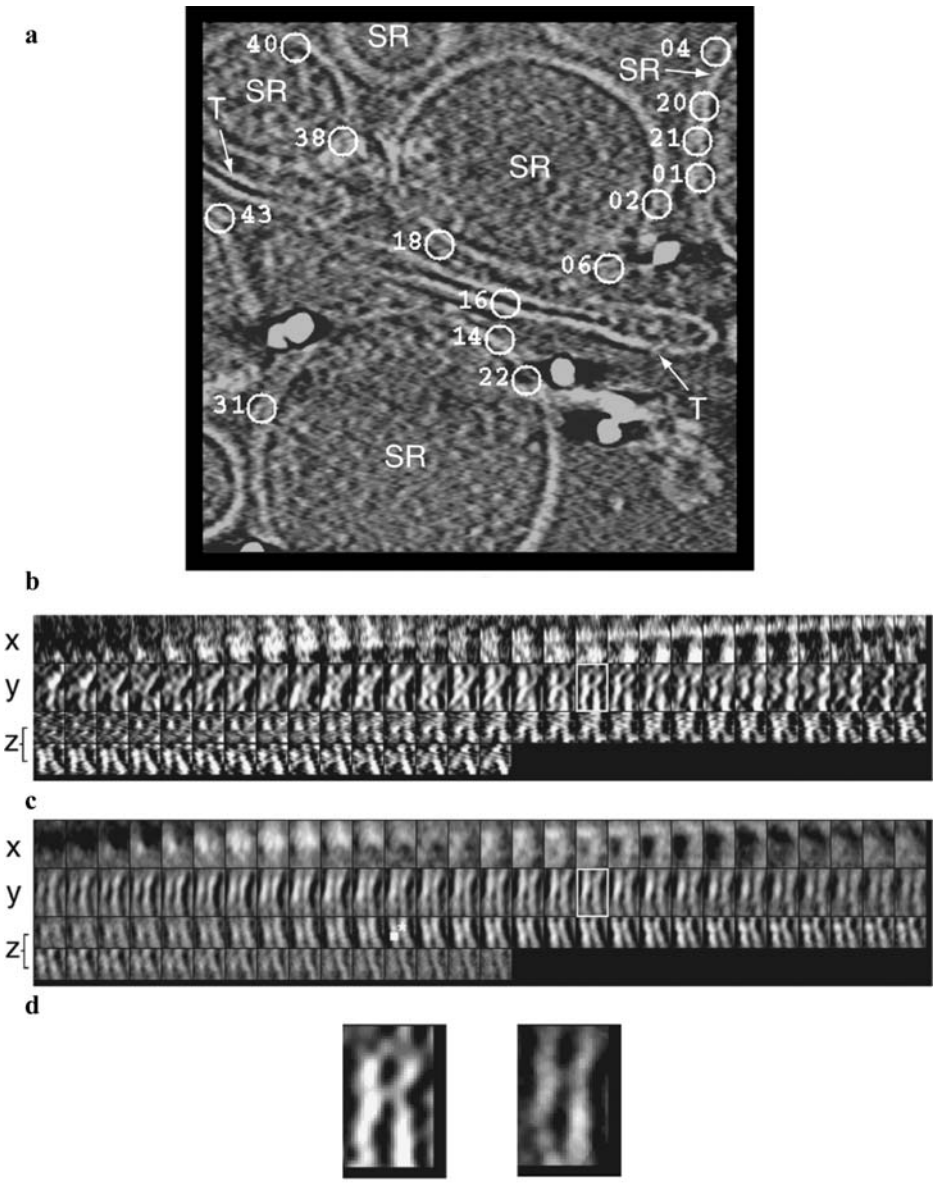


FIGURE 7. Search for ryanodine receptors in a tomographic reconstruction of SR/transverse tubule vesicles. (a) A *z*-slice of the tomogram. 'SR' and 'T' denote vesicles that are derived from sarcoplasmic reticulum and transverse tubule membrane, respectively. Circles indicate the positions of the ryanodine receptors (RyRs) as determined by the locally normalized correlation search method. (b) *x*-, *y*- and *z*-slices of the RyR motif. (c) *x*-, *y*- and *z*-slices representing the average of eight clustered RyR volumes as determined by multivariate statistical analysis. The averaged volume contains two connected structures: one (□) representing the RyR molecule protein and one (*) representing a portion of the attached membrane. (d) Left: an enlarged image of a *y*-slice of the RyR motif (boxed in (b)). Right: the corresponding *y*-slice of the average (boxed in (c)). (From Rath *et al.* (2003), reproduced with permission by Elsevier).

5. DISCUSSION

The success of the motif search depends on the resolution achieved in the tomograms. The quality of the template also plays an important role; however, a good parametric setting should result in the closest possible match with the data in the tomogram. Concerning the choice of motif search algorithms, the main concern is their performance in crowded environments. Both the dense packing of macromolecules in the cytoplasm and the large variability of the macromolecular structures represent difficult image processing problems, which will need to be addressed in the future. However, the prospect for success should be high, since similar problems have been successfully tackled in two dimensions (Zhu *et al.*, 2004).

From the algorithmic point of view, large efforts are being made to develop techniques which provide, in terms of both alleviating computational cost and providing better quality, superior performance compared with the currently used cross-correlation function. Efforts include, but are not limited to, the use of spherical harmonics, to overcome the tedious grid search over all possible rotation angles. The specific handling of the missing wedge, especially for templates that are directly derived from the tomogram itself, represents a significant problem still to be solved.

A serious problem is the lack of procedures to cross-validate the results and perform a negative control of the analysis. In other words, if a certain macromolecule is identified, how can we be certain that this is the correct identification? Is it sufficient to search the cellular tomogram with various and possibly similarly looking macromolecules and verify that a particular one gives the strongest response (i.e. the highest correlation peak)? This issue can be demonstrated on the example of the ribosome, which displays high contrast and can be visualized in cellular tomograms. Cross-correlation produces distinct peaks with the ribosome template; however, it does so also for various templates, including abstract geometric shapes such as spheres and squares. Therefore, in addition to the development of improved recognition algorithms, the development of computational or experimental ways to perform appropriate negative controls is also necessary.

ACKNOWLEDGEMENTS

The authors would like to thank Anja Seybert for a critical reading of the manuscript. The authors would also like to thank Joachim Frank for valuable discussions and comments on the manuscript.

REFERENCES

Böhm, J., Frangakis, A. S., Hegerl, R., Nickell, S., Typke, D. and Baumeister, W. (2000). Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proc. Natl Acad. Sci. USA* **97**:14245–14250.

Cong, Y., Kovacs, J. A. and Wriggers, W. (2003). 2D fast rotational matching for image processing of biophysical data. *J. Struct. Biol*. **144**:51–60.

Frangakis, A. S., Böhm, J., Forster, F., Nickell, S., Nicastro, D., Typke, D., Hegerl, R. and Baumeister, W. (2002). Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proc. Natl. Acad. Sci. USA* **99**:14153–14158.

Kovacs, J. A., Chacon, P., Cong, Y., Metwally, E. and Wriggers, W. (2003). Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta Crystallogr. D Biol. Crystallogr*. **59**:1371–1376.

Pavelcik, F., Zelinka, J. and Otwinowski, Z. (2002). Methodology and applications of automatic electron-density map interpretation by six-dimensional rotational and translational search for molecular fragments. *Acta Crystallogr. D Biol. Crystallogr.* **58**:275–283.

Rath, B. K., Hegerl, R., Leith, A., Shaikh, T. R., Wagenknecht, T. and Frank, J. (2003). Fast 3D motif search of EM density maps using a locally normalized cross-correlation function. *J. Struct. Biol.* **144**:95–103.

Roseman, A. M. (2003). Particle finding in electron micrographs using a fast local correlation algorithm. *Ultramicroscopy* **94**: 225–236.

Roseman, A. M. (2004). FindEM—a fast, efficient program for automatic selection of particles from electron micrographs. *J. Struct. Biol.* **145**:91–99.

Rosenthal, P. B. and Henderson, R. (2003). Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**:721–745.

Stewart, P. L., Fuller, S. D. and Burnett, R. M. (1993). Diference imaging of adenovivus: bridging the resolution gap between X-ray crystallography and electron microscopy. *EMBO J.* **12**:2589–2599.

Zhu, Y., Carragher, B., Glaeser, R. M., Fellman, D., Bajaj, C., Bern, M., Mouche, F., de Hass, F., Hall, R. J., Kriegman, D. J., Ludtke, S. J., Mallick, S. P., Penczek, P. A., Roseman, A. M., Sigworth, F. J., Volkmann, N. and Potter, C. S. (2004). Automatic particle selection: results of a comparative study. *J. Struct. Biol.* **145**:3–14.

15

Localization and Classification of Repetitive Structures in Electron Tomograms of Paracrystalline Assemblies

Kenneth A. Taylor, Jun Liu and Hanspeter Winkler

1. INTRODUCTION

Electron tomography offers opportunities to study structures that are not amenable to 3D imaging by any of the classical methods, such as single-particle reconstruction (Frank, 1996), helical reconstruction (Egelman, 2000; DeRosier and Moore, 1970) or electron crystallography (Glaeser,

Kenneth A. Taylor, Jun Liu and Hanspeter Winkler • Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306-4380, USA

1999) that require either a repetitive structure, or multiple copies of identical structures. Since electron tomography can produce a 3D image of a single copy of a structure, it is finding wide application in cell biology and material science. Paracrystalline specimens constitute another class of structure for which electron tomography can be particularly useful for obtaining detailed 3D images (Taylor *et al*., 1997). Paracrystals (para—Greek prefix meaning faulty) are arrays with various kinds of intrinsic disorder. Spatial averaging of such specimens usually blurs or even erases the disordered component, which may eliminate the functionally interesting feature. For this chapter, we define a paracrystalline specimen as one with partial ordering such that one component of the specimen may be highly regular while another may be irregular due to either low occupancy, lattice irregularity or both.

Striated muscle is one example of a paracrystalline structure found in nature. Striated muscles consist of hexagonal arrays of two types of filaments: thick, myosin-containing filaments, and thin, actin-containing filaments. Interactions between myosin heads and the actin filament are responsible for filament sliding and muscle shortening (Geeves and Holmes, 1999). Muscle is paracrystalline because the filaments themselves have a well-ordered structure, but have a disordered arrangement within the lattice (Squire, 1981), or the interactions between filaments, which usually involves the myosin cross-bridges, are highly variable. Structures that contain actin filaments are often poorly ordered because the actin helix can have variable twist (Egelman and DeRosier, 1992; Egelman *et al*., 1982). However, even in the instances where the actin filament has a well-ordered 28/13 helical structure, such as in insect flight muscle (IFM), the myofibrils display considerable disordering among the myosin cross-bridges. Other types of natural structures with paracrystalline ordering include the cross-linked actin arrays found in microvilli (Tilney *et al*., 1980) or various *in vitro* 2D actin assemblies (Taylor and Taylor, 1994; Taylor *et al*., 2000).

The asynchronous flight muscles of various species of the large water bug *Lethocerus* are perhaps the best ordered muscles in the animal kingdom and are therefore ideal both to study muscle contraction and to develop methods for image classification in 3D. IFM contains an hexagonal array of thick filaments with actin filaments interdigitated between thick filament pairs at pseudodiad positions in the unit cell (Fig. 1b). This filament arrangement differs from vertebrate striated muscle where the actin filaments lie at trigonal positions within the unit cell. The IFM lattice arrangement facilitates the cutting of several types of thin section that have been extremely useful for obtaining information on the arrangement of myosin heads in different muscle states (Fig. 1b). Two kinds of 25 nm thick longitudinal section can be cut parallel to the filament axis (Reedy and Reedy, 1985). One of these, the myac layer (Fig. 1a), contains alternating thick and thin filaments. Another, the actin layer, contains only actin filaments, but these divide evenly into two groups that differ by azimuthal rotation of 60° and axial translation of 12.8 nm. The third type is a 15 nm thick section cut

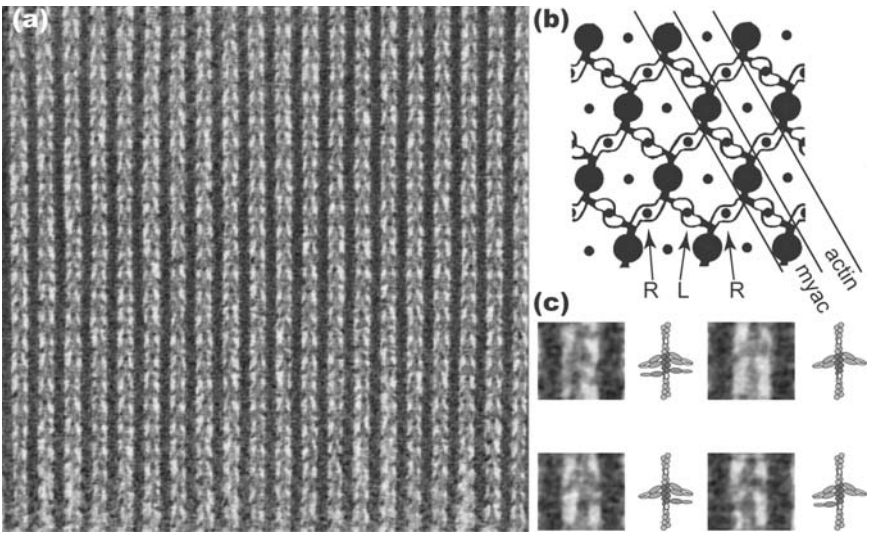


FIGURE 1. Diagram of the muscle lattice and the myac layer. (a) Electron micrograph of a 25 nm longitudinal section of rigor IFM. Protein is black and the embedding medium is white. The vertical rods of density are the thick filament backbones. Between thick filaments is an actin filament decorated with myosin heads, the cross-bridges, that originate from the neighboring thick filaments. (b) Diagram showing the thick and thin filament lattice arrangement and the typical pattern of myosin cross-bridges that occur in a 15 nm thick transverse section. Two types of cross-bridges, designated 'lead' (L) and 'rear' (R), are bound on the actin filaments. The 'myac' layer is a 25 nm longitudinal section containing alternating thick and thin filaments. An 'actin' layer is a 25 nm longitudinal section containing only actin filaments and their bound myosin cross-bridges. (c) Examples of rigor motifs. The top row contains a 'double chevron' on the left and a single chevron on the right. The bottom row contains a pair of incomplete double chevrons, with one 'rear' cross-bridge missing from either side. The rigor cross-bridge lattice comprises a mixture of these motifs spread irregularly throughout the sarcomere. Further motif variation can occur due to head occupancy and effects of lattice constraints on lead and rear cross-bridge structure.

transverse to the filament lattice in rigor muscle and is known as the 'flared-X' formation. Myac layers are ideal for studying cross-bridge formations in different muscle states because they contain all the features of the muscle, namely thick filaments, thin filaments and the connecting cross-bridges, but in a thin section that is easy to analyze. The actin filaments in myac layers are parallel to each other to within a few degrees by virtue of their organization at the muscle *Z*-disk. They have an intrinsic disorder in that they can be rotated randomly by ±180° about the filament axis (Holmes *et al*., 1980). Alignment to correct this disorder is essential if averages revealing actin monomers are to be obtained from the tomogram.

Electron micrographs of rigor IFM myac layers (Fig. 1a) show a repeating pattern of densities bridging the thick with the thin filaments. This pattern, known as the double chevron (Reedy *et al*., 1965), repeats at spacings of 38.5 nm and is produced by the binding of myosin heads to regions

of the thin filament known as 'target zones' (Reedy, 1967; Tregear *et al*., 2004). The matched axial periodicities of myosin and actin filaments are a key aspect of IFM muscle regularity. IFM myosin filaments have the characteristic 14.5 nm axial period typical of myosin filaments, but they also have a 38.5 nm helical period that describes the arrangement of myosin heads around the filament backbone. The actin filaments consist of a double helix of actin monomers with a pitch of 2×38.5 nm. There is a common axial period of 232 nm (6×38.5 nm; 16×14.5 nm) in the muscle. It is the repeating volumes based on the 38.5 nm actin filament half-period that are the object of our classification efforts.

The repeats described in this chapter are defined by the repeating actin filament motif within the muscle lattice. 'Repeats' can be equated with the term 'single-particle' used in 3D reconstruction from projections, but the two are not exactly the same. In single-particle 3D reconstruction, the images derive from different realizations of a common structure that present different orientations. In muscle, the repeats have similar orientations, but have different contents, due to different patterns of cross-bridge attachments to the actin filaments. Thus, the lattice consists not of one repeating motif, but of an unknown number of motifs that may be distributed irregularly. For example, in rigor muscle, there are four main patterns of cross-bridge attachments to actin independent of any differences in the structure of the cross-bridges (Fig. 1c). The double chevron contains a pair of opposed cross-bridges, dubbed 'lead' bridges (Reedy and Reedy, 1985), and an additional pair of opposed cross-bridges, dubbed the 'rear' bridges; single chevrons have no rear bridges, and two types of incomplete double chevrons each have a single rear bridge that can be positioned on either side of the actin filament. Further variations on this main theme can be observed, such as differences in the numbers of myosin heads (one or two) at both lead and rear bridges, and, when a cross-bridge is single-headed, there is the possibility that this head can bind either one of two actin monomers. Finally, the individual myosin heads may differ in structure due to lattice constraints in the muscle fiber. Contracting muscle has potentially many more possible combinations because the myosin heads are distributed over different steps in the actomyosin ATPase cycle.

Intact muscle has presented challenges for 3D imaging. The powerful methods of single-particle 3D reconstruction based on projections of individual molecules or assemblies that have been developed over the years are ineffective. Single-particle methods are most effective when the particles lie in random orientations and are well separated. In the muscle lattice, the repeats lie in preferred orientations and are closely apposed. Because the repeats have a preferred orientation, the object must be tilted to obtain 3D data, and, because the repeats lie in close apposition, structures become superimposed when tilted to high angle.

These problems are all overcome using electron tomography. Tomograms have been obtained of a number of different states of IFM (Chen

et al., 2001, 2002; Liu *et al.*, 2004; Schmitz *et al.*, 1996, 1997; Taylor *et al.*, 1999). However, the tomograms themselves are typically noisy 3D images and, if the information desired involves molecular conformations, averages should be obtained to improve the signal-to-noise ratio. It is also desirable to reduce the highly variable and very numerous structures of myosin cross-bridges obtainable in a single tomogram to a more tractable number of averages. A single tomogram of a myac layer from a muscle half sarcomere usually contains ~500 repeats. Higher resolution, improved signal-to-noise ratio and data reduction are all goals that can be met by using the techniques of multivariate statistical analysis (MSA), in particular correspondence analysis (Frank, 1996). In addition, MSA of 3D volumes presents several unique challenges, including 3D alignment and treatment of the missing wedge (or pyramid), which do not occur in classification of 2D projections.

This chapter will first present approaches to alignment, classification and averaging of 3D repeats in IFM. This system is in many ways ideal for development of these protocols. After describing the IFM work, other biological systems will be discussed, in particular systems in which repeats are not regularly arranged, and some of the clever solutions developed to address the problems of alignment and classification.

2. MISSING WEDGE

The majority of tomograms, but especially tomograms of frozen-hydrated specimens, are computed from single-axis tilt series. Because of limitations on the highest permissible tilt angle in the electron microscope, single-axis tilt series have a missing wedge that is usually not larger than ±30° and usually not smaller than ±20°. In principle, this can have a strong effect on the ability to align 3D volumes and to identify repeating motifs within a large population. The underlying goal of 3D volume classification is to extract structural heterogeneity. A projection of a molecule has no information on the location of features along the projection direction, but the 3D image of that same molecule does. Moreover, for the same total electron dose, the 3D image has the same signal-to-noise ratio as its projection (Hegerl and Hoppe, 1976; Hoppe and Hegerl, 1981; McEwen *et al.*, 1995), and has higher contrast. However, the missing wedge, which is an entity defined in Fourier space, can affect the relative weighting of features in the image and therefore could bias the alignment and classification in favor of grouping unit cells by orientation rather than structure (see also Chapter 14 of this volume).

The potential for the missing wedge to affect 3D volume classification can be illustrated with classification of 2D projections of identical particles that differ only in orientation. In Fourier space, the transforms of these 2D images consist of a single plane that intersects the origin of the transform,

i.e. a central section. With respect to the 3D transform, these central sections can be thought of as having a ±90° missing wedge, i.e. far more of the 3D transform is missing than is measured. If the particles have a thickness, t, their transform varies slowly over a distance of $1/t$. Similarly oriented particles will have in common an amount of data that depends on the angle between planes of finite thickness $1/t$. In this case, classification will place those particles that are closely oriented in space into a single class. In other words, classification clusters the particles by missing wedge orientation. In fact, this is exactly the desired result in single-particle reconstruction from a homogeneous population of particles. Problems arise if the particle set is heterogeneous with respect to structure. Then the classification becomes much more difficult because the individual projections may have too little information to distinguish structure differences from orientation differences (Frank, 1996).

In 3D classification, far more of the transform is measured than is missing, so the structure is much better defined. This makes 3D classification from tomograms potentially powerful for defining structural heterogeneity. MSA of 3D repeats in muscle can be more aptly described as the act of binning a continuum of structures placed around a pair of common structural elements. Inter-repeat variance in a structure like this can also come from misalignment, but minimizing alignment errors is part of the process of iterating to a consistent result. The actual goal is characterizing structure variability. However, it is not known at the moment how much of an effect the missing wedge will have on classification of the repeats; will repeats cluster based on the orientation of the missing wedge, as they do for projections, or will they cluster based on similarities in structure irrespective of the orientation of the missing wedge?

Consider three cases of volume pairs obtained from tomograms with differing orientations of the missing wedge after alignment to a common coordinate frame. The first case (Fig. 2a), which is the most optimal, would have the volumes already partially oriented as if extracted from a paracrystalline specimen so that their missing wedges superimpose. This is the case for volumes (repeats) extracted from IFM tomograms. The second case (Fig. 2b) would occur for a population of volumes that differ only by rotation about the normal to the specimen plane. This might occur with volumes selected from several different tomograms of paracrystalline 2D arrays. Each volume will have in common the in-plane projection. The alignment problem is primarily that of determining the in-plane rotation and the 3D displacement. In this case, the question is whether the 3D volumes would cluster in the same way that their 2D in-plane projections would cluster, irrespective of any differences in the positioning of features along the z-axis. The third case (Fig. 2c) is more general and has the particles randomly oriented in the specimen and, hence, after alignment, their missing wedges are randomly oriented. This is the least favorable case because both alignment and classification could be affected by the missing wedge.

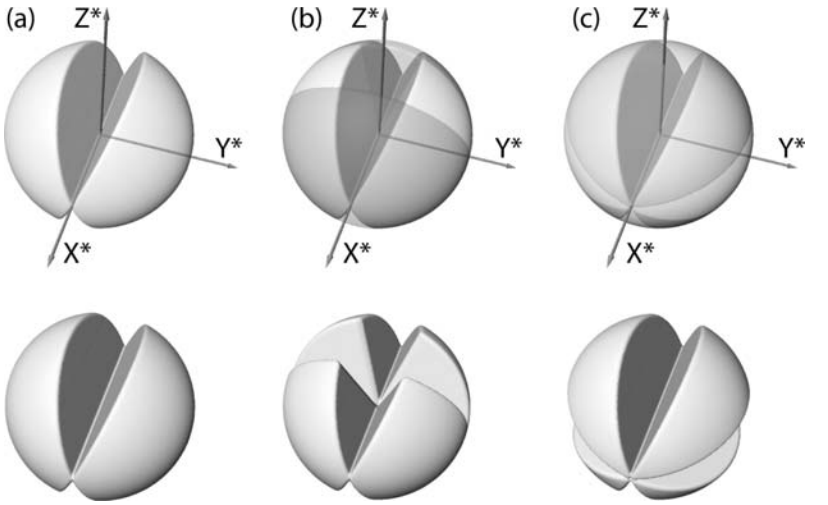


FIGURE 2. The missing wedge. Each molecular volume (repeat) extracted from a single axis tomogram has associated with it a wedge of missing data in its Fourier transform. Depending on the orientation of the repeats with respect to the tilt axis, after repeats are aligned, their missing wedges will be differently oriented. Both alignment and classification of these repeats can be affected by the missing wedge. (a) The case where all repeats have the same missing wedge orientation. Although the transform data are incomplete, the same portion is missing and the same portion is measured for each repeat. (b) The case where the missing wedges differ by rotation about the Z^* axis. Although less favorable than case (a), each repeat's transform will have in common the X^*–Y^* plane. (c) The case where the missing wedges can have any relative orientation. This worst case would occur when the missing wedge of one aligned repeat fell entirely within the measured data of another repeat. For the same tilt angle range, this case would have the smallest amount of data in common, whereas case (a) would have the largest amount of data in common. Case (b) would be intermediate because there would always be some overlap of the two missing wedges.

Technically, all atoms in the structure contribute to all Fourier coefficients, so no specimen feature should be entirely missing from a 3D reconstruction even with a ±30° missing wedge. However, in the low-resolution realm, where electron tomography operates, the contribution of some features may be concentrated in a small region of Fourier space. The best examples of this situation would be the density profile across a lipid bilayer or the mean radial density distribution of a filament. At low resolution, mean density profiles and distributions are sometimes the only features visible. The membrane profile is a 1D projection onto the normal to the membrane and thus contributes to only a single line in the Fourier transform. Tomograms of lipid vesicles or membranes reveal the lipid bilayer only along the lateral edges of the vesicle; the bilayer is entirely absent from the top and bottom (Dierksen *et al.*, 1995), where the membrane profile from this part of the vesicle would come to lie within the missing wedge. In

real space, this phenomenon is equivalent to the statement that one cannot see the contrast across the membrane without viewing along the membrane plane, which is impossible for the top and bottom of the vesicle because of tilt angle restrictions in the electron microscope. Membrane-bound molecules selected from the lateral edges would contain the membrane density, while those picked from the top and bottom surfaces would not. Classification of a collection of all these molecules using MSA would be expected to segregate those from the top and bottom from those along the lateral edges irrespective of other structural similarities.

The equator of the Fourier transform of a filament contains the information on the mean radial density distribution of a filament. This part of the transform is essentially a disk that is perpendicular to the filament axis. If this disk comes to lie within the missing wedge, which will occur when the filament axis is perpendicular to the tilt axis and the filament lies in the specimen plane, the filament will all but disappear in the tomogram (Mastronarde, 1997). These problems have provided much of the incentive toward development of double-axis (Mastronarde, 1997; Penczek *et al*., 1995) and conical tilt (Lanzavecchia *et al*., 2001) tomography.

In the IFM work, it has been possible to collect multiple tomograms from sections oriented in essentially the same way with respect to the tilt axis and, hence, the missing wedge orientation is the same over the population of repeats. This effectively factors out the missing wedge from affecting the classification or the alignment. This has greatly simplified the development of the process. Below, classification and alignment for a more general population of volumes and the possible impact of double-axis tomography on these problems will be revisited.

3. IDENTIFYING MOTIFS IN TOMOGRAMS

Selecting repeating motifs from tomograms is in many ways similar to selecting particles in projections. Manual selection is always an option, just as it is for processing projection images, but automation is obviously desirable. When some kind of preferred orientation is present, such as would occur with a paracrystalline specimen, computation of a cross-correlation function greatly simplifies the problem of locating and extracting the individual repeats. In IFM, the repeats usually have at least some similarity, because they are the products of interactions between filaments that are arranged in parallel, so that fitting a lattice to the peak locations of a cross-correlation function is normally possible. This has been done as part of a 3D 'unbending' scheme to correct for out-of-plane bending (Winkler and Taylor, 1996). On the other hand, if the repeats all lie in a reasonably flat plane, the projection of the tomogram along the 'z'-axis can often be used to determine the 'x, y' coordinates, while the average 'z' value can be obtained by inspection. In this case, any of the lattice-fitting programs used

to analyse 2D crystals would be sufficient to fit a lattice to the '*x*, *y*' coordinates (Henderson *et al*., 1986; Schmid *et al*., 1993).

When there is no lattice, and repeating motifs are essentially randomly placed in the tomogram, an automated motif search is computationally much more demanding but has the benefit of providing information on relative orientation. Speed is important in this process because the reference must be rotated over all of Euler space in fine enough increments that potential repeats are not missed because of a large difference in orientation compared with the reference. A number of efforts have been made in this area (Böhm *et al*., 2000; Frangakis *et al*., 2002; Rath *et al*., 2003). Böhm *et al*. (2000) segmented motifs using a denoising algorithm and then used cross-correlation to determine their orientations with respect to a reference. Frangakis *et al*. (2002) and Rath *et al*. (2003) used cross-correlation techniques to identify motifs in tomograms of ice-embedded specimens automatically (also see Chapter 14 of this volume).

The use of cross-correlation to localize variably oriented motifs has some potential pitfalls. The correlation peak height, which is the measure of similarity between the reference and the 'raw' volume, depends not only on the similarity in orientation and structure but also on the local normalization, which in turn is affected by the missing wedge (Frangakis *et al*., 2002). Thus, similarities in particle structure and alignment could be offset by missing wedge orientation, and thus the search may fail to identify valid repeats, i.e. it may produce false negatives. A fast, locally normalized cross-correlation function (Roseman, 2003) was specifically developed to address the normalization issue for the selection of particles for single-particle reconstruction and has been applied to the selection of 3D motifs within tomograms of ice-embedded samples (Rath *et al*., 2003). Rath *et al*. obtained the unexpected result that searching using a reference extracted from the tomogram worked better than a reference with high signal-to-noise ratio obtained by single-particle reconstruction. The reference extracted from the tomogram has a missing wedge of the same size as the other repeats within the tomogram, but which may also differ in orientation. However, it also contains the context, which includes the membrane environment. The single-particle reference had no missing wedge, but it lacked the membrane environment. This illustrates that a prime advantage of tomography is its ability to visualize molecules in the natural context. Signal-to-noise ratio in the reference may be less important than having the correct context.

4. VOLUME ALIGNMENT

Any effort toward classifying volumes requires alignment to an appropriate reference. For alignment of projections, only three degrees of freedom need be determined—two translational and one rotational—and numerous methods have evolved to determine them. The autocorrelation

function, which is invariant to the position of the particle, has been used extensively to determine the rotational alignment among a set of particles (for an early application, see Kessel *et al*. (1980). In volume alignment, there are six degrees of freedom—three translational and three rotational—so the alignment step is computationally much more expensive.

In image alignment, it is commonly thought that reference bias must be avoided in order that the result is not dependent on the choice of reference. A solution to the alignment problem is to perform classification initially using a derived function that is translationally and, if possible, rotationally invariant. In 2D image processing, the double autocorrelation function (Schatz and van Heel, 1990) and the double self-correlation function (Schatz and van Heel, 1992) have been developed as a solution to this problem. For volumes, computation of invariant functions is more difficult. Functions that are translationally and rotationally invariant and generally applicable for volumes have not been derived. However, invariant functions suitable for specific applications have been constructed.

For IFM, the alignment and classification problem is simplified because the repeats are already partially aligned within a few degrees with respect to rotations about the filament axis and with respect to displacement, by virtue of the fact that they are embedded in a lattice. Thus, Winkler and Taylor (1999) constructed a modified autocorrelation function that was translationally invariant and could be used to distinguish the 180° ambiguity of the thin filament orientation described above. The invariant function was computed as follows. First a copy of the repeat was produced and rotated 180° about the filament axis. Mutual autocorrelation functions (Van Heel *et al*., 1992) of both were computed. The sum and difference of these two functions represent the symmetric and antisymmetric parts of the structure. The antisymmetric part contains the information on the asymmetry between the left and right sides of the actin filament, which as mentioned above is key for correct classification. Similar repeats could then be aligned among themselves, thereby avoiding the possibility that a large difference in structure could influence the alignment. After members of the different classes were identified, they were aligned with each other before averaging. This step was then followed by several cycles of multireference alignment and classification.

The result of this effort was a considerable improvement in the appearance of the structures compared with the raw repeats, and thereby facilitated the construction of atomic models (Chen *et al*., 2002). In this first attempt, a major limitation was the number of repeats. Starting with 480 repeats, removal of poorly structured and preserved repeats reduced the number by 10–20%. From the remaining ~400 repeats, 25 classes were computed so that the improvement in signal-to-noise ratio was by only a factor of 4 on average in each class. This improvement was reflected in the resolution assessed by the spectral signal-to-noise ratio (Unser *et al*., 1987), which was 7 nm.

5. REFERENCE-BASED ALIGNMENT SCHEMES

To answer fundamental questions about muscle contraction, a resolution of at least 5 nm would be needed so that an atomic model of the actin filament can be positioned independently into the density without interference by the positions of the myosin cross-bridges. The scheme outlined above, which used an invariant function combined with multireference alignment of the actin filament and cross-bridge, did not result in a resolution sufficient to resolve actin monomers. Despite the potential for bias, Liu *et al*. (2004) implemented a reference-based alignment scheme to improve the resolution of the class averages. The approach was an extension, to 3D objects, of a method used earlier to visualize heterogeneous conformations in myosin and actin (Burgess *et al*., 2004).

The most invariant structure within each repeat is likely to be the actin filament. Liu *et al*. made the assumption that the structure of the actin filament was invariant in each repeat, which is reasonable at the resolution that electron tomograms can be obtained. First, the partly aligned raw repeats are aligned to a single actin filament, selected from among the population to act as a reference. The alignment involves only a small angular search within an angular cone of 4° radius about the filament axis, to remove any out-of-plane tilt of the filament segment and to correct for in-plane bends in the filament. The search is repeated after rotation of 180° about the filament axis. Then several cycles of MSA, using a mask that selects the actin filament, followed by multireference alignment, were carried out until the global average of all the actin filaments revealed the subunit structure along the actin filament. The procedure does not guarantee that the actin monomers can be resolved in the global average since that is also a function of preservation and staining. Cross-validation that the alignment is working is provided by the presence of the large troponin complex of IFM (Bullard, 1984), which should become emphasized in the averages if the alignment is good. In addition, it is known from studies of actin filaments decorated by myosin heads that the orientation of the actin filament in the regions where the myosin heads bind in rigor IFM is such that the actin monomers are positioned optimally to reveal the 2.75 nm offset between the two long pitch actin strands.

6. CLASSIFICATION

Image classification is the heart of attempts to identify structures and improve the signal-to-noise ratio. There are several schemes that are used in 2D image classification, among them K-means and hierarchical ascendant methods (see, for example, Frank, 1990). The potential benefits of one over the other have not been explored for 3D image classification. The classification of IFM cross-bridge repeats used hierarchical ascendant methods.

Recently, tomograms of negatively stained integrins were classified using K-means (Iwasaki *et al*., 2005). Some newer methods such as self-organizing maps (SOMs) have proven to be effective in clustering cross-bridge repeats (Pascual-Montano *et al*., 2002).

The advent of CCD cameras, automated electron tomography (Braunfeld *et al*., 1994; Koster *et al*., 1992) and rapid freezing/freeze substitution methods for specimen preservation has improved tomograms considerably. The combination of these improvements with single-reference alignment and cross-bridge classification has facilitated routine resolution of both actin monomers and myosin heads in rigor muscle (Fig. 3). Class averages show all the predicted arrangements of myosin heads, but with sufficient resolution that atomic models can be built into the envelopes. From these atomic models, the effect of an imposed stretch on the rigor fibers could be determined at the level of the individual myosin heads (Liu *et al*., 2004). A shell of stain surrounding the features was also clearly resolved (Fig. 3a) so that the filaments appeared negatively stained. This is probably because at the low temperatures of freeze-substitution, where the tannic acid–uranyl acetate fixative is applied, stain penetration into the protein is poor. In addition, features such as the α-helical coiled-coil S2 domain that links the myosin heads to the thick filament backbone are now routinely seen in some classes (Fig. 3b). Resolution in class averages has improved to ~4.8 nm in nearly all class averages (Fig. 3c). Further improvements in the signal-to-noise ratio can be expected with increases in the numbers of repeats that are aligned and classified.

6.1. Mask Determination

Construction of masks for 3D classification is complicated by the three-dimensionality of the repeat, but is conceptually the same as mask construction for the purpose of 2D projection classification (Frank, 1996). Construction usually consists of computing a global average, thresholding the average, low-pass filtering the resulting binary image, and thresholding it again to produce the final expanded, smoothed mask. In the muscle lattice, not all features are equally heterogeneous after alignment, and this offers the opportunity to obtain higher signal-to-noise ratio improvement of those parts of the structure that are anticipated to have the least degree of heterogeneity.

Reference-based alignment of repeats from IFM tomograms assumed that the actin filament was constant in structure. The actin filament should then be a source of low inter-repeat variance, and the cross-bridges, i.e. the desired part of the structure, the source of highest inter-repeat variance. Construction of a mask that contains just the actin filament facilitates the calculation of a reduced number of classes from the least variable part of the structure. Conversely, eliminating the actin filament from the mask facilitates calculation of a greater number of classes for the more variable parts

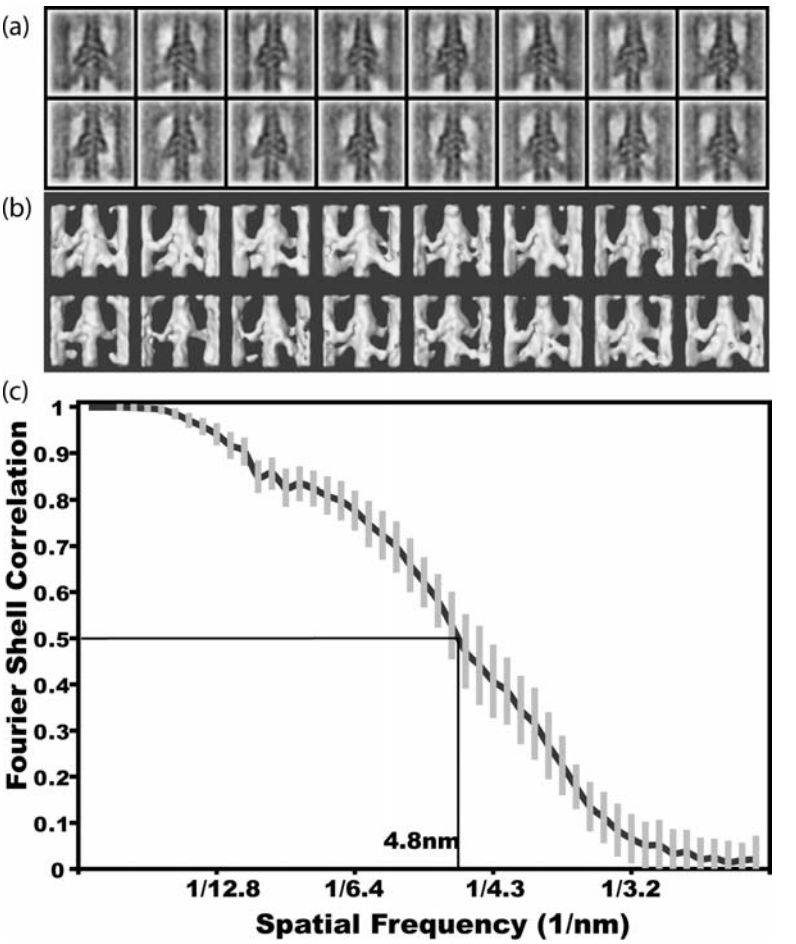


FIGURE 3. Single reference alignment of 322 repeats from stretched rigor myofibrils (Liu *et al.*, 2004). In this example, in which the tomogram has been improved by (1) collecting data on a CCD camera, (2) collecting the tilt series by automated, low-dose tomography and (3) improved specimen preservation rapid freezing/freeze substitution, there is even better definition of the actin filament structure and the individual myosin heads. (a) Central z-section from each of 16 class averages. (b) Surface view of these class averages. (c) Fourier shell correlation showing improvement in resolution to better than 5 nm for most of the class averages.

of the structure, i.e. the cross-bridges. As described below, reassembly of the actin filament and cross-bridge classes can restore the original context.

Another possible approach is to design the mask based on the variance map of the global average rather than the global average itself. This procedure tends to concentrate the classification on those parts of the structure that exhibit the highest variance. However, it has the risk that with

highly heterogeneous repeats such as occur in IFM, the variance map will be relatively featureless with regards to the most variable structures. This approach may work better for those cases where the inter-repeat variance is concentrated on a single feature.

6.2. Number of Classes to Compute

The goal of 3D repeat classification in muscle is to bin a continuum of varying structures. This begs the question of what would be the most suitable number of bins. Considerations of signal-to-noise improvement and retention of information on structure variability are conflicting goals; higher signal-to-noise improvement is achieved by computing fewer classes, but it reduces the inter-image variance retained among the classes. Conversely, retaining more image variance by computing more class averages results in less improvement in signal-to-noise ratio. Hierarchical ascendant classification (HAC) has the benefit of allowing different numbers of classes to be computed from a single classification run. However, HAC does not provide a ready estimate of the number of different classes present in the data and so requires that one examine the dendrogram and observe the relationship between classes.

Reference-based alignment using a constant feature from among the repeats tends to minimize one potential source of structure variability, the actin filament, leaving the cross-bridges as the major remaining source. Winkler and Taylor (1999) made a crude approximation of the number of possible *arrangements* of myosin heads in rigor repeats (nine) and almost doubled that number to allow for some variation in *structure* among the different arrangements. The 16 classes had between nine and 29 repeats per class. Using the numbers of repeats per class and the number of myosin heads per repeat from molecular modeling indicated 2.7 myosin heads per class on average (Chen *et al*., 2002), a number that was in excellent agreement with that found in experiments (Goody *et al*., 1985; Lovell *et al*., 1981; Thomas *et al*., 1983).

Classification of IFM cross-bridge repeats using SOMs (Pascual-Montano *et al*., 2002) has the attractive feature of being fast and well adapted to the analysis of large noisy data sets and that it can predict the number of classes. Using the same set of aligned repeats used previously in a HAC classification, SOMs predicted a smaller number of classes, but those that were predicted were nearly identical to the 16 classes computed using HAC, as described above.

Still, deciding *a priori* on the correct number of classes in such heterogeneous data remains a difficult question even for rigor muscle; it is a much more serious question for contracting muscle where the myosin head population is distributed over the entire catalytic cycle. One possible solution is separate classification of parts of the repeats. So far, all attempts at classification of IFM repeats have done so by including the cross-bridges

on both sides of the actin filament within the classification mask. However, the myosin head distribution on the left side of the actin filament is independent of the head distribution on the right hand side. The target zone in contracting muscle is three actin monomers long on one side of the actin filament and only two actins long on the other (Tregear *et al.*, 2004). In addition, in contracting muscle, the axial register of the thick filaments is lost (Taylor *et al.*, 1999) so that no correspondence exists between myosin head origins on the thick filaments on either side of the actin filament. Therefore, there is no strong justification for classifying repeats using both left and right sides of the 38.7 nm axial period in a single classification run; classification should be done on the left and right sides independently.

This approach could significantly increase the amount of structure variance recoverable from image classification while at the same time conserving the signal-to-noise improvement. For example, suppose there were nine different ways of arranging myosin heads on each side of the actin filaments within each repeat. Classification would require 81 classes to recover the original arrangement. However, if classification of each side were done separately, only nine classes would need to be computed on each side, resulting in a 3-fold improvement in signal-to-noise ratio for the same number of repeats. By recombining the separate left and right side class averages in all combinations, 81 classes can be obtained. The benefits of this approach can only be illustrated after first describing the process of reassembling tomograms using the class averages.

7. REASSEMBLY OF TOMOGRAMS

Tomograms computed from tissue samples such as muscle have a very heterogeneous molecular composition. In muscle, the actin and myosin filaments are the major proteins and their interactions are the main sources of heterogeneity. However, the spatial relationship between actin and myosin filaments and the cross-bridges are the key to understanding muscle function. In muscle, the context is therefore an important factor in understanding function, but signal-to-noise improvement is necessary if molecular models are to be built. MSA can provide the signal-to-noise improvement, but extracting the repeats from the tomogram removes them from their cellular context.

The alignment scheme used for the IFM repeats applies a single transformation to align the raw repeats to the reference. To align the class average to the original raw repeat requires only a single inverse transformation. Replacement of the raw repeat in the original tomogram by the class average to which it contributed can restore the context. The process has been named ‘mapping back’ (Liu *et al.*, 2004). We have used this process to visualize the pattern of cross-bridge repeats in their original context. In tomograms of fast-frozen, freeze-substituted IFM rigor fibers that had been

subjected to a ramp stretch, we used map-backs to visualize patterns in cross-bridge distortion (Fig. 4).

In principle, a judicious choice of alignment reference and classification mask can facilitate the averaging of different structures in a multistructure

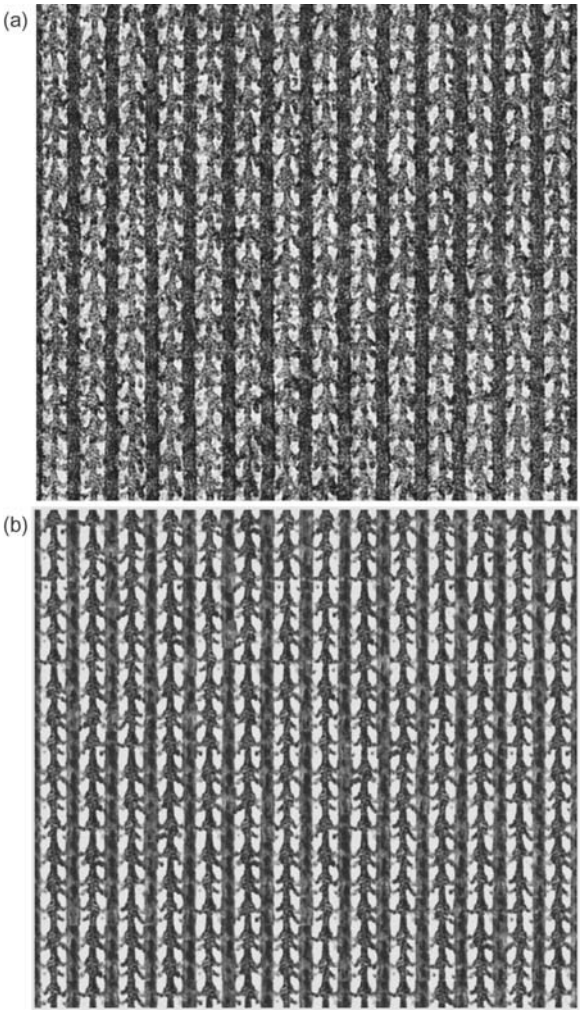


FIGURE 4. Reassembly of the raw tomogram from class averages. (a) Projection of the raw tomogram. (b) Volume-rendered view of (a) reassembled from class averages computed from its array of repeats. This is done by replacing each raw repeat by the class average to which it contributed, accounting for the possibility that the unit cells can be rotated by 180° about the actin filament axis during the alignment to account for the 28/13 helical structure of the actin filament. This process can help to recover contextual information on the lattice arrangement that is otherwise lost when the individual repeats are extracted from the tomogram.

sample. In muscle, actin filaments, myosin filaments and cross-bridges constitute three related, but partially independent structures. A plausible scheme for reassembly of a tomogram using high signal-to-noise ratio class averages might include the following. (i) As shown above, alignment on actin minimizes one source of structure variability, but does not necessarily eliminate it. However, with the proper choice of mask, classification could be done on the actin filament itself to produce class averages independent of the cross-bridges. (ii) Classification of cross-bridges independent of actin would then be done using a mask that excludes the actin (and myosin filament as well) but includes the cross-bridge features. (iii) To obtain information on the thick filament would require a separate alignment and classification scheme. One possibility would combine alignment of the individual 14.5 nm axial repeats based on the thick filament backbone features with classification based on the myosin heads. The contracting muscle could be visualized at a high signal-to-noise ratio with capture of much of the structure variance by reassembling each raw repeat with the actin filament class average to which it contributed, the cross-bridge classes to which it contributed and the classes of the neighboring thick filaments to which it contributed. These reassembled repeats could then be inserted back into their initial coordinates in the raw tomogram, thereby recovering the cellular context but with high signal-to-noise ratio averages.

8. OTHER EXAMPLES OF 3D VOLUME CLASSIFICATION

MSA of 3D images in the context of tomography has been applied to a number of other systems. The first attempt at correspondence analysis of volumes extracted from tomograms was done on volumes to which the particle symmetry was imposed, thereby reducing the effect of the missing wedge (Walz *et al.*, 1997). Later, MSA was used to evaluate the accuracy of an automated motif selection procedure (Frangakis *et al.*, 2002). Classification of 50 ryanodine receptor volumes after automated extraction from a tomogram and clustered with respect to factors 1 and 2 resulted in a class average of eight repeats (Rath *et al.*, 2003).

Electron tomography has been applied to enveloped viruses with glycoprotein surface spikes (Env). In one of these studies, MSA was used to characterize the spike proteins. Extracted and aligned Env molecular volumes from the virus surface will have missing wedges in many orientations, corresponding to case (3) in Figure 2, thereby complicating MSA and particle alignment. Förster *et al.* (2005) used the geometry of the Env particles relative to the virus center to reduce the Euler angle search range and selected 1,114 Env particles for alignment and averaging. They accounted for the missing wedge with a specially weighted cross-correlation function and by weighting coverage in Fourier space in the averages. A second application determined a structure for the Env spikes using cryotomography of intact Simian

Immunodeficiency Virus (Zhu *et al*., 2006). In this study, separate averages were obtained from 6,175 Env particles separated into top-bottom and side subgroups. Although no special cross-correlation functions and missing-wedge weightings were used, MSA was incorporated into cycles of alignment, classification and averaging and later used to examine variations in structure within the two groups.

Two applications of electron tomography of nuclear pores have been reported (Beck *et al*., 2004; Stoffler *et al*., 2003), neither of which used MSA as described here to obtain high signal-to-noise ratio averages. Stoffler *et al*. obtained single axis tomograms from isolated nuclear membranes. Beck *et al*. (2004) obtained tomograms from isolated *Dictyostelium* nuclei and grouped and averaged nuclear pore volumes based on the center of mass of the density within the pore.

9. SINGLE-AXIS VERSUS DOUBLE-AXIS TOMOGRAPHY

In this chapter, we have described the application of MSA to 3D volumes and some ways in which the method can be used to improve the molecular information obtainable from tomograms. The problems associated with the missing wedge have been described with respect to single-axis tomography. Double-axis tomography is superior to single-axis tomography because the specimen transform is more uniformly measured; instead of a missing wedge, there is a missing pyramid or, in the case of conical geometry, a missing cone.

Double-axis tomography for plastic-embedded or negatively stained specimens is relatively easy if the specimen will tolerate the additional electron dose (but see Chapter 1 of this volume). Even when a rotation holder is not available, simply extracting the grid, rotating it 90° in the holder and then collecting a second tilt series is a simple solution. The separate tilt series can be merged into a single tomogram (Mastronarde, 1997; Penczek *et al*., 1995) or merged simultaneously (Chen *et al*., 2001). Most of the published work in IFM tomography has involved single-axis tilt series (Liu *et al*., 2004; Schmitz *et al*., 1996, 1997; Taylor *et al*., 1999) but the early efforts on repeat classification used a double-axis tilt series (Chen *et al*., 2001, 2002).

Muscle fibrils have the unique advantage that the filaments all lie in the same orientation. Placing the filament axis parallel to the tilt axis, which as described above is the optimal orientation for filament visibility, is relatively easy. Is there any reason to use double-axis tilt series for this kind of specimen? Current efforts on classification of repeats in IFM are using double-axis tilt series, and the preliminary results indicate that double-axis tomography is superior. Although single-axis tilt series with the tilt axis parallel to the fiber axis gives the most accurate rendition of the filament profiles, the cross-bridges generally run perpendicular to the filament axis, so this orientation is least optimal for them. A second tilt series with the axis

perpendicular to the filaments produces superior definition of the cross-bridges even in the presence of poor filament definition. Thus, whenever possible, double-axis tilt series are recommended if all the data must come from a single specimen.

Double-axis tomography is quite demanding with ice-embedded specimens because of the radiation sensitivity of the specimen. Double-axis tomograms of ice-embedded specimens have been obtained (Nickell *et al.*, 2003), although MSA was not performed on any molecular images derived from them. However, if averages are necessary to obtain the desired information, and 3D images can be obtained over a wide range of orientations within a single tomogram, or from several tomograms with different tilt axis orientations, then single-axis tomography should be sufficient to obtain an average whose transform samples all of Fourier space or, at worst, leaves only a missing cone of data, provided that the volumes are uniformly distributed throughout orientation space. This has been the case with many of the studies that have extracted molecular images from tomograms of frozen-hydrated specimens for classification and/or averaging (Beck *et al.*, 2004; Förster *et al.*, 2005; Liu *et al.*, 2006).

The missing wedge can also affect the averages if not treated explicitly. In Fourier space, it is possible that the missing wedge, a region with no data, can overlap regions with data from another volume. Averaging without explicit treatment of the missing wedge will cause amplitude loss at high resolution. Thus, averaging in Fourier space, where weighting schemes can account explicitly for the missing wedge, would be necessary to obtain the best results. Double-axis tomography alleviates this requirement but, in some circumstances, may not eliminate it entirely. Nevertheless, reducing the missing wedge to a missing cone or pyramid will have positive effects on both alignment and classification, and is worth the additional effort if data can be obtained without heroic struggle.

10. PROSPECTS

The application of MSA to repetitive features in electron tomograms provides a way to improve the signal-to-noise ratio even in the presence of conformational heterogeneity. There are many potential applications where classification can improve the quality of averages or can distinguish subtle but functionally important differences. As described above, MSA applied to envelope viruses has succeeded in producing higher-quality images of envelope glycoproteins, but it could also be applied to density within the virus, and might be used to deduce connections between the envelope proteins and the underlying tegument (Grünewald *et al.*, 2003).

The formation of protein arrays can reduce conformational variability, but does not automatically lead to arrays with high order. Thus, electron tomography can be used to obtain a 3D image, and MSA can be used to

determine the number of conformers that are present in the arrays. This philosophy has been used to produce high-quality averages of myosin V in the inhibited state from tomograms of ice-embedded, disordered 2D arrays after application of a focus gradient correction (Winkler and Taylor, 2003). A resolution of 2.4 nm was obtained from ~6,000 molecules extracted from the tomograms (Liu *et al*., 2006). Thus, resolution approaching that which is readily obtained using negative stain preservation can be achieved, even when using ice-embedded specimens. The result illustrates the promise of cryoelectron tomography for characterizing large-amplitude motions in macromolecular assemblies, not in real time, but as a series of snapshots.

There are many other possible structures that have been difficult to image in 3D in the past but can now be approached at resolutions near 2 nm, with signal-to-noise ratios that are good enough for insightful model building.

ACKNOWLEDGEMENTS

The research on insect flight muscle was supported by NIH Grant GM30598. We thank Mary Reedy for the micrograph shown in Fig. 1. We thank Joachim Frank for his critical reading and insightful comments during the development of this chapter.

REFERENCES

Beck, M., Förster, F., Ecke, M., Plitzko, J.M., Melchior, F., Gerisch, G., Baumeister, W. and Medalia, O. (2004). Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. *Science* **306**:1387–1390.

Böhm, J., Frangakis, A. S., Hegerl, R., Nickell, S., Typke, D. and Baumeister, W. (2000). Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proc. Natl. Acad. Sci. USA* **97**:14245–14250.

Braunfeld, M. B., Koster, A. J., Sedat, J. W. and Agard, D. A. (1994). Cryo automated electron tomography: towards high-resolution reconstructions of plastic-embedded structures. *J. Microsc*. **174**:75–84.

Bullard, B. (1984). A large troponin in asynchronous insect flight muscle. *J. Muscle Res. Cell Motil.* **5**:196.

Burgess, S. A., Walker, M. L., Thirumurugan, K., Trinick, J. and Knight, P. J. (2004). Use of negative stain and single-particle image processing to explore dynamic properties of flexible macromolecules. *J. Struct. Biol*. **147**:247–258.

Chen, L. F., Blanc, E., Chapman, M. S. and Taylor, K. A. (2001). Real space refinement of actomyosin structures from sectioned muscle. *J. Struct. Biol*., **133**, 221–232.

Chen, L. F., Winkler, H., Reedy, M. K., Reedy, M. C. and Taylor, K. A. (2002). Molecular modeling of averaged rigor crossbridges from tomograms of insect flight muscle. *J. Struct. Biol*. **138**:92–104.

DeRosier, D. J. and Moore, P. B. (1970). Reconstruction of three-dimensional images from electron micrographs of structures with helical symmetry. *J. Mol. Biol*. **52**:355–369.

Dierksen, K., Typke, D., Hegerl, R., Walz, J., Sackmann, E. and Baumeister, W. (1995). Three-dimensional structure of lipid vesicles embedded in vitreous ice and investigated by automated electron tomography. *Biophys. J.* **68**:1416–1422.

Egelman, E. H. (2000). A robust algorithm for the reconstruction of helical filaments using single-particle methods. *Ultramicroscopy* **85**:225–234.

Egelman, E. H. and DeRosier, D. J. (1992). Image analysis shows that variations in actin crossover spacings are random, not compensatory. *Biophys. J.* **63**:1299–1305.

Egelman, E. H., Francis, N. and DeRosier, D. J. (1982). F-actin is a helix with a random variable twist. *Nature* **298**:131–135.

Förster, F., Medalia, O., Zauberman, N., Baumeister, W. and Fass, D. (2005). Retrovirus envelope protein complex structure *in situ* studied by cryo-electron tomography. *Proc. Natl. Acad. Sci. USA* **102**:4729–4734.

Frangakis, A. S., Bohm, J., Förster, F., Nickell, S., Nicastro, D., Typke, D., Hegerl, R. and Baumeister, W. (2002). Identification of macromolecular complexes in cryoelectron tomograms of phantom cells. *Proc. Natl Acad. Sci. USA* **99**:14153–14158.

Frank, J. (1990). Classification of macromolecular assemblies studied as 'single particles'. *Q. Rev. Biophys*. **23**:281–329.

Frank, J. (1996). *Three-dimensional Elecron Microscopy of Macromolecular Assemblies*. Academic Press, San Diego, CA.

Geeves, M. A. and Holmes, K. C. (1999). Structural mechanism of muscle contraction. *Annu. Rev. Biochem*. **68**:687–728.

Glaeser, R. M. (1999). Review: electron crystallography: present excitement, a nod to the past, anticipating the future. *J. Struct. Biol*. **128**:3–14.

Goody, R. S., Reedy, M. C., Hofmann, W., Holmes, K. C. and Reedy, M. K. (1985). Binding of myosin subfragment 1 to glycerinated insect flight muscle in the rigor state. *Biophys. J.* **47**:151–169.

Grünewald, K., Desai, P., Winkler, D. C., Heymann, J. B., Belnap, D. M., Baumeister, W. and Steven, A. C. (2003). Three-dimensional structure of herpes simplex virus from cryo-electron tomography. *Science* **302**:1396–1398.

Hegerl, R. and Hoppe, W. (1976). Influence of electron noise on three-dimensional image reconstruction. *Z. Naturforsch.* **31a**:1717–1721.

Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. and Zemlin, F. (1986). Structure of purple membrane from *Halobacterium halobium*: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. *Ultramicroscopy* **19**:147–178.

Holmes, K. C., Tregear, R. T. and Barrington Leigh, J. (1980). Interpretation of the low angle X-ray diffraction from insect muscle in rigor. *Proc. R. Soc. B* **207**:13–33.

Hoppe, W. and Hegerl, R. (1981). Some remarks concerning the influence of electron noise on 3D reconstruction. *Ultramicroscopy* **6**:205–206.

Iwasaki, K., Mitsuoka, K., Fujiyoshi, Y., Fujisawa, Y., Kikuchi, M., Sekiguchi, K. and Yamada, T. (2005). Electron tomography reveals diverse conformations of integrin alphaIIbbeta3 in the active state. *J. Struct. Biol*. **150**:259–267.

Kessel, M., Frank, J. and Goldfarb, W. (1980). Low dose electron microscopy of individual biological macromolecules. In: *Electron Microscopy at Molecular Dimensions: State of the Art and Strategies for the Future.* Baumeister, W. and Vogell, W. Eds. Springer-Verlag, Berlin, pp. 154–160.

Koster, A. J., Chen, H., Sedat, J. W. and Agard, D. A. (1992). Automated microscopy for electron tomography. *Ultramicroscopy* **46**:207–227.

Lanzavecchia, S., Cantele, F. and Bellon, P.L. (2001). Alignment of 3D structures of macromolecular assemblies. *Bioinformatics* **17**:58–62.

Liu, J., Reedy, M. C., Goldman, Y. E., Franzini-Armstrong, C., Sasaki, H., Tregear, R. T., Lucaveche, C., Winkler, H., Baumann, B. A. J., Squire, J. M., Irving, T. C., Reedy, M.K. and Taylor, K. A. (2004). Electron tomography of fast frozen, stretched rigor fibers reveals elastic distortions in the myosin crossbridges. *J. Struct. Biol.* **147**:268–282.

Liu, J., Taylor, D. W., Krementsova, E. B., Trybus, K. M. and Taylor, K. A. (2006). 3-D structure of the myosin V inhibited state by cryoelectron tomography. *Nature* **442**:208–211.

Lovell, S. J., Knight, P. J. and Harrington, W. F. (1981). Fraction of myosin heads bound to thin filaments in rigor fibrils from insect flight and vertebrate muscles. *Nature* **293**:664–666.

Mastronarde, D. N. (1997). Dual-axis tomography: an approach with alignment methods that preserve resolution. *J. Struct. Biol*. **120**:343–352.

McEwen, B. F., Downing, K. H. and Glaeser, R. M. (1995). The relevance of dose-fractionation in tomography of radiation-sensitive specimens. *Ultramicroscopy* **60**:357–373.

Nickell, S., Hegerl, R., Baumeister, W. and Rachel, R. (2003). *Pyrodictium cannulae* enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *J. Struct. Biol*. **141**:34–42.

Pascual-Montano, A., Taylor, K. A., Winkler, H., Pascual-Marqui, R. D. and Carazo, J. M. (2002). Quantitative self-organizing maps for clustering electron tomograms. *J. Struct. Biol.* **138**:114–122.

Penczek, P., Marko, M., Buttle, K. and Frank, J. (1995). Double-tilt electron tomography. *Ultramicroscopy* **60**:393–410.

Rath, B. K., Hegerl, R., Leith, A., Shaikh, T. R., Wagenknecht, T. and Frank, J. (2003). Fast 3D motif search of EM density maps using a locally normalized cross-correlation function. *J. Struct. Biol*. **144**:95–103.

Reedy, M. K. (1967). Cross-bridges and periods in insect flight muscle. *Am. Zool.* **7**:465–481.

Reedy, M. K., Holmes, K. C. and Tregear, R. T. (1965). Induced changes in orientation of the cross-bridges of glycerinated insect flight muscle. *Nature* **207**:1276–1280.

Reedy, M. K. and Reedy, M. C. (1985). Rigor crossbridge structure in tilted single filament layers and flared-X formations from insect flight muscle. *J. Mol. Biol*. **185**:145–176.

Roseman, A. M. (2003). Particle finding in electron micrographs using a fast local correlation algorithm. *Ultramicroscopy* **94**:225–236.

Schatz, M. and van Heel, M. (1990). Invariant classification of molecular views in electron micrographs. *Ultramicroscopy* **32**:255–264.

Schatz, M. and van Heel, M. (1992). Invarient recognition of molecular projections in vitreous ice preparations. *Ultramicroscopy* **45**:15–22.

Schmid, M., Dargahi, R. and Tam, M. (1993). SPECTRA: a system for processing electron images of crystals. *Ultramicroscopy* **48**:251–264.

Schmitz, H., Reedy, M. C., Reedy, M. K., Tregear, R. T. and Taylor, K. A. (1997). Tomographic three-dimensional reconstruction of insect flight muscle partially relaxed by AMPPNP and ethylene glycol. *J. Cell Biol.* **139**:695–707.

Schmitz, H., Reedy, M. C., Reedy, M. K., Tregear, R. T., Winkler, H. and Taylor, K. A. (1996). Electron tomography of insect flight muscle in rigor and AMPPNP at 23°C. *J. Mol. Biol.* **264**:279–301.

Squire, J. M. (1981). *The Structural Basis of Muscle Contraction*. Plenum Press, New York.

Stoffler, D., Feja, B., Fahrenkrog, B., Walz, J., Typke, D. and Aebi, U. (2003). Cryo-electron tomography provides novel insights into nuclear pore architecture: implications for nucleocytoplasmic transport. *J. Mol. Biol*. **328**:119–130.

Taylor, K. A., Schmitz, H., Reedy, M. C., Goldman, Y. E., Franzini-Armstrong, C., Sasaki, H., Tregear, R. T., Poole, K. J. V., Lucaveche, C., Edwards, R. J., Chen, L. F., Winkler, H. and Reedy, M. K. (1999). Tomographic 3-D reconstruction of quick frozen, Ca^{2+}-activated contracting insect flight muscle. *Cell* **99**:421–431.

Taylor, K. A., Tang, J., Cheng, Y. and Winkler, H. (1997). The use of electron tomography for structural analysis of disordered protein arrays. *J. Struct. Biol.* **120**:372–386.

Taylor, K. A. and Taylor, D. W. (1994). Formation of two-dimensional complexes of F-actin and crosslinking proteins on lipid monolayers: demonstration of unipolar alpha-actinin-F-actin crosslinking. *Biophys. J.* **67**:1976–1983.

Taylor, K. A., Taylor, D. W. and Schachat, F. (2000). Isoforms of alpha-actinin from cardiac, smooth, and skeletal muscle form polar arrays of actin filaments. *J. Cell Biol*. **149**:635–646.

Thomas, D. D., Cooke, R. and Barnett, V. A. (1983). Orientation and rotational mobility of spin-labelled myosin heads in insect flight muscle in rigor. *J. Muscle Res. Cell Motil*. **4**:367–378.

Tilney, L. G., Derosier, D. J. and Mulroy, M. J. (1980). The organization of actin filaments in the stereocilia of cochlear hair cells. *J. Cell Biol*. **86**:244–259.

Tregear, R. T., Reedy, M. C., Goldman, Y. E., Taylor, K. A., Winkler, H., Franzini-Armstrong, C., Sasaki, H., Lucaveche, C. and Reedy, M. K. (2004). Cross-bridge number, position, and angle in target zones of cryofixed isometrically active insect flight muscle. *Biophys. J*. **86**:3009–3019.

Unser, M., Trus, B. L. and Steven, A. C. (1987). A new resolution criterion based on spectral signal-to-noise ratios. *Ultramicroscopy* **23**:39–51.

Van Heel, M., Schatz, M. and Orlova, E. (1992). Correlation functions revisited. *Ultramicroscopy* **46**:307–316.

Walz, J., Typke, D., Nitsch, M., Koster, A. J., Hegerl, R. and Baumeister, W. (1997). Electron tomography of single ice-embedded macromolecules: three-dimensional alignment and classification. *J. Struct. Biol*. **120**:387–395.

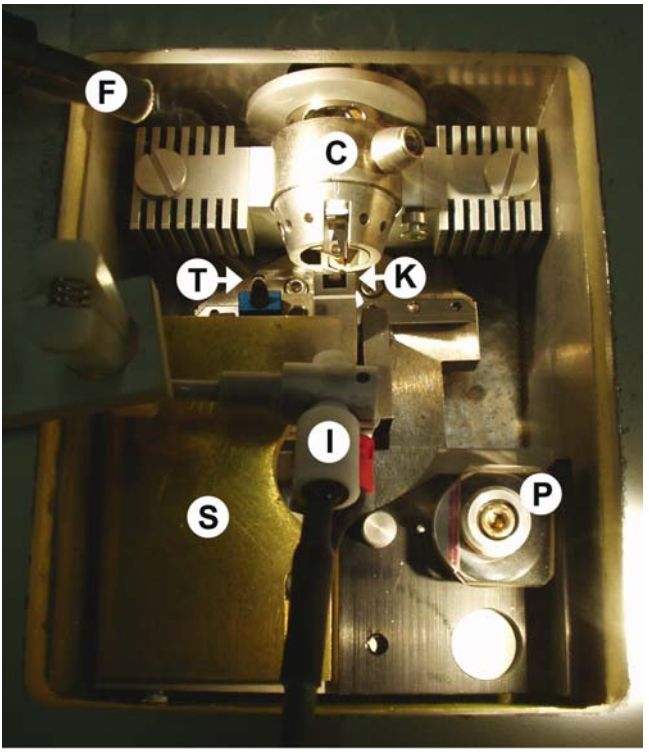
Winkler, H. and Taylor, K. A. (1996). Three-dimensional distortion correction applied to tomographic reconstructions of sectioned crystals. *Ultramicroscopy* **63**:125–132.

Winkler, H. and Taylor, K. A. (1999) Multivariate statistical analysis of three-dimensional cross-bridge motifs in insect flight muscle. *Ultramicroscopy* **77**:141–152.

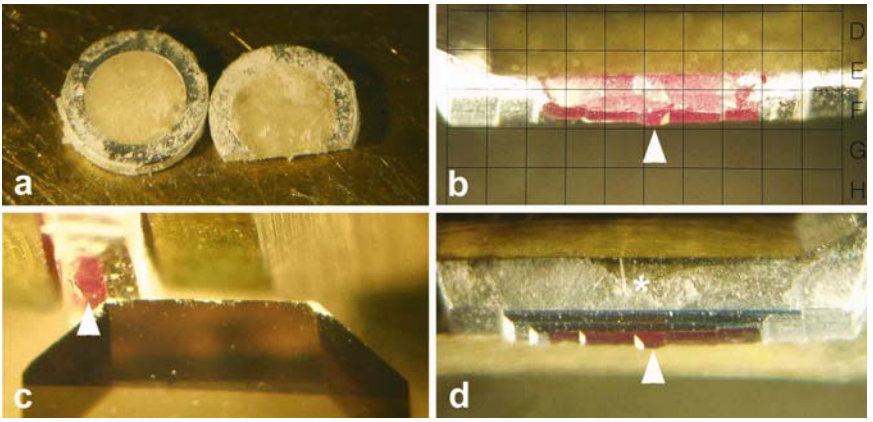
Winkler, H. and Taylor, K. A. (2003). Focus gradient correction applied to tilt series image data used in electron tomography. *J. Struct. Biol*. **143**:24–32.

Zhu, P., Liu, J., Bess, J. J., Chertova, E., Lifson, J. D., Grisé, H., Ofek, G., Taylor, K. A. and Roux, K. H. (2006) Distribution and three-dimensional structure of AIDS virus envelope spikes. *Nature* **441**:847–852.

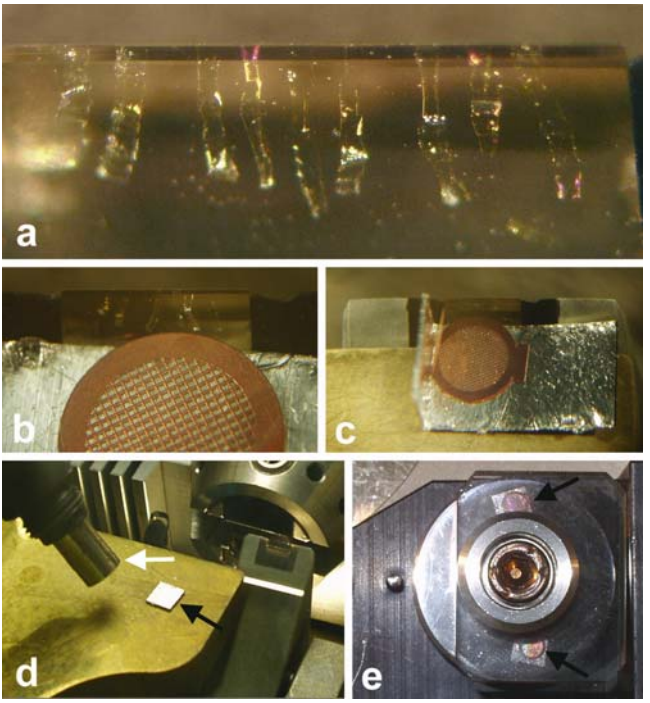
Index



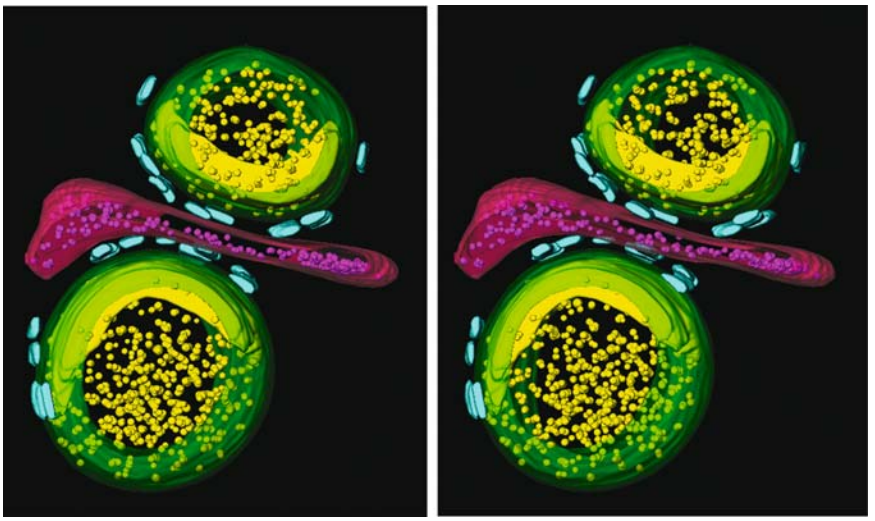
Chapter 2, FIGURE 1. Cryochamber of an UCT ultramicrotome with an EM-FCS cryo kit (Leica, Vienna, Austria). The chamber is cooled to –160°C and is filled with cold, dry nitrogen vapor. One of the two auxiliary fiber-optic illuminators is seen, labeled (F). The specimen is held in the chuck (C). The knife stage is shown positioned for trimming, with the diamond trimming tool (T) adjacent to the specimen. For microtomy, the sectioning knife (K) is moved in front of the specimen, and a brass shelf (S) is pushed forward so that it is close to the knife edge. The shelf, fabricated in-house, provides a surface for collecting sections. Both diamond knives (T and K) are from Diatome (Biel, Switzerland). The head of the ionizer unit (I; Static Line II, Haug, Biel, Switzerland) is mounted ~3 cm from the knife edge. The glass screw press tool is mounted at (P).



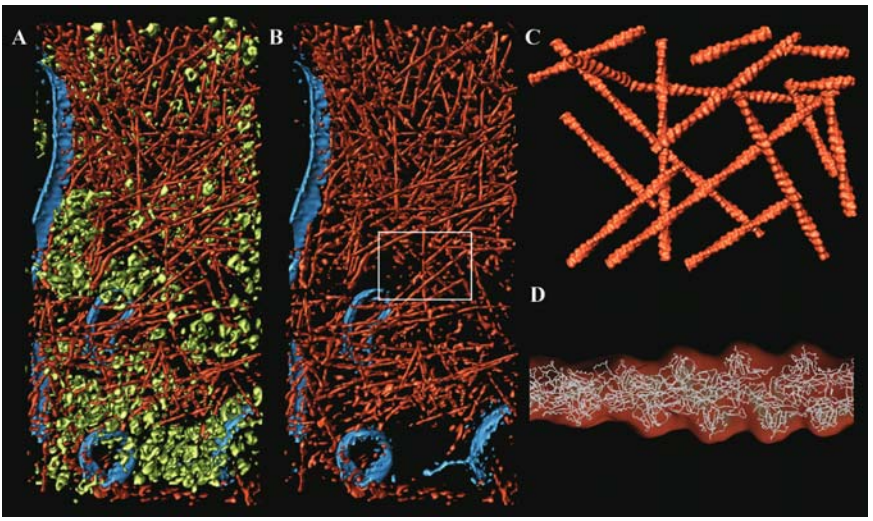
Chapter 2, FIGURE 2. Trimming of high-pressure frozen specimens. (a) A 3 mm diameter aluminum specimen carrier for the HPM010 high-pressure freezer (Bal-Tec, Balzers, Liechtenstein) is shown, before and after trimming. The upper portion of the carrier has been removed, revealing the specimen and allowing inspection for air bubbles and suitable areas for microtomy. (b) A top view of the trimmed specimen carrier shows the trimmed block face, which is shaped like a low mesa (arrowhead), 100 µm square, as seen within a 250 µm graticle square of the stereomicroscope. (c) With the chuck rotated 90° from its position in (b), the diamond trimming tool is used to cut a 45° bevel in the metal on the bottom of the specimen carrier, leaving a thin ridge of specimen from which the mesa (arrowhead) has been cut. (d) Bottom view of the trimmed specimen carrier. The specimen is well supported during microtomy by the remaining metal bevel (*). The mesa is indicated by an arrowhead.



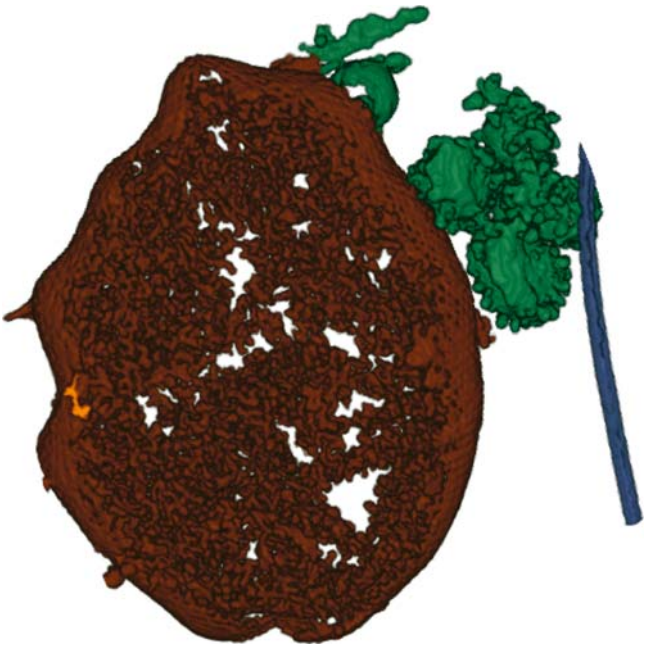
Chapter 2, FIGURE 3. Collection and mounting of sections. (a) A ribbon of sections is cut, and then the knife is moved laterally. A series of ribbons is thus collected on the knife edge. Faint gold interference color can be seen in most sections, indicating a thickness of ~140 nm. In places, a purple color is seen, corresponding to a thickness of ~180 nm. Some sections are curled at the edges, or wrinkled. (b) A Quantifoil grid is placed on a fresh piece of indium foil, just adjacent to the knife edge, and sections are transferred to the center of the grid by means of a dog hair. (c) When folding grids are used, the indium foil is pre-folded to a 90° angle and, after placement of the sections, the indium foil is used to push the grid closed. (d) After the grid is latched, the indium foil is folded closed and the 'envelope' (arrow) is pressed with the polished metal rod (white arrow at left). (e) Quantifoil grids (arrows) are placed, still supported on indium foil, in the glass press tool. An impression of the grid squares in the indium foil indicates the evenness of pressing.



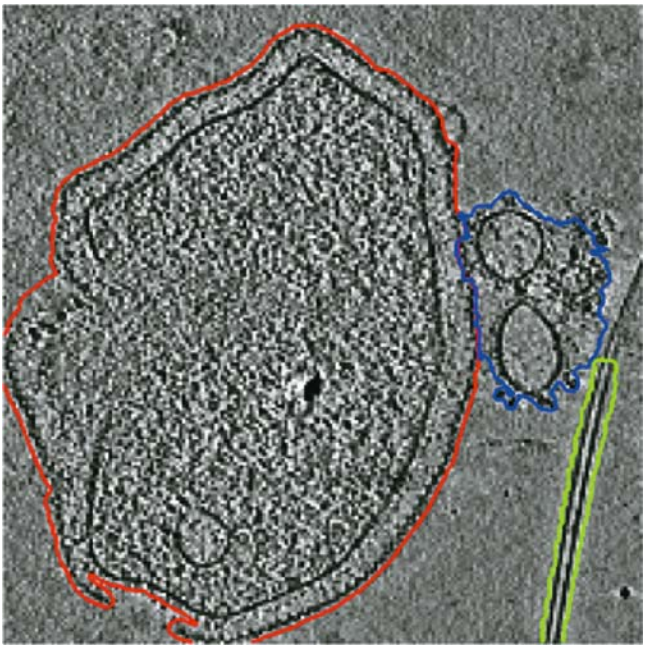
Chapter 4, FIGURE 5. Isolated triad junctions. Stereo-pairs of a surface-rendered representation of a tomographic reconstruction of a triad junction. Sarcoplasmic reticulum- (SR) and transverse tubule-derived vesicles are shown in green and red, respectively. Yellow spheres in the lumen of the SR represent calsequestrin, and continuous yellow slabs near junctional surfaces of the SR represent the condensed calsequestrin. Blue structures correspond to feet/ryanodine receptors. (From Wagenknecht *et al*. (2002), reproduced with permission of the Biophysical Society).



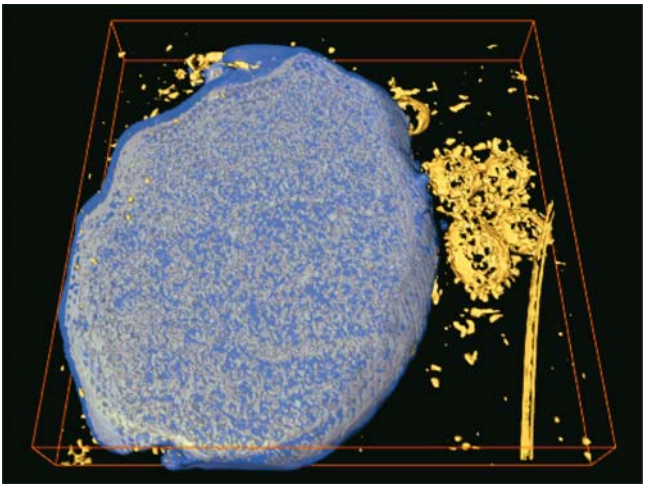
Chapter 4, FIGURE 6. Cytoskeletal network of a *Dictyostelium discoideum* cell. (A) Three-dimensional volume rendering representation of a cortical region of the cytoplasm ($400 \times 870 \times 97$ nm) showing the actin filament network (reddish), membranes (blue) and cytoplasmic macromolecular complexes (green), many of them resembling 80S ribosomes in size and shape. (B) Visualization of the network of actin filaments. (C) Idealized representation of the actin network region marked in (B) at a resolution of 2 nm using the 3D density map obtained from the atomic model of an actin filament. (D) Docking of the atomic model of actin into the density map of an actin filament. (Panel A adapted from Medalia *et al.* (2002b); panels B–D reproduced from Kurner *et al.* (2002) with permission by Elsevier).



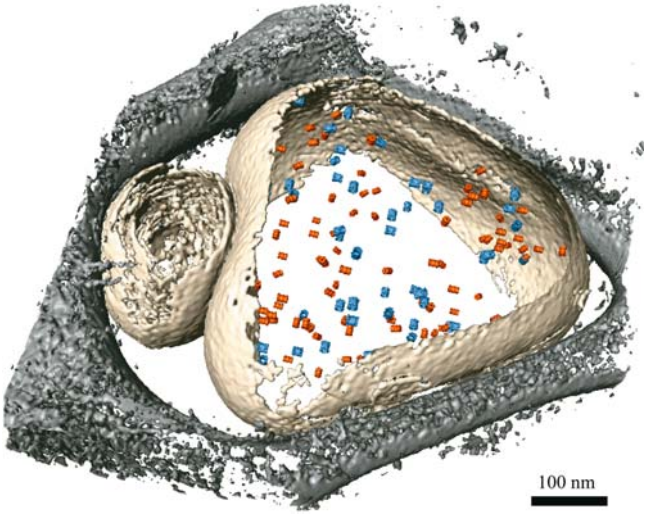
Chapter 12, FIGURE 11. Segmentation of a slice from *Pyrodictium* using level sets according to the procedure described by Bajaj *et al*. (2003).



Chapter 12, FIGURE 13. Segmentation of a slice from the tomogram of *Pyrodictium* using the eigenvector method. The original slice is overlaid with contour plots of the segmentation. The three main features on the image are seen to be separated from one another. (From Frangakis *et al.* (2002), reproduced with permission of Elsevier).



Chapter 12, FIGURE 14. Eigenvector-based segmentation presented in Fig. 13, visualized in three dimensions. (From Frangakis *et al.* (2002), reproduced with permission of Elsevier).



Chapter 14, FIGURE 6. Segmented and denoised 3D image of a vesicle containing two types of macromolecules, visualized in different colors. In the isosurface visualization, two vesicles (in beige) surrounded by the carbon film (in gray) are shown. In the larger vesicle, the two different macromolecular complexes were located and positively identified, and were then replaced with their low-pass-filtered X-ray structure. Proteasomes are shown in orange and thermosomes in blue. (From Frangakis *et al.* (2002), reproduced wth permission of the National Academy of Sciences).